Vragenlijst Zicht op Algoritmes

#	Vraag	Antwoord
1	Wat zijn de kritische primaire/werk processen, alsook de belangrijkste producten/diensten die onder de verantwoordelijkheid van uw departement vallen en waar het gebruik van algoritmes[1], impact heeft?	Klik hier als u tekst wilt invoeren
2	Welke programma's en projecten (nog niet in "productie" maar in het kader van ontwikkeling, innovatie, proeftuinen, labs e.d.) zijn er binnen het departement en de verbonden uitvoeringsorganisaties die betrekking hebben op algoritmes met een voorspellend en/of voorschrijvend karakter?	Klik hier als u tekst wilt invoeren
3	Kunt u de belangrijkste voorspellende en voorschrijvende algoritmes die binnen het departement en de verbonden uitvoeringsorganisaties in gebruik zijn benoemen (max 10)? Het verzoek is of u voor elk van deze algoritme(s) de bijlage op de volgende pagina wilt invullen.	Algoritme 1: <tekstveld> Algoritme 2: <tekstveld> Algoritme 3: <tekstveld> Algoritme 4: <tekstveld> Algoritme 5: <tekstveld> Algoritme 6: <tekstveld> Algoritme 7: <tekstveld> Algoritme 8: <tekstveld> Algoritme 9: <tekstveld> Algoritme 10: <tekstveld>
4	Op welke wijze wordt sturing op en de beheersing van algoritmes vormgegeven en hoe zijn de verantwoordelijkheden belegd? Hierbij doelen wij op de verantwoordelijkheid voor de kwaliteitsaspecten in algemene zin, beheer/onderhoud, het voldoen aan geldende wet/regelgeving maar ook inzicht in en controle op goede werking.	Klik hier als u tekst wilt invoeren
5	Van welke normen- en of toetsingskaders maakt u bij de ontwikkeling, implementatie, beheer van algoritmes gebruik?	Klik hier als u tekst wilt invoeren
6	Wat wilt u in dit kader van dit onderzoek nog aan de Algemene Rekenkamer meegeven (ruimte voor suggesties, aspecten die u belangrijk vindt etc.)?	Klik hier als u tekst wilt invoeren

[1] Algoritmes in "productie": We doelen hierbij dus niet op algoritmes in laboratorium, test of pilot omgevingen. Vraag 2 heeft betrekking op die context.

Bijlage vraag 3: beschrijving van algoritmes

Hieronder kunt u een beschrijving geven van de belangrijkste voorspellende en voorschrijvende algoritmes die binnen het departement en de verbonden uitvoeringsorganisaties in gebruik zijn. Per algoritme dat is opgegeven bij vraag 3, vragen wij u om een tabel met vragen te beantwoorden. De vragen worden op de volgende pagina's herhaald.

Een algoritme is in dit kader "belangrijk" op het moment dat er burgers/bedrijven direct geraakt worden en/of er sprake is van impact op de financiële stromen van het departement en de verbonden uitvoeringsorganisaties.

ALGORITME 1		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[2]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[3]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[2] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[3] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 2		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[4]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[5]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[4] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[5] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 3		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[6]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[7]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[6] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[7] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 4		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[8]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[9]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[8] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[9] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 5		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[10]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[11]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[10] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[11] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 6		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[12]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[13]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[12] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[13] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 7		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[14]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[15]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[14] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[15] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 8		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[16]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[17]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[16] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[17] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 9		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[18]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[19]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[18] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[19] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

ALGORITME 10		
1	Wat is de naamgeving/typering van het algoritme of het systeem waar het algoritme deel van uitmaakt?	Klik hier als u tekst wilt invoeren
2	In welk primair/werkproces of ten behoeve van welke dienst of product speelt het algoritme een rol?	Klik hier als u tekst wilt invoeren
3	Wat doet het algoritme/combinatie van algoritmes/systeem (kernachtige omschrijving)?	Klik hier als u tekst wilt invoeren
4	Wat is de impact op burgers en/of bedrijven?	Klik hier als u tekst wilt invoeren
5	Wat is het financiële belang van de werking van het algoritme voor financiën/financiële stromen van het departement/uitvoeringsorganisatie?	Klik hier als u tekst wilt invoeren
6	Is het algoritme voorspellend of voorschrijvend?[20]	Kies een item
7	Is er sprake van automatische besluitvorming?	Kies een item
8	Welke data/databronnen wordt gebruikt (beschrijving in hoofdlijnen)?	Klik hier als u tekst wilt invoeren
9	Welk techniek is gebruikt?[21]	Klik hier als u tekst wilt invoeren
10	Indien er sprake is van een lerend algoritme, hoe vaak wordt er geleerd/getraind?	Kies een item
11	Welke software/applicatie is gebruikt?	Kies een item Klik hier voor invoeren toelichting
12	Overige opmerkingen/toelichting	Klik hier als u tekst wilt invoeren

[20] *Voorspellend:* het algoritme doet een bepaalde voorspelling of berekent een bepaalde waarschijnlijkheid die wordt gebruikt in werkprocessen en/of besluitvorming. Bij het komen tot een uiteindelijke beslissing of actie speelt de mens nog een rol. *Voorschrijvend:* het algoritme bepaalt en/of dicteert de beslissing/actie of uitvoering.

[21] Een korte typering van gebruikte technieken. Voorbeelden: gebruikte statistische methoden/technieken, AI/ML technieken, onderdeel van systeem/applicatie, combinaties van technieken.

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
van Algemene Zaken
Postbus 20001
2500 EA DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte heer Huijts,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is de heer [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
van Buitenlandse Zaken
Postbus 20061
2500 EB DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nll
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte mevrouw Brandt,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Binnenlandse Zaken en Koninkrijksrelaties
Postbus 20011
2500 EA Den Haag

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 4139
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 25 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte heer Schurink,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Wen

Secretaris van de
Algemene Rekenkamer

VERTROUWELIJK
Secretaris-generaal van het Ministerie
Economische Zaken en Klimaat
Postbus 20401
2500 EK DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte heer Camps,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
Infrastructuur en Waterstaat
Postbus 20901
2500 EX DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte mevrouw Ongering,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
van Justitie en Veiligheid
Postbus 20301
2500 EH DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte heer Riedstra,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
Landbouw, Natuur en Voedselkwaliteit
Postbus 20401
2500 EK DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte heer Goet,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
van Defensie
Postbus 20701
2500 ES DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte mevrouw Van Craaikamp,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted]. Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]. Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

Secretaris-generaal van het Ministerie
van Financiën
Postbus 20201
2500 EE DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nll
W www.rekenkamer.nl

DATUM 24 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte heer Van den Dungen,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
Onderwijs, Cultuur en Wetenschap
Postbus 16375
2500 BJ DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte mevrouw Hammersma,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted]. Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
Sociale Zaken en Werkgelegenheid
Postbus 90801
2509 LV DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte mevrouw Mulder,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is d[redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] et projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Secretaris van de
Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie
Volksgezondheid, Welzijn en Sport
Postbus 20350
2500 EJ DEN HAAG

Lange Voorhout 8
Postbus 20015
2500EA Den Haag
T [070] 342 43 44
E voorlichting@rekenkamer.nl
W www.rekenkamer.nl

DATUM 26 februari 2020
BETREFT aankondiging onderzoek Zicht op algoritmes

Geachte heer Gerritsen,

Graag informeer ik u over het onderzoek "Zicht op algoritmes" dat wij recent zijn gestart en uitvoeren bij alle ministeries. Algoritmes hebben een steeds grotere impact op het functioneren en handelen van de overheid en daarmee op burgers en bedrijven. Voor de Algemene Rekenkamer is het daarom belangrijk een scherper inzicht te krijgen in welke typen algoritmes de overheid toepast, voor welke activiteiten, wat hun impact is op de maatschappij en op welke wijze ze het beste kunnen worden getoetst. In lijn met onze strategie, "Inzicht als basis voor vertrouwen", gaan we op onafhankelijke wijze beoordelen hoe algoritmes bij de rijksoverheid in de praktijk functioneren en welke verbeteringen mogelijk zijn.

De volgende vragen staan centraal in dit onderzoek:

1. Bij welke activiteiten en processen worden welke soorten algoritmes toegepast door de rijksoverheid en door de daaraan verbonden organisaties; wat zijn de effecten en risico's hiervan?
2. Hoe is de besturing / governance en kwaliteitsbeheersing van algoritmes bij de rijksoverheid en bij de daaraan verbonden organisaties vormgegeven?
3. In hoeverre worden de risico's van een aantal te selecteren algoritmes beheerst, gelegd langs de meetlat van een toetsingskader?

UW KENMERK
ONS KENMERK 20001205 R/S
BIJLAGEN

We toetsen een aantal algoritmes die in praktijk wordt toegepast en impact heeft op burgers en bedrijven. Dit doen we met een te ontwikkelen toetsingskader dat we op basis van bestaande normen en 'best practices' opstellen. Dit toetsingskader kan een basis leggen om vervolgens breder ingezet worden binnen de rijksoverheid en daarbuiten. Ons onderzoek zal enerzijds een brede inventarisatie kennen bij alle departementen en anderzijds een verdieping in een beperkt aantal concrete organisaties in hun toepassing van algoritmes in processen die impact hebben op burgers en/of bedrijven.

Onze contactpersoon op uw ministerie wordt op de hoogte gesteld van het feit dat wij dit onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in januari 2021 te publiceren.

Het onderzoek valt onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden bij de uitvoering van dit onderzoek staan in meer detail beschreven in artikel 7.18 van deze wet.

De verantwoordelijke directeur is [redacted] Met vragen kunt u terecht bij de projectleider van het onderzoek [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted] Het projectvoorstel voor het onderzoek zal binnenkort aan onze contactpersonen bij het ministerie gestuurd worden waarbij we een nadere mondelinge toelichting zullen aanbieden.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

[redacted]

drs. C. (Cornelis) van der Werf

Project Expl*AI*n

Explaining decisions made with AI

Draft guidance for consultation

Part 2:

Explaining AI in practice

ico.
Information Commissioner's Office

The Alan Turing Institute

About this guidance

What is the purpose of this guidance?

This guidance helps you with the practicalities of explaining AI-assisted decisions and providing explanations to individuals. It shows you how to go about selecting the appropriate explanation for your sector and use case, how to choose an appropriately explainable model, and which tools you can use to extract explanations from less interpretable models.

How should we use this guidance?

This guidance is primarily for technical teams, however DPOs and compliance teams will also find it useful. It goes through the steps you can take to explain AI-assisted decisions to individuals. It starts with how you can choose which explanation type is most relevant for your use case, and what information you should put together for each explanation type. For most of the explanation types, you can derive this information from your organisational governance decisions and documentation.

However, given the central importance of understanding the underlying logic of the AI system for AI-assisted explanations, we provide technical teams with a comprehensive guide to choosing appropriately interpretable models. This depends on the use case. We also indicate how to use supplementary tools to extract elements of the model's workings in 'black box' systems. Finally, we show you how you can deliver your explanation, containing the relevant explanation types you have chosen, in the most useful way for the decision recipient.

What is the status of this guidance?

This guidance is issued in response to the commitment in the Government's AI Sector Deal, but it is not a statutory code of practice under the Data Protection Act 2018.

This is practical guidance that sets out good practice for explaining decisions to individuals that have been made using AI systems processing personal data.

Why is this guidance from the ICO and The Alan Turing Institute?

The ICO is responsible for overseeing data protection in the UK, and The Alan Turing Institute ("The Turing") is the UK's national institute for data science and artificial intelligence.

In October 2017, Professor Dame Wendy Hall and Jérôme Pesenti published their independent review on growing the AI industry in the UK. The second of the report's recommendations to support uptake of AI was for the ICO and The Turing to:

"…develop a framework for explaining processes, services and decisions delivered by AI, to improve transparency and accountability."

In April 2018, the government published its AI Sector Deal. The deal tasked the ICO and The Turing to:

"…work together to develop guidance to assist in explaining AI decisions."

The independent report and the Sector Deal are part of ongoing efforts made by national and international regulators and governments to address the wider implications of transparency and fairness in AI decisions impacting individuals, organisations, and wider society.

Summary of the steps to take

We have set out a number of steps to help you provide explanations of your AI decisions. These offer a systematic approach to selecting, extracting and delivering explanations and they should help in navigating the detailed technical recommendations in this part. However, we recognise that in practice some steps may be concurrent rather than consecutive, and organisations may wish develop their own plan for doing this.

1. **Select priority explanations by considering the domain, use case and impact on the individual**

Start by getting to know the different types of explanation in Part 1 of this guidance. This should help you to separate out the different aspects of an AI-assisted decision that people may want you to explain. While we have identified what we think are the key types of explanation that people will need, there may be additional relevant explanations in the context of your

organisation, and the way you do (or plan to) use AI to make decisions about people. Or perhaps some of the explanations we identify are not of particular relevance to your organisation and the people you make decisions about.

That's absolutely fine. The explanations we identify are intended to underline the fact that there are many different aspects to explanations, and to get you thinking about what those aspects are, and whether or not they are relevant to your customers. You may think the list we have created works for your organisation or you might want to create your own.

Either way, we recommend that your approach to explaining AI-assisted decisions should be informed by the importance of putting the principles of transparency and accountability into practice, and of paying close attention to context and impact.

Next, think about the specifics of the context within which you are deploying your AI decision-support system. The domain you work in, the particular use case and the impact on the person will further help you choose the relevant explanations. In most cases, it will be useful for you to include rationale and responsibility in your priority explanations.

It is likely that you will identify multiple explanations to prioritise for the AI-assisted decisions you make. Make a list of these and document the justification for your choices.

While you have identified the explanations that are most important in the context of your AI decision-support system, this does not mean that the remaining explanations should be discarded.

Choosing what to prioritise is not an exact science, and while your choices may reflect what the majority of the people you make decisions about want to know, it's likely that other individuals will still want and benefit from the explanations you have not prioritised. These will probably also be useful for your own accountability or auditing purposes.

It therefore makes sense that all the explanations you have identified as relevant are made available to the people subject to your AI-assisted decisions. You should consider how to prioritise the remaining explanations based on the contextual factors you identified, and how useful they might be for people.

Speak with colleagues involved in the design/procurement, testing and deployment of AI decision-systems to get their views. If possible, speak with your customers.

2. Collect the information you need for each explanation type

For each explanation, it will be useful for you to gather the information you need for its process-based and outcome-based explanation. The process-based explanation will help you explain how your general decision-making is structured for each explanation, while the outcome-based explanation helps you explain what happened in the case of a specific decision. Create your own list of the types of explanations you have determined are relevant and formalise this in a policy or procedure.

3. Build your rationale explanation to provide meaningful information about the underlying logic of your AI system

It will be useful to understand the inner workings of your AI system, particularly to be able to comply with certain parts of the GDPR. The model you choose should be at the right level of interpretability for your use case and the impact it will have on the decision recipient. If you use a 'black box' model, make sure the supplementary explanation techniques you use provide a reliable and accurate representation of the system's behaviour.

4. Translate the rationale of your system's results into useable and easily understandable reasons

You should determine how you are going to convey your model's statistical results to users and decision recipients as understandable reasons.

A central part of delivering an explanation is communicating how the statistical inferences, which were the basis for your model's output, played a part in your thinking. This involves translating the mathematical rationale of the explanation extraction tools into plain, easily understandable language to justify the outcome.

For example, say your extracted rationale explanation provides you with:

- information about the relative importance of features that influence your model's results; and
- a more global understanding of how this specific decision fits with the model's linear and monotonic constraints.

These factors should then be translated into simple, everyday language that can be understood by non-technical stakeholders. Transforming your model's logic from quantitative rationale into intuitive reasons should lead you to present information as clearly and meaningfully as possible. You could do this through textual clarification, visualisation media, graphical representations, summary tables, or any combination of these.

The main thing is to make sure that there is a simple way to describe or explain the result to an individual. If the decision is fully automated, you may use software to do this. Otherwise this will be through a person who is responsible for translating the result (the implementer – see below).

5. Prepare implementers to deploy your AI system

When human decision-makers are meaningfully involved in an AI-assisted outcome they must be appropriately trained and prepared to use your model's results responsibly and fairly.

Training should include conveying basic knowledge about the nature of machine learning, and about the limitations of AI and automated decision-support technologies. It should also encourage users (the implementers) to view the benefits and risks of deploying these systems in terms of their role in helping humans to come to judgements, rather than replacing that judgement.

If the system is wholly automated and provides a result directly to the decision recipient, it should be set up to provide understandable explanations to them.

6. Consider contextual factors when you deliver your explanation

Consider contextual factors (domain, impact, data, urgency, audience) to help you determine how you should deliver the explanation to the individual.

Again, you may feel that some of the factors we identify (or aspects of them) are simply not relevant to what you do, or that there are additional issues to consider that are unique to the circumstances of your AI model and the decisions it helps you make.

What's important is that you give thought to all the different things that may have an effect on what people will find useful to know about the AI-assisted

decisions you make, and what they might want to do with that knowledge. As a result of this, draw up a list of the relevant factors.

7. Consider how to present your explanation

Finally, you should think about how you will present your explanation of an AI-assisted decision to an individual, whether you are doing this via a website or app, in writing or in person.

A layered approach can be helpful because it presents people with the most relevant information about the decision, while making further explanations easily accessible if they are required. The explanations you have identified as priorities can go in the first layer, while the others can go into a second layer.

You should also think about what information to provide in advance of a decision, and what information to provide to individuals about a decision in their particular case.

Step 1: Select priority explanations by considering the domain, use case and impact on the individual

At a glance

- Getting to know the different types of explanation will help you identify the dimensions of an explanation that decision recipients will find useful.
- In most cases, explaining AI-assisted decisions involves identifying what is happening in your AI system and who is responsible. That means you should prioritise the rationale and responsibility explanation types.
- The setting and sector you are working in is important in figuring out what kinds of explanation you should be able to provide. You should therefore consider domain context and use case.
- In addition, consider the potential impacts of your use of AI to determine which other types of explanation you should provide.
- This will also help you think about how much information is required, and how comprehensive it should be.
- Choosing what to prioritise is not an exact science, and while your choices may reflect what the majority of the people you make decisions about want to know, it's likely that other individuals will still benefit from the explanations you have not prioritised. These will probably also be useful for your own accountability or auditing purposes.

Checklist

☐ We have prioritised rationale and responsibility explanations. We have therefore put in place and documented processes that optimise the end-to-end transparency and accountability of our AI model.

☐ We have considered the setting and sector in which our AI model will be used, and how this affects the types of explanation we provide.

☐ We have considered the potential impacts of our system, and how these affect the scope and depth of the explanation we provide.

In more detail

Introduction

You should consider what types of explanation you need before you start the design process for your AI system, or procurement of a system if you are outsourcing it. You can think of this as 'explanation-by-design'. It involves operationalising the principles we set out in 'The basics of explaining AI'. The following considerations will help you to decide which explanation types you should choose.

Familiarise yourself with the different types of explanation

We introduced the different types of explanation in Part 1 of this guidance, 'The basics of explaining AI'. Making sure you are aware of the range of explanations will provide you with the foundations for considering the different dimensions of an explanation that decision recipients will find useful.

Prioritise rationale and responsibility explanation

It is likely that most explanations of AI-assisted decisions will involve knowing both what your system is doing and who is responsible. In other words, they are likely to involve both rationale and responsibility explanations.

To set up your AI use case to cover these explanations, it is important to consider how you are going to put in place and document processes that:

- optimise the end-to-end transparency and accountability of your AI model. This means making sure your organisation's policies, protocols and procedures are lined up to ensure that when you design and deploy your AI system, you do this in a way that makes it possible to provide clear and accessible process-based explanations; and

- ensure that the intelligibility and interpretability of your AI model is prioritised from the outset. This also means that the explanation you offer to affected individuals appropriately covers the other types of explanation, given the use case and possible impacts of your system.

Consider domain or sector context and use case

When you are trying to work out what kinds of explanation you provide, a good starting point is to consider the setting and sector in which it will be used.

In certain safety-critical/high-stakes and highly regulated domains, sector-specific standards for explanations may largely dictate the sort of information you need to provide to affected individuals.

For instance, AI applications that are employed in safety-critical domains like medicine will have to be set up to provide the safety and performance explanation in line with the established standards and expectations of that sector. Likewise, in a high-stakes setting like criminal justice, where biased decision-making is a significant concern, the fairness explanation will play an important and necessary role.

Understanding your AI application's domain context and setting may also give you useful information about public expectations regarding the content and scope of explanations that have been previously offered in relevant decisions. Doing due diligence and researching these sorts of sector-specific expectations will help you to draw on background knowledge as you weigh which types of AI explanation to include as part of your model's design and implementation processes.

Consider potential impacts

Paying attention to the setting in which your model will be deployed will also put you in a good position to consider its potential impacts. This will be especially useful for selecting your explanations, because it will key you in to

the relevance of impact-specific explanations that should be included as part of your more general explanation of your AI system.

Assessing the potential impact of your AI model on the basis of its use case will help you to determine the extent to which you need to include fairness, safety and performance and more general impact explanations, together with the scope and depth of these types of explanation.

Assessing your AI model's potential impact will also help you understand how comprehensive your explanation needs to be. This includes the risks of deploying the system, and the risks for the person receiving the AI-assisted decision. It will allow you to make sure that the scope and depth of the explanations you are going to be able to offer line up with the real-world impacts of the specific case. For example, an AI system that triages customer service complainers in a luxury goods retailer will have a different (and much lower) explanatory burden than one that triages patients in a hospital critical care unit.

Once you have worked through these considerations, you should choose the most appropriate explanations for your use case (in addition to the rationale and responsibility explanations you have already prioritised). You should document these choices and why you made them.

Prioritise remaining explanations

Once you have identified the other explanations that are relevant to your use case, you should make these available to the people subject to your AI-assisted decisions. You should also document why you made these choices.

See more on the types of explanation in the link below for 'The basics of explaining AI' and the information you need to put together for each one in Step 2.

> The basics of explaining AI

Examples for choosing suitable explanation types

AI-assisted recruitment

An AI system is deployed as a job application filtering tool for a company that is looking to hire someone for a vacancy. This system classifies decision recipients (who receive either a rejection or an invitation to interview) by processing social or demographic data related to individual human attributes and social patterns that are implied in the CVs that have been submitted. A resulting concern might be that bias is 'baked into' the dataset, and that discriminatory features or their proxies might have been used in the model's training and processing. For example, the strong correlation in a dataset between 'all-male' secondary schools attended and successful executive placement in higher paying positions might lead a model trained on this data to discriminate against non-male applicants when it renders recommendations about granting job interviews related to positions of a certain higher paying and executive-level profile.
Which explanation types should you choose in this case?

- **Prioritise rationale and responsibility explanations**: it is highly likely that you will need to include the responsibility and rationale explanations, to tell the individual affected by the AI-assisted hiring decision who is responsible for the decision, and why the decision was reached.
- **Consider domain or sector context and use case**: the recruitment and human resources domain context suggests that bias should be a primary concern in this case.
- **Consider potential impacts**: considering the impact of the AI system on the applicant relates to whether they think the decision was justified, and whether they were treated fairly. Your explanation should be comprehensive enough for the applicant to understand the risks involved in your use of the AI system, and how you have mitigated these risks.
- **Prioritise other explanation types**: This example demonstrates how understanding the specific area of use (the domain) and the particular nature of the data is important for knowing which type of explanation is required for the decision recipient. In this case, a fairness explanation is required because the decision recipient wants to know that they have not been discriminated against. This discrimination could be due to the legacies of discrimination and historical patterns of inequity that may have influenced an AI system trained on biased social and demographic data. In addition, the individual may want an impact explanation to understand how

the recruiter thought about the AI tool's impact on the individual whose data it was processing. A data explanation might also be helpful to understand what data was used to determine whether the candidate would be invited to interview.

AI-assisted medical diagnosis

An AI system utilises image recognition algorithms to support a radiologist to identify cancer in scans. It is trained on a dataset containing millions of images from patient MRI scans and learns by processing billions of corresponding pixels. It is possible that the system may fail unexpectedly when confronted with unfamiliar data patterns or unforeseen environmental anomalies (objects it does not recognise). Such a system failure might lead to catastrophic physical harm being done to an affected patient.

Which explanation types should you choose in this case?

- **Prioritise rationale and responsibility explanations**: it is highly likely that you will need to include the responsibility and rationale explanations, to tell the individual affected by the AI-assisted diagnostic decision who is responsible for the decision, and why the decision was reached.
- **Consider domain or sector context and use case**: the medical domain context suggests that demonstrating the safety and optimum performance of the AI system should be a primary concern in this case.
- **Consider potential impacts**: the impact of the AI system on the patient is high if the system makes an incorrect diagnosis. Your explanation should be comprehensive enough for the patient to understand the risks involved in your use of the AI system, and how you have mitigated these risks.
- **Prioritise other explanation types**: The safety and performance explanation provides justification, when possible, that an AI system is sufficiently robust, accurate, secure and reliable, and that codified procedures of testing and validation have been able to certify these attributes.

Step 2: Collect the information you need for each explanation type

At a glance

- For each type of explanation you should provide:
    - process-based explanations which give you information on the governance of your AI system across its design and deployment; and
    - outcome-based explanations which tell you what happened in the case of a particular decision.
- Your rationale explanation should cover how the system performed and turned inputs into outputs, as well as how the outputs are translated into understandable reasons.
- Your responsibility explanations should identify who is responsible at each stage of the design and deployment of your AI system.
- The data explanation should outline what data was used and why, as well as where it came from.
- Your fairness explanation should reflect that:
    - you made sure the AI system was trained and tested on representative, relevant, accurate and generalisable datasets;
    - you can justify how you built the model architecture;
    - the system does not have a discriminatory effect on those affected by the decision; and
    - the system is deployed by users who are trained to implement it responsibly.
- Safety and performance explanations should cover how you have guaranteed the accuracy, reliability, security and robustness of your system.
- Finally, impact explanations should show how you have considered the impact your AI system has on the individuals affected, as well as wider society.
- The data that you collect and pre-process before inputting it into your system also has an important role to play in the ability to derive each explanation type.

Checklist

☐ We have identified the people within our organisation that are responsible for providing explanations and what exactly they are responsible for.

☐ Our policies, protocols and procedures make it possible to provide clear and accessible process-based explanations when we design and deploy our AI system.

☐ We have considered the setting and sector in which our AI system will be used.

☐ We have considered the potential impacts of our AI system.

☐ We have thought about which other explanation types to include, as well as the depth of the information we will provide in the explanation.

☐ We have documented the information required for process-based and outcome-based explanations for each explanation type.

How collecting and pre-processing the data impacts explanation:

☐ Our data is representative of those we will make decisions about, reliable, relevant and up-to-date.

☐ We have checked with a domain expert to ensure that the data we are using is appropriate and adequate.

☐ We know where the data has come from, the purpose it was originally collected for, and how it was collected.

☐ Where we are using synthetic data, we know how it was created and what properties it has.

☐ We know what the risks are of using the data we have chosen to use, as well as the risks to data subjects of having their data included.

☐ We have labelled the data we are using in our AI system with information including what it is, where it is from, and the reasons why we have included it.

☐ Where we are using unstructured or high-dimensional data, we are clear about why we are doing this and the impact of this on explainability.

☐ We have ensured as far as possible that the data does not reflect past discrimination, whether based explicitly on protected characteristics or possible proxies.

☐ We have mitigated possible bias through pre-processing techniques such as re-weighting, up-weighting, masking, or excluding features and their proxies.

☐ It is clear who within our organisation is responsible for data collection and pre-processing.

In more detail

Building the different explanations

The main aim of explaining fully automated or AI-assisted decisions is justifying a particular result to the individual whose interests are affected by it. In this part, that means making the reasoning behind the outcome of that decision clear, and demonstrating how you were responsible when you chose the processes to design and deploy the system that led to the decision.

We have therefore divided each type of explanation into the subcategories of 'process' and 'outcome':

- **Process-based explanations** of AI systems are about demonstrating that you have followed good governance processes and best practices throughout your design and use.

  For example, if you are trying to explain the fairness and safety of a particular AI-assisted decision, one component of your explanation will involve establishing that you have taken adequate measures across the system's production and deployment to ensure that its outcome is fair and safe.
- **Outcome-based explanations** of AI systems are about clarifying the results of a specific decision. They involve explaining the reasoning behind a particular algorithmically-generated result in plain, easily understandable, and everyday language.

  If there is meaningful human involvement in the decision-making process, you also have to make clear to the affected individual how and why a human judgement that is assisted by an AI output was reached.

  In addition, you may also need to confirm that the actual outcome of an AI decision meets the criteria that you established in your design process to ensure that the AI system is being used in a fair, safe, and ethical way.

The list of explanations below helps you put together the information you will need to be able to build the different explanations.

While we include the rationale explanation here, due to its central importance in AI explanations, we go into further detail in the following sections about how to derive it from a technical perspective. The rationale explanation helps you understand the underlying logic of your AI system, and helps you comply with Articles 13, 14 and 15 of the GDPR.

Rationale explanation

What you need to show

- How the system performed and behaved to get to that decision outcome.
- How the different components in the AI system led it to transform inputs into outputs in a particular way, so you can communicate which features, interactions, and parameters were most significant.

- How these technical components of the logic underlying the result can provide supporting evidence for the decision reached.
- How this underlying logic can be conveyed as easily understandable reasons to decision recipients.
- How you have thought about their impacts on the lives of affected individuals and society.

What information goes into this explanation

- Process-based explanation:
    - Explain how the procedures you have set up help you provide meaningful explanations of the underlying logic of your AI model's results.
    - Ensure that these are appropriate given the model's particular domain context and its possible impacts on the affected decision recipients and wider society.
    - Demonstrate that you have thought about how you are going to set up your AI system and its data collection and pre-processing, model selection, explanation extraction, and explanation delivery procedures so that your system is appropriately interpretable and explainable.

This explanation might answer:

    - Have we selected an algorithmic model, or set of models, that will provide a degree of interpretability that corresponds with its impact on affected individuals?
    - Are the supplementary explanation tools that we are using to help make our complex system explainable good enough to provide meaningful and accurate information about its underlying logic?

- Outcome-based explanation:

    - Explain the formal and logical rationale of the AI system − how the AI system is verified against its formal specifications, so you can verify that the AI system will operate reliably and behave correctly.
    - Explain the technical rationale of the AI system or its output − how the AI model's components (its variables, rules and procedures) transform inputs into outputs, so you know what role these components play in producing the AI system's

output. By understanding the roles and functions of the individual components of the AI system, it is possible to identify the features and parameters that most influence a particular output/decision.

- Explain the translation of the AI system's workings – transforming its input and output variables, parameters and so on into accessible everyday language, so that it becomes clear what role these factors play in reasoning about the real-world problem that the model is trying to address or solve.
- Explain the application of the statistical result to the individual concerned – an application of the reasoning behind the result which takes into account the uniqueness of the specific circumstances, background and personal qualities of affected individuals.
- Explain the justification of the impacts of the use of the AI system, so that you remain accountable both to the individuals about whom you make decisions and to their communities.

The GDPR also makes reference to providing meaningful information about the logic involved in automated decision-making under Articles 13, 14 and 15.

In order to be able to derive your rationale explanation, you need to know how your algorithm works. See Step 3 for more detail about how to do this.

Responsibility explanation

What you need to show

- Identify who is accountable at each stage of the AI system's design and deployment, from defining outcomes for the system at its initial phase of design, through to providing the explanation to the affected individual at the end.
- Define the mechanisms by which each of these people will be held accountable, as well as how you have made the design and implementation processes of your AI system traceable and auditable.

What information goes into this explanation

- Process-based explanation:
  - Detail the roles and functions across your organisation that are involved in the various stages of developing and implementing

your AI decision system, including any human involvement in the decision-making.
    - Explain broadly what the roles do, why they are important, and where overall responsibility lies for management of the AI model – who is ultimately accountable.
    - Explain who is responsible at each step from the design of an AI system through to its implementation to make sure that there is effective accountability throughout.

- Outcome-based explanation:

    - Cover information on how to request a human review of an AI-enabled decision or object to the use of AI, including details on who to contact, and what the next steps will be (eg how long it will take, what the human reviewer will take into account, how they will present their own decision and explanation).
    - Provide a way for individuals to directly contact the role or team responsible for the review. You do not need to identify a specific person in your organisation. One person involved in this should be someone who implemented the decision, and who used the statistical results of a decision-support system to come to a determination about an individual.

Data explanation

What you need to show

- What data was used when you trained your AI system.
- What data you used in a particular decision.

What information goes into this explanation

- Process-based explanation:
    - Detail the source of the training/ test data.
    - Explain how you boost 'explainability' (eg labelling), assess and improve its quality.
    - Explain how you ensure the data is representative.
    - Explain how you ensure bias and discrimination have been mitigated.

- Outcome-based explanation:

    - Clarify the input data used for a specific decision, and the sources of that data.
    - Document the handling and preparation of training and test data, so that a clear and meaningful picture of data handling and use can be provided to affected individuals and other relevant parties.

Fairness explanation

What you need to show

- Dataset fairness: It is trained and tested on properly representative, relevant, accurate, and generalisable datasets.

How?

    - Make sure your data sample is representative of all those affected.
    - Ensure your data is sufficient in terms of its quantity and quality, so it represents the underlying population and the phenomenon you are modelling.
    - Ensure your data is assessed and recorded through suitable, reliable and impartial sources of measurement and has been sourced through sound collection methods.
    - Make sure your data is up-to-date and accurately reflects the characteristics of individuals, populations and the phenomena you are trying to model.
    - Make sure your data is relevant by calling on domain experts to help you understand, assess and use the most appropriate sources and types of data to serve your objectives.

- Design fairness: It has model architectures that do not include target variables, features, processes, or analytical structures (correlations, interactions, and inferences) which are unreasonable or unjustifiable.

How?

    - When defining the problem at the start of the AI project, identify how structural biases can play a factor in translating your objectives into target variables and measurable proxies. These biases could also influence what system designers expect

target variables to measure and what they statistically represent.

- In data pre-processing, take into account the sector or organisational context in which you are operating, as you may introduce bias into your classification process. When this process is automated or outsourced, review what has been done, maintain oversight, and use certification. You should also attach information on the context and metadata to the datasets, so that those coming to the pre-processed data later on have access to the relevant properties when they undertake bias mitigation.
- When you determine which features are relevant as input variables for your model, be aware that the choices you make about grouping or separating and including or excluding features, as well as more general judgements about the comprehensiveness or coarseness of the total set of features, may have consequences for protected groups of people.
- Bias can come in when tuning parameters and setting metrics at the modelling, testing and evaluation stages – ie into the trained model. Your AI development team should iterate the model and peer review it to help ensure that how they choose to adjust the dials and metrics of the model are in line with your objectives of mitigating bias.
- Look out for hidden proxies for discriminatory features in your trained model, as these may act as influences on your model's output. Designers should also look into whether the significant correlations and inferences determined by the model's learning mechanisms are justifiable.

- Outcome fairness: It does not have discriminatory or inequitable impacts on the lives of the people they affect.

How?

- This depends on the definitions of fairness you choose. For example, data scientists can apply different formalised fairness criteria to choose how specific groups in a selected set will receive benefits in comparison to others in the same set, or how the accuracy or precision of the model will be distributed among subgroups. This can be done by reweighting model parameters; embedding trade-offs in a classification procedure; or re-tooling algorithmic results to adjust for outcome preferences.

- Implementation fairness: It is deployed by users sufficiently trained to implement it responsibly and without bias.

How?

  - To avoid automation bias (over-relying on the outputs of AI systems) or automation-distrust bias (under-relying on AI system outputs because of a lack of trust in them) you should train implementers of AI system outputs on how to use them in the specific context in which they are being used. That is, they should understand the individual circumstances of the individual to which that output is being applied.

What information goes into this explanation

This explanation is about providing people with appropriately simplified and concise information on the considerations, measures and testing you carry out. Fairness considerations come into play through the whole lifecycle of an AI model, from inception to deployment, monitoring and review.

- Process-based explanation:
  - Detail your chosen measures to mitigate risks of bias and discrimination at the data collection, preparation, model design and testing stages.
  - Detail the results of your initial (and ongoing) fairness testing and external validation – proving that your chosen fairness measures are working in practice. You could do this by showing that different groups of people receive similar outcomes, or that protected characteristics have not played a factor in the results.

- Outcome-based explanation:
  - Explain how your organisation has decided to define fairness by the criteria it has selected in its formal model(s). It should then be possible to explain how these fairness criteria were implemented in the case of a particular decision or output.
  - Include the relevant fairness metrics and performance measurements in the delivery interface of your model.
  - Explain how others similar to the individual were treated, ie whether they received the same decision outcome as the individual. For example, you could use information generated from counter-factual scenarios to show whether or not someone with similar characteristics, but of a different ethnicity or

gender, would receive the same decision outcome as the individual.

Safety and performance explanation

What you need to show

- Accuracy: the proportion of examples for which your model generates a correct output. This component may also include other related performance measures such as precision, sensitivity (true positives), and specificity (true negatives). Individuals may want to understand how accurate, precise, and sensitive the output was in their particular case.
- Reliability: how dependably the AI system does what it was intended to do. If it did not do what it was programmed to carry out, individuals may want to know why, and whether this happened in the process of producing the decision that affected them.
- Security: the system is able to protect its architecture from unauthorised modification or damage of any of its component parts; the system remains continuously functional and accessible to its authorised users and keeps confidential and private information secure, even under hostile or adversarial conditions.
- Robustness: the system functions reliably and accurately under harsh conditions. Individuals may want to know how well the system works when things go wrong, how this has been anticipated and tested, and how the system has been immunised from adversarial attacks.

What information goes into this explanation

- Process-based explanation:
    - Accuracy: how you measure it and why you chose those measures, eg maximising precision to reduce the risk of false negatives; what you did at the data collection stage to ensure your training data was up-to-date and reflective of the characteristics of the people you are now making AI-assisted decisions about; what kinds of external validation you have undertaken to test and confirm your model's accuracy; what the overall accuracy rate of the system was at testing stage, and what you do to monitor this (eg measuring for concept drift over time).
    - Reliability: how you measure it and why you chose those measures, which helps the individual to understand how

confident you are in the system's consistency and therefore its safety.

- Security: how you measure it and why you chose those measures, eg who is able to access the AI system; how you manage the security of confidential and private information.
- Robustness: how you measure it and why you chose those measures, eg how you've stress-tested the system to understand how it responds to adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications).
- Summarise the type of AI model(s) used (eg decision tree, random forest, neural network), the AI software, software development kit, or programme used (eg TensorFlow, scikit-learn, H2O), and the technical approach you use to extract rationale explanations from your model (eg sensitivity analysis, SHAP, LIME).

- Outcome-based explanation:

It is unlikely (or even, in some cases, impossible) for you to be able to conduct testing on the accuracy of your AI model's predictions or classifications at the individual level (eg for particular decisions).

- In the case of accuracy and the other performance metrics, however, you should include in your model's delivery interface the results of your cross-validation (training/testing splits) and any external validation carried out.
- You may also include relevant information related to your system's confusion matrix (the table that provides the range of performance metrics) and ROC curve (receiver operating characteristics)/AUC (area under the curve) – with guidance for users and affected individuals that makes the meaning of these measurement methods, and specifically the ones you have chosen to use, easily accessible and understandable. This should also include a clear representation of the uncertainty of the results (eg confidence intervals and error bars).
- Provide information for components other than accuracy that confirms that the AI system operated securely, reliably, and in accordance with its intended design in the case of a specific decision.

Impact explanation

What you need to show

- Demonstrate that you have thought about how your AI system will potentially affect individuals and wider society, and make the process you have gone through to determine these possible impacts plain to affected individuals.

What information goes into this explanation

- Process-based explanation:
    - Summarise the considerations you made, how you made them, and the measures and steps you took to mitigate possible negative effects on society, and to amplify the positive effects.
    - Include information about how you plan to monitor and re-assess impacts while your system is deployed.
- Outcome-based explanation:
    - Explain the intention and purpose behind the AI model − you should say what the system is being used to help make decisions about and what you were optimising for when designing and developing it.
    - Explain the consequences for the individual of the different possible decision outcomes, eg if the decision has favourable and unfavourable outcomes, what will this mean for the individual, and what happens next for them?
    - Explain the impacts on the wellbeing of wider society − be explicit about the considerations you have made regarding the effect of your system on both communities in which it is being deployed and society as whole.

How collecting and pre-processing data impacts the explanation

How you collect and pre-process the data you use in your chosen model has a bearing on the quality of the explanation you can offer to decision recipients. Below we set out some of the things you should think about, and how this can contribute to the information you provide to individuals for each explanation type.

Rationale

Understanding the logic of an AI model, or of a specific AI-assisted decision, is much simpler when the features (the input variables from which the model draws inferences and that influence a decision) are already interpretable by humans, for example, someone's age or location. Limit your pre-processing of that data so that it isn't transformed through extensive feature engineering into more abstract features that are difficult for humans to understand.

Careful, transparent, and well-informed data labelling practices will set up your AI model to be maximally interpretable. If you are using data that is not already naturally labelled, there will be a stage at which you will have humans labelling the data with relevant information. At this stage you should ensure that the information recorded is as rich and meaningful as possible. Ask those charged with labelling data to not only tag and annotate what a piece of data is, but also the reasons for that tag. For example, rather than 'this x-ray contains a tumour', say 'this x-ray contains a tumour because…'. Then, when your AI system classifies new x-ray images as tumours, you will be able to look back to the labelling of the most similar examples from the training data to contribute towards your explanation of the decision rationale.

Of course, all of the above isn't always possible. The domain in which you wish to use AI systems may require the collection and use of unstructured, high-dimensional data (where there are countless different input variables interacting with each other in complex ways).

In these cases, you should justify and document the need to use such data. You should also use the guidance in the next step to assess how best to obtain an explanation of the rationale through appropriate model selection and approaches to explanation extraction.

Responsibility

Responsibility explanations are about telling people who, or which part of your organisation, is responsible for overall management of the AI model. This is primarily to make your organisation more accountable to the individuals it makes AI-assisted decisions about.

But you may also want to use this as an opportunity to be more transparent with people about which parts of your organisation are responsible for each stage of the development and deployment process, including data collection and preparation.

Of course, it may not be feasible for your customers to have direct contact with these parts of your organisation (depending on your organisation's size and how you interact with customers). But informing people about the different business functions involved will make them more informed about the process. This may increase their trust and confidence in your use of AI-assisted decisions because you are being open and informative about the whole process.

If you are adopting a layered approach to the delivery of explanations, it is likely that this information will sit more comfortably in the second or third layer − where interested individuals can access it, without overloading others with too much information. See Step 7 for more on layering explanations.

Data

The data explanation is, in part, a catch-all for giving people information about the data used to train your AI model.

There is a lot of overlap therefore with information you may already have included about data collection and preparation in your rationale, fairness and safety and performance explanations.

However, there are other aspects of the data collection and preparation stage, which you could also include. For example:

- the source of the training data;
- how it was collected;
- assessments about its quality; and
- steps taken to address quality issues, such as completing or removing data.

While these may be more procedural aspects (less directly linked to key areas of interest such as fairness and accuracy) there remains value in providing this information to people. As with the responsibility explanation, the more insight individuals have on the AI model that makes decisions about them, the more confident they are likely to be in interacting with these systems and trusting your use of them.

Fairness

Fairness explanations are about giving people information on the steps taken to mitigate risks of discrimination both in the production and implementation of your AI system and in the results it generates. They shed light on how individuals have been treated in comparison to others. Some of the most

important steps to mitigate discrimination and bias arise at the data collection stage.

For example, when you collect data, you should have a domain expert to assess whether it is sufficiently representative of the people you will make AI-assisted decisions about.

You should also consider where the data came from, and assess to what extent it reflects past discrimination, whether based explicitly on protected characteristics such as race, or on possible proxies such as post code. You may need to modify the data to avoid your AI model learning and entrenching this bias in its decisions. Pre-processing techniques such as re-weighting, up-weighting, masking, or even excluding features may be used to mitigate implicit discrimination in the dataset and to prevent bias from entering into the training process. If you exclude features, you should also ensure that you exclude proxies or related features.

Considerations and actions such as these, that you take at the data collection and preparation stages, should feed directly into the fairness explanations you give to individuals. Ensure that you appropriately document what you do at these early stages so you can reflect this in your explanation.

Safety and performance

The safety and performance explanation is concerned with the actions and measures you take to ensure that your AI system is accurate, secure, reliable and robust.

The accuracy component of this type of explanation is mainly concerned with the actions and measures you take at the modelling, testing, and monitoring stages of developing an AI model. It involves providing people with information about the accuracy rate of a model, and about the various accuracy related measures you used.

Impact

The impact explanation involves telling people about how an AI model, and the decisions it makes, may impact them as individuals, communities, and members of wider society. It involves making decision recipients aware of what the possible positive and negative effects of an AI model's outcomes are for people taken singly and as a whole. It also involves demonstrating that your organisation has put appropriate forethought into mitigating any potential harm and pursuing any potential societal benefits.

Information on this will come from considerations you made as part of your impact or risk assessment (eg a data protection impact assessment). But it will also come from the practical measures you took throughout the development and deployment of the AI model to act on the outcome of the impact assessment.

This includes what you do at the data collection and preparation stage to mitigate risks of negative impact and amplify the possibility of positive impact on society.

While you may have covered such steps in your fairness and accuracy explanations (eg ensuring the collection of representative and up-to-date datasets), the impact explanation type is a good opportunity to clarify in simple terms how this affects the impact on society (eg by reducing the likelihood of systematic disadvantaging of minority groups, or improving the consistency of decision-making for all groups).

> For an introduction to the explanation types, see 'The basics of explaining AI'. For further details on how to take measures to ensure these kinds of fairness in practice and across your AI system's design and deployment, see the fairness section of Understanding Artificial Intelligence Ethics and Safety, a guidance produced by the Office for AI, the Government Digital Service, and The Alan Turing Institute.

Step 3: Build your rationale explanation to provide meaningful information about the underlying logic of your AI system

At a glance

- Deriving the rationale explanation is key to understanding your AI system and helps you comply with parts of the GDPR. It requires looking 'under the hood' and helps you gather information you need for some of the other explanations, such as safety and performance and fairness. However, this is a complex task that requires you to know when to use more and less interpretable models and how to understand their outputs.
- To choose the right AI model for your explanation needs, you should think about the domain you are working in, and the potential impact of the deployment of your system on individuals and society.
- Following this, you should consider:
    - the costs and benefits of replacing your current system with a newer and potentially less explainable AI model;
    - whether the data you use requires a more or less explainable system;
    - whether your use case and domain context encourage choosing an inherently interpretable system;
    - if your processing needs lead you to select a 'black box' model; and
    - whether your supplementary interpretability tools are appropriate in your context.
- To extract explanations from inherently interpretable models, look at the logic of the model's mapping function by exploring it and its results directly.
- To extract explanations from 'black box' systems, there are many techniques you can use. Make sure that they provide a reliable and accurate representation of the system's behaviour.

Checklist

Selecting an appropriately explainable model:

☐ We know what the interpretability/transparency expectations and requirements are in our sector or domain.

☐ In choosing our AI model, we have taken into account the specific type of application and the impact of the model on decision recipients.

☐ We have considered the costs and benefits of replacing the existing technology we use with an AI system.

☐ Where we are using social or demographic data, we have considered the need to choose a more interpretable model.

☐ Where we are using biophysical data, for example in a healthcare setting, we have weighed the benefits and risks of using opaque or less interpretable models.

☐ Where we are using a 'black box' system, we have considered the risks and potential impacts of using them.

☐ Where we are using a 'black box' system we have also determined that the case we will use it for and our organisational capacity both support the responsible design and implementation of these systems.

☐ Where we are using a 'black box' system we have considered which supplementary interpretability tools are appropriate for our use case.

☐ Where we are using 'challenger' models alongside more interpretable models, we have established that we are using them lawfully and responsibly, and we have justified why we are using them.

☐ We have considered how to measure the performance of the model and how best to communicate those measures to implementers, and decision recipients.

☐ We have mitigated any bias we have found in the model.

☐ We have made it clear how the model has been tested, including which parts of the data have been used to train the model, and which have been used to test it, and which have formed the holdout data.

☐ We have a record of each time the model is updated, how each version has changed, and how this affects the model's outputs.

☐ It is clear who within our organisation is responsible for validating the explainability of our AI system.

Tools for extracting a rationale explanation:

All the explanation extraction tools we use:

☐ Convey the model's results reliably and clearly.

☐ Help implementers of AI-assisted decisions to exercise better-informed judgements.

☐ Offer affected individuals plausible and easily understandable accounts of the logic behind the model's output.

For interpretable AI models:

☐ We are confident in our ability to extract easily understandable explanations from models such as regression-based and decision/rule-based systems, Naïve Bayes, and K nearest neighbour.

For supplementary explanation tools to interpret 'black box' AI models:

☐ We are confident that they are suitable for our application.

☐ We recognise that they will not give us a full picture of the opaque model.

☐ In selecting the supplementary tool, we have prioritised the need for it to provide a reliable, accurate and close approximation of the logic behind our AI system's behaviour, for both local and global

explanations.

Combining supplementary explanation tools to produce meaningful information about your AI system's results:

☐ We have included a visualisation of how the model works.

☐ We have included an explanation of variable importance and interaction effects, both global and local.

☐ We have included counterfactual tools to explore alternative possibilities and actionable recourse.

In more detail

Introduction

The rationale explanation is key to understanding your AI system and helps you comply with parts of the GDPR. It requires detailed consideration because it is about how the AI system works, and can help you obtain an explanation for the underlying logic of the AI model you decide to use.

Selecting an appropriately explainable model

Where do we start?

Before you consider the technical factors, you should consider:

Domain: Consider the specific standards, conventions, and requirements of the domain in which your AI system will be applied.

For example, in the financial services sector, rigorous justification standards for credit and loan decisions largely dictate the need to use fully transparent and easily understandable AI decision-support systems. Likewise, in the medical sector, rigorous safety standards largely dictate the extensive levels of performance testing, validation and assurance that are demanded of

treatments and decision-support tools. Such domain specific factors should actively inform the choices you make about model complexity and interpretability.

Impact: Think about the type of application you are building and its potential impacts on affected individuals.

For example, there is a big difference between a computer vision system that sorts handwritten employee feedback forms and one that sorts safety risks at a security checkpoint. Likewise, there is also a difference between a complex random forest model that triages applicants at a licensing agency and one that triages sick patients in an accident and emergency department.

Higher-stakes or safety-critical applications will require you to be more thorough in how you consider whether prospective models can appropriately ensure outcomes that are non-discriminatory, safe, and supportive of individual and societal wellbeing.

Low-stakes AI models that are not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data are likely to mean there is less need for you to dedicate extensive resources to developing an optimally performing but highly interpretable system.

Draw on the appropriate domain knowledge, policy expertise and managerial vision in your organisation. You need to consider these when your team is looking for the best-performing model.

What are the technical factors that we should consider when selecting a model?

You should also discuss a set of more technical considerations with your team before you select a model.

Existing technologies: consider the costs and benefits of replacing current data analysis systems with newer systems that are possibly more resource-intensive and less explainable.

One of the purposes of using an AI system might be to replace an existing algorithmic technology that may not offer the same performance level as the more advanced machine learning techniques that you are planning to deploy.

In this case, you may want to carry out an assessment of the performance and interpretability levels of your existing technology. This will provide you with a baseline against which you can compare the trade-offs of using a more advanced AI system. This could also help you weigh the costs and benefits of building or using a more complex system that requires more support for it to be interpretable, in comparison to the costs and benefits of using a simpler model.

It might also be helpful to look into which AI systems are being used in your application area and domain. This should help you to understand the resource demands that building a complex but appropriately interpretable system will place on your organisation.

> For more information on the trade-offs involved in using AI systems, see the ICO's AI Auditability Framework blogpost on trade-offs.

Data: integrate a comprehensive understanding of what kinds of data you are processing into considerations about the viability of algorithmic techniques.

To select an appropriately explainable model, you need to consider what kind of data you are processing and what you are processing it for.

There are two groups of data that it is helpful to consider:

i. Data that refers to demographic characteristics, measurements of human behaviour, social and cultural characteristics of people.
ii. Biological or physical data, such as biomedical data used for research and diagnostics (ie data that does not refer to demographic characteristics or measurements of human behaviour);

With these in mind, there are certain things to consider:

- In cases where social or demographic data (group i. above) is being processed you may come across issues of bias and discrimination. Here, you should prioritise selecting an optimally interpretable model, and avoid 'black box' systems.
- More complex systems may be appropriate in cases where biological or physical data (group ii. above) is being processed, only for the purposes of gaining scientific insight (eg radiological diagnostics), or

operational functionality (eg computer vision for vehicle navigation). However, where the application is high impact or safety-critical, you should weigh the safety and performance (accuracy, security, reliability and robustness) of the AI system heavily in selecting the model. Note, though, that bias and discrimination issues may arise in processing biological and physical data, for example in the representativeness of the datasets on which these models are trained and tested.

- In cases where both these groups of data are being processed and the processing directly affects individuals, you should consider concerns about both bias and safety and performance when you are selecting your model.

Another distinction you should consider is between conventional data (eg a person's payment history or length of employment at a given job) and unconventional data (eg sensor data – whether raw or interlinked with other data to generate inferences – collected from a mobile phone's gyroscope, accelerometer, battery monitor, or geolocation device or text data collected from social media activity).

In cases where unconventional data is being used to support decisions that affect individuals, you should bear the following in mind:

- This data can be considered to be of the same type as group i. data above, and treated the same way (as it gives rise to the same issues).
- You should select transparent and explainable AI systems that yield interpretable results, rather than black box models.
- You can justify its use by indicating what attribute the unconventional data represents in its metadata, and how such an attribute might be a factor in evidence-based reasoning.

  For example, if GPS location data is included in a system that analyses credit risk, the metadata must indicate what interpretively significant feature such data is supposed to indicate about the individual whose data is being processed.

Interpretable algorithms: when possible and application-appropriate, draw on standard and maximally interpretable algorithmic techniques.

In high impact, safety-critical or other potentially sensitive environments, you are likely to need an AI system that maximises accountability and

transparency. In some cases, this will mean you prioritise choosing standard but sophisticated non-opaque techniques.

These techniques (some of which are outlined in the table below) may include decision trees/rule lists, linear regression and its extensions like generalised additive models, case-based reasoning, or logistic regression. In many cases, reaching for the 'black box' model first may not be appropriate and may even lead to inefficiencies in project development. This is because more interpretable models are also available, which perform very well but do not require supplemental tools and techniques for facilitating interpretable outcomes.

Careful data pre-processing and iterative model development can hone the accuracy of interpretable systems in ways that may make the advantages gained by the combination of their performance and transparency outweigh the benefits of less transparent approaches.

'Black box' AI systems: when considering the use of opaque algorithmic techniques, make sure that the supplementary interpretability tools that will be used to explain the model are appropriate to meet the domain-specific risks and explanatory needs that may arise from deploying it.

For certain data processing activities it may not be feasible to use straightforwardly interpretable AI systems. For example, the most effective machine learning approaches will likely be opaque when AI applications are sought for classifying images, recognising speech, or detecting anomalies in video footage. The feature spaces of these kinds of AI systems grow exponentially to hundreds of thousands or even millions of dimensions. At this scale of complexity, conventional methods of interpretation no longer apply.

For clarity, we define a 'black box' model as any AI system whose inner workings and rationale are opaque or inaccessible to human understanding. These systems may include neural networks (including recurrent and convolutional neural nets), ensemble methods (an algorithmic technique such as the random forest method that strengthens an overall prediction by combining and aggregating the results of several or many different base models), and support vector machines (a classifier that uses a special type of mapping function to build a divider between two sets of features in a high dimensional space). The main kinds of opaque models are described in more detail below.

You should only use 'black box' models if you have thoroughly considered their potential impacts and risks in advance, and the members of your team have determined that your use case and your organisational capacities/resources support the responsible design and implementation of these systems.

Likewise, you should only use them if supplemental interpretability tools provide your system with a domain-appropriate level of explainability that is reasonably sufficient to mitigate its potential risks and to provide a solid basis for providing affected decision recipients with meaningful information about the rationale of any given outcome. A range of the supplementary techniques and tools that assist in providing some access to the underlying logic of 'black box' models is explored below.

As part of the process-based aspect of the rationale explanation of your AI system, you should document and keep a record of any deliberations that go into your organisation's selection of a 'black box' model.

Hybrid methods – use of 'challenger' models: in cases where an interpretable model is selected to ensure explainable data processing, you should carry out parallel use of opaque 'challenger' models for purposes of feature engineering/selection, insight, or comparison in a transparent, responsible, and lawful manner.

Our research has shown that, while some organisations in highly regulated areas like banking and insurance are continuing to select interpretable models in their customer-facing AI decision-support applications, they are increasingly using more opaque 'challenger' models alongside these, for the purposes of feature engineering/selection, comparison, and insight.

'Black box' challenger models are trained on the same data that trains transparent production models and are used both to benchmark the latter, and in feature engineering and selection.

When challenger models are employed to craft the feature space, ie to reduce the number of variables (feature selection) or to transform/combine/bucket variables (feature engineering), they can potentially reduce dimensionality and show additional relationships between features. They can therefore increase the interpretability of the production model.

If you use challenger models for this purpose, the process should be made explicit and documented. Moreover, any highly engineered features that are

drawn from challenger models and used in production models must be properly justified and annotated in the metadata to indicate what attribute the combined feature represents and how such an attribute might be a factor in evidence-based reasoning.

When challenger models are used to process the data of affected decision recipients – even for benchmarking purposes – they should be properly recorded and documented. If the insights from this challenger model's processing are incorporated into any dimension of actual decision-making (for instance, the comparative benchmarking results are shared with implementers/users, who are making decisions), you should treat them as core production models, document them, and hold them to the same explainability standards.

What types of models are we choosing between?

To help you get a better picture of the spectrum of algorithmic techniques, the following table lays out some of the basic properties, potential uses, and interpretability characteristics of the most widely used algorithms at present.

The first eleven techniques listed are considered to be largely interpretable, although for some of them, like the regression-based and tree-based algorithms, this depends on the number of input features that are being processed. The final four techniques are more or less considered to be 'black box' algorithms.

Algorithm type	**Basic description**	**Possible uses**	**Interpretability**
Linear regression (LR)	Makes predictions about a target variable by summing weighted input/predictor variables.	Advantageous in highly regulated sectors like finance (eg credit scoring) and healthcare (predict disease risk given eg lifestyle and existing health conditions) because it's	High level of interpretability because of linearity and monotonicity. Can become less interpretable with increased number of features (ie high dimensionality).

		simpler to calculate and have oversight over.	
Logistic regression	Extends linear regression to classification problems by using a logistic function to transform outputs to a probability between 0 and 1.	Like linear regression, advantageous in highly regulated and safety-critical sectors, but in use cases that are based in classification problems such as yes/no decisions on risks, credit, or disease.	Good level of interpretability but less so than LR because features are transformed through a logistic function and related to the probabilistic result logarithmically rather than as sums.
Regularised regression (LASSO and Ridge)	Extends linear regression by adding penalisation and regularisation to feature weights to increase sparsity/ reduce dimensionality.	Like linear regression, advantageous in highly regulated and safety-critical sectors that require understandable, accessible, and transparent results.	High level of interpretability due to improvements in the sparsity of the model through better feature selection procedures.
Generalised linear model (GLM)	To model relationships between features and target variables that do not follow normal (Gaussian) distributions a GLM	This extension of LR is applicable to use cases where target variables have constraints that require the exponential family	Good level of interpretability that tracks the advantages of LR while also introducing more flexibility.

	introduces a link function that allows for the extension of LR to non-normal distributions.	set of distributions (for instance, if a target variable involves number of people, units of time or probabilities of outcome, the result has to have a non-negative value).	Because of the link function, determining feature importance may be less straightforward than with the additive character of simple LR, a degree of transparency may be lost.
Generalised additive model (GAM)	To model non-linear relationships between features and target variables (not captured by LR), a GAM sums non-parametric functions of predictor variables (like splines or tree-based fitting) rather than simple weighted features.	This extension of LR is applicable to use cases where the relationship between predictor and response variables is not linear (i.e where the input-output relationship changes at different rates at different times) but optimal interpretability is desired.	Good level of interpretability because, even in the presence of non-linear relationships, the GAM allows for clear graphical representation of the effects of predictor variables on response variables.
Decision tree (DT)	A model that uses inductive branching methods to split data into interrelated decision nodes which terminate in classifications or	Because the step-by-step logic that produces DT outcomes is easily understandable to non-technical users (depending on number of	High level of interpretability if the DT is kept manageably small, so that the logic can be followed end-to-end. The

	predictions. DT's moves from starting 'root' nodes to terminal 'leaf' nodes, following a logical decision path that is determined by Boolean-like 'if-then' operators that are weighted through training.	nodes/ features), this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where volume of relevant features is reasonably low.	advantage of DT's over LR is that the former can accommodate non-linearity and variable interaction while remaining interpretable.
Rule/decision lists and sets	Closely related to DT's, rule/decision lists and sets apply series of if-then statements to input features in order to generate predictions. Whereas decision lists are ordered and narrow down the logic behind an output by applying 'else' rules, decision sets keep individual if-then statements unordered and largely independent, while weighting them so that rule voting can occur in generating predictions.	As with DT's, because the logic that produces rule lists and sets is easily understandable to non-technical users, this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where the clear and fully transparent justification of outcomes is a priority.	Rule lists and sets have one of the highest degrees of interpretability of all optimally performing and non-opaque algorithmic techniques. However, they also share with DT's the same possibility that degrees of understandability are lost as the rule lists get longer or the rule sets get larger.
Case-based reasoning	Using exemplars drawn from prior	CBR is applicable in any domain	CBR is interpretable-by-

(CBR)/ Prototype and criticism	human knowledge, CBR predicts cluster labels by learning prototypes and organising input features into subspaces that are representative of the clusters of relevance. This method can be extended to use maximum mean discrepancy (MMD) to identify 'criticisms' or slices of the input space where a model most misrepresents the data. A combination of prototypes and criticisms can then be used to create optimally interpretable models.	where experience-based reasoning is used for decision-making. For instance, in medicine, treatments are recommended on a CBR basis when prior successes in like cases point the decision maker towards suggesting that treatment. The extension of CBR to methods of prototype and criticism has meant a better facilitation of understanding of complex data distributions, and an increase in insight, actionability, and interpretability in data mining.	design. It uses examples drawn from human knowledge in order to syphon input features into human recognisable representations. It preserves the explainability of the model through both sparse features and familiar prototypes.
Supersparse linear integer model (SLIM)	SLIM utilises data-driven learning to generate a simple scoring system that only requires users to add, subtract, and multiply a few numbers in order to make a prediction. Because SLIM produces such a sparse and	SLIM has been used in medical applications that require quick and streamlined but optimally accurate clinical decision-making. A version called Risk-Calibrated SLIM (RiskSLIM) has been applied to	Because of its sparse and easily understandable character, SLIM offers optimal interpretability for human-centred decision-support. As a manually completed scoring system, it also ensures the active

	accessible model, it can be implemented quickly and efficiently by non-technical users, who need no special training to deploy the system.	the criminal justice sector to show that its sparse linear methods are as effective for recidivism prediction as some opaque models that are in use.	engagement of the interpreter-user, who implements it.
Naïve Bayes	Uses Bayes rule to estimate the probability that a feature belongs to a given class, assuming that features are independent of each other. To classify a feature, the Naïve Bayes classifier computes the posterior probability for the class membership of that feature by multiplying the prior probability of the class with the class conditional probability of the feature.	While this technique is called naïve for reason of the unrealistic assumption of the independence of features, it is known to be very effective. Its quick calculation time and scalability make it good for applications with high dimensional feature spaces. Common applications include spam filtering, recommender systems, and sentiment analysis.	Naïve Bayes classifiers are highly interpretable, because the class membership probability of each feature is computed independently. The assumption that the conditional probabilities of the independent variables are statistically independent, however, is also a weakness, because feature interactions are not considered.
K-nearest neighbour (KNN)	Used to group data into clusters for purposes of either classification or prediction, this technique identifies	KNN is a simple, intuitive, versatile technique that has wide applications but works best with smaller	KNN works off the assumption that classes or outcomes can be predicted by looking at the

	a neighbourhood of nearest neighbours around a data point of concern and either finds the mean outcome of them for prediction or the most common class among them for classification.	datasets. Because it is non-parametric (makes no assumptions about the underlying data distribution), it is effective for non-linear data without losing interpretability. Common applications include recommender systems, image recognition, and customer rating and sorting.	proximity of the data points upon which they depend to data points that yielded similar classes and outcomes. This intuition about the importance of nearness/proximity is the explanation of all KNN results. Such an explanation is more convincing when the feature space remains small, so that similarity between instances remains accessible.
Support vector machines (SVM)	Uses a special type of mapping function to build a divider between two sets of features in a high dimensional feature space. An SVM therefore sorts two classes by maximising the margin of the decision boundary between them.	SVM's are extremely versatile for complex sorting tasks. They can be used to detect the presence of objects in images (face/no face; cat/no cat), to classify text types (sports article/arts article), and to identify genes of interest in bioinformatics.	Low level of interpretability that depends on the dimensionality of the feature space. In context-determined cases, the use of SVM's should be supplemented by secondary explanation tools.

Artificial neural net (ANN)	Family of non-linear statistical techniques (including recurrent, convolutional, and deep neural nets) that build complex mapping functions to predict or classify data by employing the feedforward—and sometimes feedback—of input variables through trained networks of interconnected and multi-layered operations.	ANN's are best suited to complete a wide range of classification and prediction tasks for high dimensional feature spaces—ie cases where there are very large input vectors. Their uses may range from computer vision, image recognition, sales and weather forecasting, pharmaceutical discovery, and stock prediction to machine translation, disease diagnosis, and fraud detection.	The tendencies towards curviness (extreme non-linearity) and high-dimensionality of input variables produce very low-levels of interpretability in ANN's. They are considered to be the epitome of 'black box' techniques. Where appropriate, the use of ANN's should be supplemented by secondary explanation tools.
Random Forest	Builds a predictive model by combining and averaging the results from multiple (sometimes thousands) of decision trees that are trained on random subsets of shared features and training data.	Random forests are often used to effectively boost the performance of individual decisions trees, to improve their error rates, and to mitigate overfitting. They are very popular in high-dimensional problem areas like genomic medicine	Very low levels of interpretability may result from the method of training these ensembles of decision trees on bagged data and randomised features, the number of trees in a given forest, and the possibility that individual trees may have

		and have also been used extensively in computational linguistics, econometrics, and predictive risk modelling.	hundreds or even thousands of nodes.
Ensemble methods	As their name suggests, ensemble methods are a diverse class of meta-techniques that combines different 'learner' models (of the same or different type) into one bigger model (predictive or classificatory) in order to decrease the statistical bias, lessen the variance, or improve the performance of any one of the sub-models taken separately.	Ensemble methods have a wide range of applications that tracks the potential uses of their constituent learner models (these may include DT's, KNN's, Random Forests, Naïve Bayes, etc.).	The interpretability of Ensemble Methods varies depending upon what kinds of methods are used. For instance, the rationale of a model that uses bagging techniques, which average together multiple estimates from learners trained on random subsets of data, may be difficult to explain. Explanation needs of these kinds of techniques should be thought through on a case-by-case basis.

> Further reading on algorithm types

Tools for extracting explanations

Extracting and delivering meaningful explanations about the underlying logic of your AI model's results involves both technical and non-technical components.

At the technical level, to be able to offer an explanation of how your model reached its results, you need to:

- become familiar with how AI explanations are extracted from intrinsically interpretable models;
- get to know the supplementary explanation tools that may be used to shed light on the logic behind the results and behaviours of 'black box' systems; and
- learn how to integrate these different supplementary techniques in a way that will enable you to provide meaningful information about your system to its users and decision recipients.

At the non-technical level, extracting and delivering meaningful explanations involves establishing how conveying your model's results can reliably and clearly enable users and implementers to:

- exercise better-informed judgements; and
- offer plausible and easily understandable accounts of the logic behind its output to affected individuals and concerned parties.

Technical dimensions of AI interpretability

Before going into detail about how to set up a strategy for explaining your AI model, you need to be aware of a couple of commonly used distinctions that will help you and your team to think about what is possible and desirable for an AI explanation.

- Local vs global explanation

The distinction between the explanation of single instances of a model's results and an explanation of how it works across all of its outputs is often characterised as the difference between local explanation and global

explanation. Both types of explanation offer potentially helpful support for providing significant information about the rationale behind an AI system's output.

A **local explanation** aims to interpret individual predictions or classifications. This may involve identifying the specific input variables or regions in the input space that had the most influence in generating a particular prediction or classification.

Providing a **global explanation** entails offering a wide-angled view that captures the inner-workings and logic of that model's behaviour in sum and across predictions or classifications. This kind of explanation can capture the overall significance of features and variable interactions for model outputs and significant changes in the relationship of predictor and response variables across instances. It can also provide insights into dataset-level and population-level patterns, which are crucial for both big picture and case-focused decision-making.

- Internal/model intrinsic vs. external/post-hoc explanation

Providing an **internal or model intrinsic explanation** of an AI model involves making the way its components and relationships function intelligible. It is therefore closely related to, and overlaps to some degree with, global explanation - but it is not the same. An internal explanation makes insights available about the parts and operations of an AI system **from the inside**. These insight can help your team understand why the trained model does what it does, and how to improve it.

Similarly, when this type of internal explanation is applied to a 'black box model', it can shed light on that opaque model's operation by breaking it down into more understandable, analysable, and digestible parts. For example, in the case of an artificial neural network (ANN) it can break it down into interpretable characteristics of its vectors, features, interactions, layers, parameters etc. This is often referred to as 'peeking into the black box'.

Whereas internal explanations can be drawn from both interpretable and opaque AI systems, **external or post-hoc explanations** are more applicable to 'black box' systems where it is not possible to fully access the internal underlying rationale due to the model's complexity and high dimensionality.

Post-hoc explanations attempt to capture essential attributes of the observable behaviour of a 'black box' system by subjecting it to a number of different techniques that reverse-engineer explanatory insights. Post-hoc approaches can do a number of different things:

- test the sensitivity of the outputs of an opaque model to perturbations in its inputs;
- allow for the interactive probing of its behavioural characteristics; or
- build proxy-based models that utilise simplified interpretable techniques to gain a better understanding of particular instances of its predictions and classifications, or of system behaviour as a whole.

Getting familiar with AI explanations through interpretable models

For AI models that are basically interpretable (such as regression-based and decision/rule-based systems, Naïve Bayes, and K nearest neighbour), the technical aspect of extracting a meaningful explanation is relatively straightforward - draw on the intrinsic logic of the model's mapping function by looking directly at it and at its results.

For instance, in decision trees or decision/rule lists, the logic behind an output will depend on the interpretable relationships of weighted conditional (if-then) statements. In other words, each node or component of these kinds of models is, in fact, operating **as a reason**. Extracting a meaningful explanation from them therefore factors down to following the path of connections between these reasons.

Note, though, that if a decision tree is excessively deep or a given decision list is overly long, it will be challenging to interpret the logic behind their outputs. Human-scale reasoning, generally speaking, operates on the basis of making connections between only a few variables at a time, so a tree or a list with thousands of features and relationships will be significantly harder to follow and thus less interpretable. In these more complex cases, an interpretable model may lose much of its global as well as its local explainability.

Similar advantages and disadvantages have long been recognised in the explainability of regression-based models. Clear-cut interpretability has made this class of algorithmic techniques a favoured choice in high-stakes and highly regulated domains because many of them possess linearity, monotonicity, and sparsity/non-complexity:

Characteristics of regression-based models that allow for optimal explainability and transparency

- **Linearity:** Any change in the value of the predictor variable is directly reflected in a change in the value of the response variable at a constant rate b. The interpretable prediction yielded by the model can therefore be directly inferred from the relative significance of the parameter/weights of the predictor variable and have high inferential clarity and strength.
- **Monotonicity:** When the value of the predictor changes in a given direction, the value of the response variable changes consistently either in the same or opposite direction. The interpretable prediction yielded by the model can therefore be directly inferred. This monotonicity dimension is a highly desirable interpretability condition of predictive models in many heavily regulated sectors, because it incorporates reasonable expectations about the consistent application of sector specific selection constraints into automated decision-making systems.
- **Sparsity/non-complexity**: The number of features (dimensionality) and feature interactions is low enough and the model of the underlying distribution is simple enough to enable a clear understanding of the function of each part of the model in relation to its outcome.

In general, it is helpful to get to know the range of techniques that are available for the explanation of interpretable AI models such as those listed above. These techniques not only make the rationale behind models like logic-based decision trees, lists, and sets, and of regression-based systems readily interpretable; they also form the basis of many of the supplementary explanation tools that are widely used to make 'black box' models more interpretable.

Technical strategies for explaining 'black box' AI models through supplementary explanation tools

If, after considering domain, impact, and technical factors, you have chosen to use a 'black box' AI system, your next step is to incorporate appropriate supplementary explanation tools into building your model.

There is no comprehensive or one-size-fits-all technical solution for making opaque algorithms interpretable. The supplementary explanation strategies available to support interpretability may shed light on significant aspects of a model's global processes and components of its local results.

However, often these strategies operate as imperfect approximations or as simpler surrogate models, which do not fully capture the complexities of the original opaque system. This means that overreliance on supplementary tools may be misleading.

With this in mind, 'fidelity' may be a suitable primary goal for your technical 'black box' explanation strategy. In order for your supplementary tool to achieve a high level of fidelity, it should provide a reliable and accurate approximation of the system's behaviour.

For practical purposes, you should think both locally and globally when choosing the supplementary explanation tools that will achieve fidelity.

Thinking locally is a priority, because the primary concern of AI explainability is to make the results of specific data processing activity clear and understandable to affected individuals. This is local explanation.

Even so, it is just as important to provide supplementary global explanations of your AI system. Understanding the relationship between your system's component parts (its features, parameters, and interactions) and its behaviour as a whole will often be a critical factor in setting up an accurate local explanation. It will also be essential to securing your AI system's fairness, safety and optimal performance.

This sort of global understanding may also provide crucial insights into your model's more general potential impacts on individuals and wider society, as well as allow your team to improve the model, so that concerns raised by such global insights can be properly addressed.

Below we provide you with a table containing some of the more widely used supplementary explanation strategies and tools. Keep in mind, though, that this is a rapidly developing field, so remaining up to date with the latest tools will mean that you and technical members of your team will need to move beyond the basic information we are offering here.

Supplementary explanation strategy	**What is it and what is it useful for?**	**Limitations**

Surrogate models (SM)	SM's build a simpler interpretable model (often a decision tree or rule list) from the dataset and predictions of an opaque system. The purpose of the SM is to provide an understandable proxy of the complex model that estimates that model well, while not having the same degree of opacity. They are good for assisting in processes of model diagnosis and improvement and can help to expose overfitting and bias. They can also represent some non-linearities and interactions that exist in the original model.	As approximations, SM's often fail to capture the full extent of non-linear relationships and high-dimensional interactions among features. There is a seemingly unavoidable trade-off between the need for the SM to be sufficiently simple so that it is understandable by humans, and the need for that model to be sufficiently complex so that it can represent the intricacies of how the mapping function of a 'black box' model works as a whole. That said, the R^2 measurement can provide a good quantitative metric of the accuracy of the SM's approximation of the original complex model.
Global/local? Internal/post-hoc?	For the most part, SM's may be used both globally and locally. As simplified proxies, they are post-hoc.	
Partial Dependence Plot (PDP)	A PDP calculates and graphically represents the marginal effect of one or two input features on the output of an opaque model by probing the dependency relation between the input variable(s) of	While PDP's allow for valuable access to non-linear relationships between predictor and response variables, and therefore also for comparisons of model behaviour with domain-informed expectations of reasonable relationships

	interest and the predicted outcome across the dataset, while averaging out the effect of all the other features in the model. This is a good visualisation tool, which allows a clear and intuitive representation of the nonlinear behaviour for complex functions (like random forests and SVM's). It is helpful, for instance, in showing that a given model of interest meets monotonicity constraints across the distribution it fits.	between features and outcomes, they do not account for interactions between the input variables under consideration. They may, in this way, be misleading when certain features of interest are strongly correlated with other model features. Because PDP's average out marginal effects, they may also be misleading if features have uneven effects on the response function across different subsets of the data—ie where they have different associations with the output at different points. The PDP may flatten out these heterogeneities to the mean.
Global/local? Internal/post-hoc?	PDP's are global post-hoc explainers that can also allow deeper causal understandings of the behaviour of an opaque model through visualisation. These insights are, however, very partial and incomplete both because PDP's are unable to represent feature interactions and heterogenous effects, and because they are unable to graphically represent more than a couple of features at a time (human spatial thinking is limited to a few dimensions, so only two variables in 3D space are easily graspable).	
Individual Conditional	Refining and extending PDP's, ICE plots graph	When used in combination with PDP's, ICE plots can

Expectations Plot (ICE)	the functional relationship between a single feature and the predicted response for an individual instance. Holding all features constant except the feature of interest, ICE plots represent how, for each observation, a given prediction changes as the values of that feature vary. Significantly, ICE plots therefore disaggregate or break down the averaging of partial feature effects generated in a PDP by showing changes in the feature-output relationship for each specific instance, ie observation-by-observation. This means that it can both detect interactions and account for uneven associations of predictor and response variables.	provide local information about feature behaviour that enhances the coarser global explanations offered by PDP's. Most importantly, ICE plots are able to detect the interaction effects and heterogeneity in features that remain hidden from PDP's in virtue of the way they compute the partial dependence of outputs on features of interest by averaging out the effect of the other predictor variables. Still, although ICE plots can identify interactions, they are also liable to missing significant correlations between features and become misleading in some instances. Constructing ICE plots can also become challenging when datasets are very large. In these cases, time-saving approximation techniques such as sampling observation or binning variables can be employed (but, depending on adjustments and size of the dataset, with an unavoidable impact on explanation accuracy).
Global/local? Internal/post-hoc?	ICE plots offer a local and post-hoc form of supplementary explanation.	

Accumulated Local Effects Plots (ALE)	As an alternative approach to PDP's, ALE plots provide a visualisation of the influence of individual features on the predictions of a 'black box' model by averaging the sum of prediction differences for instances of features of interest in localised intervals and then integrating these averaged effects across all of the intervals. By doing this, they are able to graph the accumulated local effects of the features on the response function as a whole. Because ALE plots use local differences in prediction when computing the averaged influence of the feature (instead of its marginal effect as do PDP's), it is able to better account for feature interactions and avoid statistical bias. This ability to estimate and represent feature influence in a correlation-aware manner is an advantage of ALE plots. ALE plots are also more computationally	A notable limitation of ALE plots has to do with the way that they carve up the data distribution into intervals that are largely chosen by the explanation designer. If there are too many intervals, the prediction differences may become too small and less stably estimate influences. If the intervals are widened too much, the graph will cease to sufficiently represent the complexity of the underlying model. While ALE plots are good for providing global explanations that account for feature correlations, the strengths of using PDP's in combination with ICE plots should also be considered (especially when there are less interaction effects in the model being explained). All three visualisation techniques shed light on different dimensions of interest in explaining opaque systems, so the appropriateness of employing them should be weighed case-by-case.

	tractable than PDP's because they are able to use techniques to compute effects in smaller intervals and chunks of observations.	
Global/local? Internal/post-hoc?	ALE plots are a global and post-hoc form of supplementary explanation.	
Global Variable Importance	The global variable importance strategy calculates the contribution of each input feature to model output across the dataset by permuting the feature of interest and measuring changes in the prediction error; if changing the value of the permuted feature increases the model error, then that feature is considered to be important. Utilising global variable importance to understand the relative influence of features on the performance of the model can provide significant insight into the logic underlying the model's behaviour. This method also provides valuable understanding about non-linearities in the complex model that	While permuting variables to measure their relative importance, to some extent, accounts for interaction effects, there is still a high degree of imprecision in the method with regard to which variables are interacting and how much these interactions are impacting the performance of the model. A bigger picture limitation of global variable importance comes from what is known as the 'Rashomon effect'. This refers to the variety of different models that may fit the same data distribution equally well. These models may have very different sets of significant features. Because the permutation-based technique can only provide explanatory insight with regard to a single model's performance, it is unable to address this wider

	is being explained.	problem of the variety of effective explanation schemes.
Global/local? Internal/post-hoc?	Global variable importance is a form of global and post-hoc explanation.	
Global Variable Interaction	The global variable interaction strategy computes the importance of variable interactions across the dataset by measuring the variance in the model's prediction when potentially interacting variables are assumed to be independent. This is primarily done by calculating an 'H-statistic' where a no-interaction partial dependence function is subtracted from an observed partial dependence function in order to compute the variance in the prediction. This is a versatile explanation strategy, which has been employed to calculate interaction effects in many types of complex models including ANN's and Random Forests. It can be used to calculate interactions between	While the basic capacity to identify interaction effects in complex models is a positive contribution of global variable interaction as a supplementary explanatory strategy, there are a couple of potential drawbacks to which you may want to pay attention. First, there is no established metric in this method to determine the quantitative threshold across which measured interactions become significant. The relative significance of interactions is useful information as such, but there is no way to know at which point interactions are strong enough to exercise effects. Second, the computational burden of this explanation strategy is very high, because interaction effects are being calculated combinatorially across all the data points. This means

	two or more variables and also between variables and the response function as a whole. It has been effectively used, for example, in biological research to identify interaction effects among genes.	that as the number of data points increase, the number of necessary computations increase exponentially.
Global/local? Internal/post-hoc?	Global variable interaction is a form of global and post-hoc explanation.	
Sensitivity Analysis and Layer-Wise Relevance Propagation (LRP)	Sensitivity analysis and LRP are supplementary explanation tools used for artificial neural networks. Sensitivity analysis identifies the most relevant features of an input vector by calculating local gradients to determine how a data point has to be moved to change the output label. Here, an output's sensitivity to such changes in input values identifies the most relevant features. LRP is another method to identify feature relevance that is downstream from sensitivity analysis. It uses a strategy of moving backward through the layers of a	Both sensitivity analysis and LRP identify important variables in the vastly large feature spaces of neural nets. These explanatory techniques find visually informative patterns by mathematically piecing together the values of individual nodes in the network. As a consequence of this piecemeal approach, they offer very little by way of an account of the reasoning or logic behind the results of an ANNs' data processing. Recently, more and more research has focused on attention-based methods of identifying the higher-order representations that are guiding the mapping functions of these kinds of

	neural net graph to map patterns of high activation in the nodes and ultimately generates interpretable groupings of salient input variables that can be visually represented in a heat or pixel attribution map.	models as well as on interpretable CBR methods that are integrated into ANN architectures and that analyse images by identifying prototypical parts and combining them into a representational wholes. These newer techniques are showing that some significant progress is being made in uncovering the underlying logic of some ANN's.
Global/local? Internal/post-hoc?	Sensitivity analysis and salience mapping are forms of local and post-hoc explanation, although the recent incorporation of CBR techniques is moving neural net explanations toward a more internal basis of interpretation.	
Local Interpretable Model-Agnostic Explanation (LIME) and anchors	LIME works by fitting an interpretable model to a specific prediction or classification produced by an opaque system. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under scrutiny.	While LIME appears to be a step in the right direction, in its versatility and in the availability of many iterations in very useable software, a host of issues that present challenges to the approach remains unresolved. For instance, the crucial aspect of how to properly define the proximity measure for the 'neighbourhood' or 'local region' where the explanation applies remains unclear, and small changes in the scale of the chosen

	LIME does this by generating a simple linear regression model by weighting the values of the data points, which were produced by randomly perturbing the opaque model, according to their proximity to the original prediction or classification. The closest of these values to the instance being explained are weighted the heaviest, so that the supplemental model can produce an explanation of feature importance that is locally faithful to that instance. Note that other interpretive models like decision trees may be used as well.	measure can lead to greatly diverging explanations. Likewise, the explanation produced by the supplemental linear model can quickly become unreliable, even with small and virtually unnoticeable perturbations of the system it is attempting to approximate. This challenges the basic assumption that there is always some simplified interpretable model that successfully approximates the underlying model reasonably well near any given data point. LIME's creators have largely acknowledged these shortcomings and have recently offered a new explanatory approach that they call 'anchors'. These 'high precision rules' incorporate into their formal structures 'reasonable patterns' that are operating within the underlying model (such as the implicit linguistic conventions that are at work in a sentiment prediction model), so that they can establish suitable and faithful boundaries of their explanatory coverage of its predictions or classifications.
Global/local?	LIME offers a local and post-hoc form of supplementary	

Internal/post-hoc?	explanation.	
Shapley Additive ExPlanations (SHAP)	SHAP uses concepts from cooperative game theory to define a 'Shapley value' for a feature of concern that provides a measurement of its influence on the underlying model's prediction. Broadly, this value is calculated by averaging the feature's marginal contribution to every possible prediction for the instance under consideration. The way SHAP computes marginal contributions is by constructing two instances: the first instance includes the feature being measured, while the second leaves it out by substituting a randomly selected stand-in variable for it. After calculating the prediction for each of these instances by plugging their values into the original model, the result of the second is subtracted from that of the first to determine the marginal contribution of the feature. This procedure	Of the several drawbacks of SHAP, the most practical one is that such a procedure is computationally burdensome and becomes intractable beyond a certain threshold. Note, though, some later SHAP versions do offer methods of approximation such as Kernel SHAP and Shapley Sampling Values to avoid this excessive computational expense. These methods do, however, affect the overall accuracy of the method. Another significant limitation of SHAP is that its method of sampling values in order to measure marginal variable contributions assumes feature independence (ie that values sampled are not correlated in ways that might significantly affect the output for a particular calculation). As a consequence, the interaction effects engendered by and between the stand-in variables that are used as substitutes for left-out features are necessarily unaccounted for

	is then repeated for all possible combinations of features so that the weighted average of all of the marginal contributions of the feature of concern can be computed. This method then allows SHAP, by extension, to estimate the Shapley values for all input features in the set to produce the complete distribution of the prediction for the instance. While computationally intensive, this means that for the calculation of the specific instance, SHAP can axiomatically guarantee the consistency and accuracy of its reckoning of the marginal effect of the feature. This computational robustness has made SHAP attractive as an explainer for a wide variety of complex models, because it can provide a more comprehensive picture of relative feature influence for a given instance than any other post-hoc explanation tool.	when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced, because the complexity of multivariate interactions in the underlying model may not be sufficiently captured by the simplicity of this supplemental interpretability technique. This drawback in sampling (as well as a certain degree of arbitrariness in domain definition) can cause SHAP to become unreliable even with minimal perturbations of the model it is approximating. There are currently efforts being made to account for feature dependencies in the SHAP calculations. The original creators of the technique have introduced Tree SHAP to, at least partially, include feature interactions. Others have recently introduced extensions of Kernel SHAP.

Global/local? Internal/post-hoc?	SHAP offers a local and post-hoc form of supplementary explanation.	
Counterfactual Explanation	Counterfactual explanations offer information about how specific factors that influenced an algorithmic decision can be changed so that better alternatives can be realised by the recipient of a particular decision or outcome. Incorporating counterfactual explanations into a model at its point of delivery allows stakeholders to see what input variables of the model can be modified, so that the outcome could be altered to their benefit. For AI systems that assist decisions about changeable human actions (like loan decisions or credit scoring), incorporating counterfactual explanation into the development and testing phases of model development may allow the incorporation of actionable variables, ie input variables that will afford decision subjects	While counterfactual explanation offers a useful way to contrastively explore how feature importance may influence an outcome, it has limitations that originate in the variety of possible features that may be included when considering alternative outcomes. In certain cases, the sheer number of potentially significant features that could be at play in counterfactual explanations of a given result can make a clear and direct explanation difficult to obtain and selected sets of possible explanations seem potentially arbitrary. Moreover, there are as yet limitations on the types of datasets and functions to which these kinds of explanations are applicable. Finally, because this kind of explanation concedes the opacity of the algorithmic model outright, it is less able to address concerns about potentially harmful feature interactions and questionable covariate relationships that may be

	with concise options for making practical changes that would improve their chances of obtaining the desired outcome. In this way, counterfactual explanatory strategies can be used as way to incorporate reasonableness and the encouragement of agency into the design and implementation of AI systems.	buried deep within the model's architecture. It is a good idea to use counterfactual explanations in concert with other supplementary explanation strategies—that is, as one component of a more comprehensive explanation portfolio.
Global/local? Internal/post-hoc?	Counterfactual explanations are a local and post-hoc form of supplementary explanation strategy.	
Self-Explaining and Attention-Based Systems	Self-explaining and attention-based systems actually integrate secondary explanation tools into the opaque systems so that they can offer runtime explanations of their own behaviours. For instance, an image recognition system could have a primary component, like a convolutional neural net, that extracts features from its inputs and classifies them while a secondary component,	Automating explanations through self-explaining systems is a promising approach for applications where users benefit from gaining real-time insights about the rationale of the complex systems they are operating. However, regardless of their practical utility, these kinds of secondary tools will only work as well as the explanatory infrastructure that is actually unpacking their underlying logics. This explanatory layer must remain accessible to human

	like a built-in recurrent neural net with an 'attention-directing' mechanism translates the extracted features into a natural language representation that produces a sentence-long explanation of the result to the user. Research into integrating 'attention-based' interfaces is continuing to advance toward potentially making their implementations more sensitive to user needs, explanation-forward, and humanly understandable. Moreover, the incorporation of domain knowledge and logic-based or convention-based structures into the architectures of complex models are increasingly allowing for better and more user-friendly representations and prototypes to be built into them.	evaluators and be understandable to affected individuals. Self-explaining systems, in other words, should themselves remain optimally interpretable. The task of formulating a primary strategy of supplementary explanation is still part of the process of building out a system with self-explaining capacity. Another potential pitfall to consider for self-explaining systems is their ability to mislead or to provide false reassurance to users, especially when humanlike qualities are incorporated into their delivery method. This can be avoided by not designing anthropomorphic qualities into their user interface and by making uncertainty and error metrics explicit in the explanation as it is delivered.
Global/local? Internal/post-hoc?	Because self-explaining and attention-based systems are secondary tools that can utilise many different methods of explanation, they may be global or local, internal or post-hoc, or a combination of any of them.	

> Further reading on supplementary techniques

Combining and integrating supplementary explanation strategies

The main purpose of using supplementary explanation tools is to make the underlying rationale of the results of an AI system's data processing both optimally interpretable and more easily intelligible to those who use the system and to decision recipients.

For this reason, it is a good idea to think about using different explanation strategies in concert. You can combine explanation tools to enable affected individuals to make sense of the reasoning behind an AI-assisted decision with as much clarity and precision as possible.

With this in mind, it might be helpful to think about how to combine these different strategies into a portfolio of tools for explanation extraction that best serves the needs of your particular AI system, and that is most appropriate for providing meaningful information about the rationale of its results.

Keeping in mind the various strategies we have introduced in the table above, there are three interrelated layers of technical rationale that might be considered as especially significant components to include in such a portfolio:

- visualisation of how the model works;
- explanation of variable importance and interaction effects, both global and local; and
- counterfactual tools to explore alternative possibilities and actionable recourse.

Here are some questions that may assist you in thinking about how to integrate these layers of explanation extraction:

Visualisation of how the model works

How might graphical tools like ALE plots or a combination of PDP's and ICE plots make the logic behind both the global and the local behaviour of our model clearer to users, implementers, auditors and decision recipients? How might these tools be used to improve the model and to ensure that it operates in accordance with reasonable expectations?

How can domain knowledge and understanding the use case inform the insights derived from visualisation techniques? How might this knowledge

inform the integration of visualisation techniques with other explanation tools?

What are the most effective ways that such visualisations can be presented and explained to users and decision recipients so as to help them build a mental model of how the system works, both as a whole and in specific instances? How can they be used to enhance evidence-based reasoning?

Are other visualisation techniques available (like heat maps, interactive querying tools for ANN's, or more traditional 2D tools like principle components analysis) that would also be helpful to enhance the interpretability of our system?

Understanding of the role of variables and variable interactions

How can global measures of feature importance and feature interactions be utilised to help users and decision recipients better understand the underlying logic of the model as a whole?

How might they provide reassurance that the model is yielding results that are in line with reasonable expectations?

How might they support and enhance the information being provided in the visualisation tools?

How might measures of variable importance and interaction effects be used to confirm that our AI system is operating fairly and is not harming or discriminating against affected stakeholders?

Which local, post-hoc explanation tools—like LIME, SHAP, LOCO (Leave-One-Covariate-Out), etc—are reliable enough in the context of our particular AI system to be useful as part of its portfolio of explanation extraction tools?

Have we established through model exploration and testing that using these local explanation tools will help us to provide meaningful information that is informative rather than misleading or inaccurate?

Understanding of how the behaviours or circumstances that influence an AI-assisted decision would need to be changed to change that decision

Are counterfactual explanations appropriate for the use case of our AI application? If so, have alterable features been included in the input space that can provide decision recipients with reasonable options to change their behaviour in order to obtain different results?

Have we used a solid understanding of global feature importance, correlations, and interaction effects to set up reasonable and relevant options for the possible alternative outcomes that will be explored in our counterfactual explanation tool?

Step 4: Translate the rationale of your system's results into useable and easily understandable reasons

At a glance

- Once you have extracted the rationale of the underlying logic of your AI model, you will need to take the statistical output and incorporate it into your wider decision-making process.
- Implementers of the outputs from your AI system will need to recognise the factors that they see as legitimate determinants of the outcome they are considering.
- For the most part, the AI systems we consider in this guidance will produce statistical outputs that are based on correlation rather than causation. You therefore need to sense-check whether the correlations that the AI model produces make sense in the case you are considering.
- Decision recipients should be able to easily understand how the statistical result has been applied to their particular case.

Checklist

☐ We have taken the technical explanation delivered by our AI system and translated this into reasons that can be easily understood by the decision recipient.

☐ We have used tools such as textual clarification, visualisation media, graphical representations, summary tables, or a combination, to present information about the logic of the AI system's output.

☐ We have justified how we have incorporated the statistical inferences from the AI system into our final decision and rationale explanation.

Where there is a 'human in the loop' we have trained our implementers to:

☐ Understand the associations and correlations that link the input data to the model's prediction or classification.

☐ Interpret which correlations are consequential for providing a meaningful explanation by drawing on their domain knowledge or the decision recipient's specific circumstances.

☐ Combine the chosen correlations and outcome determinants with what they know of the individual affected to come to their conclusion.

☐ Apply the AI model's results to the individual case at hand, rather than uniformly across decision recipients.

☐ Where our decision-making is fully automated, we have made sure that our AI system is set up to provide understandable explanations to individuals.

In more detail

Introduction

The non-technical dimension to rationale explanation involves working out how you are going to convey your model's results in a way that is clear and understandable to users, implementers and decision recipients.

This involves presenting information about the logic of the output as clearly and meaningfully as possible. You could do this through textual clarification, visualisation media, graphical representations, summary tables, or any combination of them. The main thing is to make sure that there is a simple way for the implementer to describe the result to an affected individual.

However, it is important to remember that the technical rationale behind an AI model's output is only one component of the decision-making and explanation process. It reveals the statistical inferences (correlations) that implementers must then incorporate into their wider deliberation before they reach their ultimate conclusions and explanations.

Integrating statistical associations into their wider deliberations means implementers should be able to recognise the factors that they see as legitimate determinants of the outcome they are considering. They must be able to pick out, amongst the model's correlations, those associations that they think reasonably explain the outcome given the specifics of the case. They then need to be able to incorporate these legitimate determining factors into their thinking about the AI-supported decision, and how to explain it.

It is likely they will need training in order to do this.

Understand the statistical rationale

Once you have extracted your explanation, either from an inherently interpretable model or from supplementary tools, you should have a good idea of both the relative feature important and significant feature interactions. This is your local explanation, which you should combine with a more global picture of the behaviour of the model across cases. Doing this should help clarify where there is a meaningful relationship between the predictor and response variables.

Understanding the relevant associations between input variables and an AI model's result (ie its statistical rationale) is the first step in moving from the model's mathematical inferences to a meaningful explanation. However, on their own, these statistical inferences are not direct indicators of what determined the outcome, or of significant population-level insights in the real world.

As a general rule, the kinds of AI and machine learning models that we are exploring in this guidance generate statistical outputs that are based on **correlational** rather than **causal** inference. In these models, a set of relevant input features, X, is linked to a target or response variable, Y, where there is an established association or correlation between them. While it is justified, then, to say that the components of X are correlated (in some unspecified way) with Y, it is not justified (on the basis of the statistical inference alone) to say that the components of X cause Y, or that X is a

direct determinant of Y. This is a version of the well-known phrase 'correlation does not imply causation'.

Further steps need to be taken to assess the role that these statistical associations should play in a reasonable explanation, given the particulars of the case being considered.

Sense-check correlations and identify legitimate determining factors in a case-by-case manner

Next, you need to determine which of the statistical associations that the model's results have identified as important are legitimate and reasonably explanatory in the case you are considering. The challenge here is that there is no simple technical tool you can use to do this.

The model's prediction and classification results are observational rather than experimental, and they have been designed to minimise error rather than to be informative about causal structures. This means it is difficult to draw out an explanation.

You will therefore need to interpret and analyse which correlations and associations are consequential for providing a meaningful explanation. You can do this by drawing on your knowledge of the domain you are working in, and the decision recipient's specific circumstances.

This should help you do two things:

- Sense-check which correlations are relevant to an explanation. This involves not only ensuring that these correlations are not spurious or caused by hidden variables, but also determining how applicable the statistical generalisations are to the affected individual's specific circumstances.

  For example, a job candidate, who has spent several years in a full-time family care role, has been eliminated by an AI model because it identifies a strong statistical correlation between long periods of unemployment and poor work performance. This suggests that the correlation identified may not reasonably apply in this case. If such an outcome were challenged, the model's implementer would have to sense-check whether such a correlation should play a significant role given the decision recipient's particular circumstances. They would

also have to consider how other factors should be weighed in justifying that outcome.

- Identifying relevant determining factors involves picking out the features and interactions that could reasonably make a real-world difference when considering how they contribute to the outcome, as it specifically applies to the decision recipient under consideration.

  For example, a model predicts that a patient has a high chance of developing lung cancer in their lifetime. The features and interactions that have significantly contributed to this prediction include family history. The doctor knows that the patient is a non-smoker and has a family history of lung cancer, and concludes that, given risks arising from shared environmental and genetic factors, family history should be considered as a strong determinant in this patient's case.

Integrate your chosen correlations and outcome determinants into your reasoning

The final step involves integrating the correlations you have identified as most relevant into your reasoning. You should consider how this particular set of factors that influenced the model's result, combined with the specific context of the decision recipient, can support your overall conclusion on the outcome.

Similarly, implementers should be able to make their reasoning explicit and intelligible to affected individuals. Decision recipients should be able to easily understand how the statistical result has been applied to their particular case, and why the implementer assessed the outcome as they did. This could be through a plain-language explanation, or any other format they require to be able to make sense of the decision.

Step 5: Prepare implementers to deploy your AI system

At a glance

- In cases where decisions are not fully automated, implementers need to be meaningfully involved.
- This means that they need to be appropriately trained to use the model's results responsibly and fairly.
- Their training should cover:
    - the basics of how machine learning works;
    - the limitations of AI and automated decision-support technologies;
    - the benefits and risks of deploying these systems, particularly how they help humans come to judgements rather than replacing that judgement; and
    - how to manage cognitive biases.
- Where decisions are fully automated and provide a result directly to the decision recipient, you should set up the AI system to provide understandable explanations.

Checklist

☐ Where there is a 'human in the loop' we have trained our implementers to:

☐ Understand the associations and correlations that link the input data to the model's prediction or classification.

☐ Interpret which correlations are consequential for providing a meaningful explanation by drawing on their domain knowledge or the decision recipient's specific circumstances.

☐ Combine the chosen correlations and outcome determinants with what they know of the individual affected to come to their conclusion.

☐ Apply the AI model's results to the individual case at hand, rather than uniformly across decision recipients.

☐ Where our decision-making is fully automated, we have made sure that our AI system is set up to provide understandable explanations to individuals.

In more detail

Introduction

When human decision-makers are meaningfully involved in deploying an AI-assisted outcome (ie where the decision is not fully automated), you should make sure that they have been appropriate trained and prepared to use your model's results responsibly and fairly.

Implementer training

Implementer training should therefore include conveying basic knowledge about the statistical and probabilistic character of machine learning, and about the limitations of AI and automated decision-support technologies. This training should avoid any anthropomorphic (or human-like) portrayals of AI systems. It should also encourage the implementers to view the benefits and risks of deploying these systems in terms of their role in helping humans come to judgements, rather than replacing that judgement.

Further, training should address any cognitive or judgemental biases that may occur when implementers use AI systems in different settings. This should be based on the use-case, highlighting, for example, where over-reliance or over-compliance with the results of computer-based system can occur (known as automation bias). Cognitive biases may include overconfidence in a prediction based on the historical consistency of data, illusions that any clustering of data points necessarily indicates significant

insights, and discounting social patterns that exist beyond the statistical result.

Individuals are likely to expect that decisions produced about them do not treat them in terms of demographic probabilities and statistics. Inferences that are drawn from a model's results should therefore be applied to the particular circumstances of the decision recipient.

Fully automated systems

While it is usually safer to have a trained human to translate the result of an AI system for the affected individual, in many cases these processes will be automated, in which case you will have to ensure the AI system is set up to provide understandable explanations to individuals.

Step 6: Consider contextual factors when you deliver your explanation

At a glance

- Several contextual factors will have an effect on the purpose for which an individual wishes to use an explanation, and on how you should deliver your explanation. The factors are the:
  - domain you work in;
  - impact on the individual;
  - data used;
  - urgency of the decision; and ;
  - audience it is being presented to.

Checklist

☐ We have considered the contextual factors that affect what a decision recipient will find useful in an explanation.

☐ We have formulated all of our explanation types in a way that is most useful for the decision recipient, taking into account any reasonable adjustments.

In more detail

Introduction

The previous steps have shown you how to gather the information you need for each explanation type, and given further details on extracting rationale explanations in particular. However, there are also several factors relating to the context within which an AI-assisted decision is made that have an effect on the type of explanation which people will find useful and the purposes they wish to use it for.

From the primary research we carried out, particularly with members of the public, we identified five key contextual factors affecting why people want explanations of AI-assisted decisions. These contextual factors are set out below, along with suggestions of which explanations to prioritise in delivering an explanation of an AI-assisted decision given the factor.

Domain factor

What is this factor?

By 'domain', we mean the setting or the sector in which you deploy your AI model to help you make decisions about people. This can affect the explanations people want. For instance, what people want to know about AI-assisted decisions made in the criminal justice domain can differ significantly from other domains such as healthcare.

Likewise, domain or sector specific explanation standards can affect what people expect out of an explanation. For example, a person receiving an AI-assisted mortgage decision will expect to learn about the reasoning behind the determination in a manner that accords with established lending standards and practices.

Which explanations should we prioritise?

Considering the domain factor is perhaps the most crucial determiner of what explanations should be included and prioritised when communicating with affected individuals. If your AI system is operating in a safety-critical setting, decision recipients will obviously want appropriate safety and performance explanations. However, if your system is operating in a domain where bias and discrimination concerns are prevalent, they will likely want you to provide a fairness explanation.

In lower-stakes domains such as e-commerce, it is unlikely that people will, on average, want or expect extensive explanations of the fairness or performance of the outputs of recommender systems. Even so, in these

lower impact domains, provisions should be made for explaining the basic rationale and responsibility components (as well as all other relevant explanation types) of any decision system that affects people.

For example 'low' impact applications such as product recommendations and personalisation - eg of advertising or content - may give rise to sensitivities around targeting particular demographics, or ignoring others (eg advertising leadership roles targeted at men) raise obvious issues of fairness and impact on society, increasing the importance of explanations addressing these issues.

Impact factor

What is this factor?

The 'impact' factor is about the effect an AI-assisted decision can have on an individual and wider society. Varying levels of severity and different types of impact can change what explanations people will find useful, and the purpose the explanation serves.

Are the decisions safety-critical, relating to life or death situations (most often in the healthcare domain)? Do the decisions affect someone's liberty or legal status? Is the impact of the decision less severe but still significant – eg denial of a utility or targeting of a political message? Or is the impact more trivial – eg being directed to a specific ticket counter by an AI system that sorts queues in an airport?

Which explanations should we prioritise?

In general, where an AI-assisted decision has a high impact on an individual, explanations such as fairness, safety and performance, and impact are often important, because individuals want to be reassured as to the safety of the decision, to trust that they are being treated fairly, and to understand the consequences.

However, the rationale and responsibility explanations can be equally as important depending on the other contextual factors at play – for instance if the features of the data used by the AI model are changeable, or the inferences drawn are open to interpretation and can be challenged.

Considering impact as a contextual factor is not straightforward. There is no hard and fast rule. It should be done on a case by case basis, and considered in combination with all the other contextual factors.

Data factor

What is this factor?

'Data' as a contextual factor relates to both the data used to train and test your AI model, as well as the input data at the point of the decision. The type of data used by your AI model can influence an individual's willingness to accept or contest an AI-assisted decision, and the actions they take as a result of it.

This factor suggests that you should think about the nature of the data your model is trained on and uses as inputs for its outputs when it is deployed. You should consider whether the data is biological or physical (eg biomedical data used for research and diagnostics), or if it is social data that relates to demographic characteristics or measurements of human behaviour.

You should also consider whether an individual can change the outcome of a decision. If the factors that go into your decision are ones that can be influenced by changes to someone's behaviour or lifestyle, it is more likely that individuals that don't agree with the outcome may want to make such changes.

For example, if a bank loan decision was made based on a customer's financial activity, the customer may want to alter their spending behaviour to change that decision in the future. This will affect the type of explanation an individual wants. However, if the data is less flexible, such as biophysical data, it will be less likely that an individual will disagree with the output of the AI system. For example in healthcare, an output that is produced by an AI system on a suggested diagnosis based on genetic data about a patient is more 'fixed' – this is not something the patient can easily change.

Which explanations should we prioritise?

It will often be useful to prioritise the rationale explanation, for both social data and biophysical data. Where social data is used, individuals receiving an unfavourable decision can understand the reasoning and learn from this to appropriately adapt their behaviour for future decisions. For biophysical data, this can help people understand why a decision was made about them.

However, where biophysical data is used such as in medical diagnoses, individuals often prefer to simply know what the decision outcome means for them, and to be reassured about the safety and reliability of the decision. In these cases it makes sense to prioritise the impact and safety and performance explanations to meet these needs.

On the other hand, where the nature of the data is social, or subjective, individuals are more likely to have concerns about what data was taken into account for the decision, and the suitability or fairness of this in influencing an AI-assisted decision about them. In these circumstances, the data and fairness explanations will help address these concerns by telling people what the input data was, where it was from, and what measures you put in place to ensure that using this data to make AI-assisted decisions does not result in bias or discrimination.

Urgency factor

What is this factor?

The 'urgency' factor concerns the importance of receiving, or acting upon the outcome of an AI-assisted decision within a short timeframe. What people want to know about a decision can change depending on how little or much time they have to reflect on it.

The urgency factor recommends that you give thought to how urgent the AI-assisted decision is. Think about whether or not a particular course of action is often necessary after the kind of decisions you make, and how quickly that action needs to be taken.

Which explanations should we prioritise?

Where urgency is a key factor in the context within which your AI-assisted decision is made, it is more likely that individuals will want to know what the consequences are for them, and to be reassured that the AI model helping to make the decision is safe and reliable. As such, the impact and safety and performance explanations are suitable in these cases. This is because these explanations will help individuals to understand how the decision affects them, what happens next, and what measures and testing were implemented to maximise and monitor the safety and performance of the AI model.

Audience factor

What is this factor?

'Audience' as a contextual factor is about the individuals to whom you are explaining an AI-assisted decision. The groups of people you make decisions about, and the individuals within those groups have an effect on what type of explanations are meaningful or useful for them.

What level of expertise (eg about AI) do they have in relation to what the decision is about? Are a broad range of people subject to decisions you make (eg the UK general public), thus indicating that there might also be a broad range of knowledge or expertise? Or are the people you make decisions about limited to a smaller subset (eg your employees), suggesting they may be more informed on the things you are making decisions about? Also consider the decision recipients require any reasonable adjustments in how they receive the explanation (Equality Act 2010).

Which explanations should we prioritise?

If the people about whom you are making AI-assisted decisions are likely to have some domain expertise, you might consider using the rationale explanation. This is because you can be more confident that they can understand the reasoning and logic of an AI model, or a particular decision, due to being more familiar with the topic of the decisions. Additionally, if people subject to your AI-assisted decisions have some technical expertise, or are likely to be interested in the technical detail underpinning the decision, the safety and performance explanation will help.

Alternatively, where you think it's often likely the people you make AI-assisted decisions about do not have any specific expertise or knowledge about either the topic of the decision, or its technical aspects, other explanation types such as responsibility, or particular aspects of the safety and performance explanation may be more helpful. This is so that people can be reassured about the safety of the system, and know who to contact to query or question an AI decision.

Of course, even for those with little knowledge of an area about which an AI-assisted decision is made, the rationale explanation can still be useful to help illuminate the reasons why a decision was made in plain and simple terms. But there may also be occasions where the data used and inferences drawn by an AI model are particularly complex (see the 'data' factor above), and individuals would rather delegate the rationale explanation to a relevant domain expert to review and come to their own informed conclusions about the validity or suitability of the reasons for the decision (eg a doctor in a healthcare setting).

Step 7: Consider how to present your explanation

At a glance

- How you present your explanation depends on the way in which you make AI-assisted decisions, and on how people might expect you to deliver explanations you make without using AI.
- You can 'layer' your explanation by proactively providing individuals first with the explanations you have prioritised, and making additional explanations available in further layers. This helps to avoid information (or explanation) overload.
- You should think of delivering your explanation as a conversation, as opposed to a one-way process. People should be able to discuss a decision with a competent human being.
- Providing your explanation at the right time is also important.
- To increase trust and awareness of your use of AI, you can proactively engage with your customers by making information available about how you use AI systems to help you make decisions.

Checklist

☐ We have presented our explanation in a layered way, giving the most relevant explanation type(s) upfront, and providing the other types in additional layers.

☐ We have made it clear how decision recipients can contact us if they would like to discuss the AI-assisted decision with a human being.

☐ We have provided the decision recipient with the process-based and relevant outcome-based explanation for each explanation type, in advance of making a decision.

☐ We have proactively made information about our use of AI available in order to build trust with our customers and stakeholders.

In more detail

Introduction

You should determine the most appropriate method of delivery based on the way in which you make AI-assisted decisions about people, and how they might expect you to deliver explanations of decisions you make without using AI. This might be verbally, face to face, in hard-copy or electronic format. Think about any reasonable adjustments you might need to make for people under the Equality Act 2010. The timing for delivery of explanations will also affect the way you deliver the explanation.

If you deliver explanations in hard-copy or electronic form, you may also wish to consider whether there are design choices that can help make what you're telling people more clear and easy to understand. For example, in addition to text, simple graphs and diagrams may help explain certain explanations such as rationale and safety and performance. Depending on the size and resources of your organisation, you may be able to draw on the expertise of user experience and user interface designers.

Layering explanations

Based on the guidance we've provided above, and engagement with industry, we think it makes sense to build a 'layered' explanation.

By layered we mean proactively providing individuals with the prioritised explanations (the first layer), and making the additional explanations available on a second, and possibly third, layer. If you deliver your explanation on a website, you can use expanding sections, tabs, or simply link to webpages with the additional explanations.

The purpose of this layered approach is to avoid information (or explanation) fatigue. It means you won't overload people. Rather, they are provided with what is likely to be the most relevant and important information, while still having clear and easy access to other explanatory information, should they wish to know more about the AI decision.

Explanation as a dialogue

However you choose to deliver your explanations to individuals, it is important to think of this exercise as a conversation as opposed to a one-way process. By providing the priority explanations, you are the initiating a conversation, not ending it. Individuals should not only have easy access to additional explanatory information (hence layered explanations), but they should also be able to discuss the AI-assisted decision with a human being. This ties in with the responsibility explanation and having a human reviewer. However, as well as being able to contest decisions, it's important to provide a way for people to talk about and clarify explanations with a competent human being.

Explanation timing

It is important to provide explanations of AI-assisted decisions to individuals at the right time.

Delivering an explanation is not just about telling people the result of an AI decision. It is equally about telling people how decisions are made in advance.

What explanation can I provide in advance?

In Step 2 we provided two categories for each type of explanation: process-based and outcome-based.

You can provide the process-based explanations in advance of a specific decision. In addition, there will be some outcome-based explanations that you can provide in advance, particularly those related to:

- Responsibility - who is responsible for taking the decision that is supported by the result of the AI system, for reviewing and for implementing it;
- Impact – how you have assessed the potential impact of the model on the individual and the wider community; and
- Data - what data was input into the AI system to train, test, and validate it.

There will also be some situations when you can provide the same explanation in advance of a decision as you would afterwards. This is

because in some sectors it is possible to run a simulation of the model's output. For example, if you applied for a loan some organisations could explain the computation and tell you which factors matter in determining whether or not your application would be accepted. In cases like this, the distinction between explanations before and after a decision is less important. However, in many situations this won't be the case.

What should I do?

After you have prioritised the explanations (see Step 1), you should provide the relevant process-based explanations before the decision, and the outcome-based explanations if you are able to.

What explanation can I provide after a decision?

You can provide the full explanation after the decision, however there are some specific outcome-based explanations that you will not have been able to explain in advance - ie rationale, fairness, and safety and performance of the system, which are specific to a particular decision and are likely to be queried after the decision has been made. These explain the underlying logic of the system that led to the specific decision or output, whether the decision recipient was treated fairly compared with others who were similar, and whether the system functioned properly in that particular instance.

> The basics of explaining AI

Example

In this example, clinicians are using an AI system to help them detect cancer.

Example: Explanations in health care - cancer diagnosis		
Before decision	**Process-based explanation**	Responsibility – who is responsible for ensuring the AI system used in detecting cancer works in the intended way. Rationale – what steps you have taken to ensure that the components or measurements used in the model make sense for detecting cancer and can be made understandable to

		affected patients. Fairness – what measures you have taken to ensure the model is fair, prevents discrimination and mitigates bias (this may be less relevant here where biophysical data is being used). Safety and performance – what measures you have taken to ensure the model chosen to detect cancer is secure, accurate, reliable and robust, and how it has been tested, verified and validated. Impact – what measures you have taken to ensure that the AI model does not negatively impact the patient in how it has been designed or used. Data – how you have ensured that the source(s), quantity, and quality of the data used to train the system is appropriate for the type(s) of cancer detection for which you are utilising your model.
	Outcome-based explanation	Responsibility – who is responsible for taking the diagnosis resulting from the AI system's output, implementing it, and providing an explanation for how the diagnosis came about, and who the patient can go to in order to query the diagnosis. Impact – how the design and use of the AI system in the particular case of the patient will impact the patient. For example, if the system detects cancer but the result is a false positive, this could have a significant impact on the mental health of the patient. Data – the patient's data that will be used in this particular instance.

After decision	**Outcome-based explanation**	Rationale – whether the AI system's output, ie what it has detected as being cancerous or not, makes sense in the case of the patient, given the doctor's domain knowledge. Fairness – whether the model has produced results consistent with those it has produced for other patients with similar characteristics. Safety and performance – how secure, accurate, reliable and robust the AI model has been in the patient's particular case, and which safety and performance measures were used to test this.

Why is this important?

Not only is this a good way to provide an explanation to an individual when they might need it, it is also a way to comply with the law.

Articles 13-14 of the GDPR require that you proactively provide individuals with '…meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject…' in the case of solely automated decisions with a legal or similarly significant effect.

Article 15 of the GDPR also gives individuals a right to obtain this information at any time on request.

This is also good practice for systems where there is a 'human in the loop'.

The process-based and outcome-based explanations relating to the rationale of the AI system, and the outcome-based explanation relating to the AI system's impact on the individual, fulfil this requirement of the GDPR.

It is up to you and your organisation to determine the most appropriate way to deliver the explanations you choose to provide.

However, you might consider what the most direct and helpful way would be to deliver explanations that you can provide in advance of a decision. You should consider where individuals are most likely to go to find an explanation or information on how you make decisions with support of AI systems.

You should think about using the same platform for providing an advance explanation that you will use to provide the ultimate decision. This means that the information that an individual needs is in one place. You should also ensure that the explanation is prominent, to make it easier for individuals to find it.

Proactive engagement

How can we build trust?

Proactively making information available about how you use AI systems to help you make decisions is a good way to increase awareness among your customers. This will help them know more about when and why you use an AI system and how it works.

By being open and inclusive in how you share this information, you can increase the trust your customers have in how you operate, and build confidence in your organisation using AI to help them get a better service.

In the primary research we conducted, we found that the public is looking for more engagement from organisations and awareness raising about how they use AI for decision-making. By being proactive, you can use this engagement to help you fulfil the principle of being transparent.

What should we proactively share?

Among the things you could consider sharing are the following:

- What is AI?

This helps to demystify the technologies involved. It might be useful to outline these technologies, and provide a couple of examples of where AI is used in your sector.

> A good example is this animation about machine learning produced by researchers at the University of Oxford.

> ‘What is Machine Learning?’ animation

- How can it be used for decision-making?

  This outlines the different ways AI is useful for supporting decision-making – this tells people what the tools do. You could provide some examples of how you use it to help you make decisions.

- What are the benefits?

  This should lay out how AI can be beneficial, specifically for the individuals that are affected by the decisions you make. For example, if you are a service provider, you can outline how it can personalise your services so that your customers can get a better experience. The benefits you outline could also explore ways that the AI tools available can be better than more traditional decision-support tools. Examples could help you to make this clear.

- What are the risks?

  You should be honest about how AI can go wrong in your sector, for example how it can lead to discrimination or misinformation, and how you will mitigate this. This helps to set people’s expectations about what AI can do in their situation, and helps them understand what your organisation will do to look after them.

  You should also provide information about people’s rights under the GDPR, for example the right to object or challenge the use of AI, and the right to obtain human review or intervention.

- Why does our organisation use AI for decisions?

  This should clearly and comprehensively outline why you have chosen to use AI systems in your particular organisation. It should expand on the more general examples you have provided above for how it improves the service you offer compared with other approaches (if applicable), and what the benefits are for your customers.

- Where/when do we do this?

Here you can describe which parts of your organisation and in which parts of the decision-making process you are using AI. You should make this as informative as possible. You could also outline what measures you have put in place to ensure that the AI system you are using in each of these areas is designed in a way to maximise the benefits and minimise the risks. In particular, you should be clear about whether there is a 'human in the loop' or whether the AI is solely automated. In addition, it might be helpful to show how you are managing the system's use to make sure it is maximising the interests of your customers.

- Who can I speak to about it?

  You could provide an email address or helpline for interested members of the public to contact in order to get more information on how you are using AI. Those answering these queries should have good knowledge of AI and how you are using it, and be able to explain it in a clear, open and accessible way. The amount of detail you provide should be proportionate to the information people ask for.

How should we share this?

There are many different ways you could proactively share information with your customers and stakeholders:

- Your usual communications to customers and stakeholders, such as regular newsletters or customer information.
- Providing a link to a dedicated part of your organisation's website outlining the sections above.
- Flyers and leaflets distributed in your offices and to those of other relevant or partner organisations.
- An information campaign or other initiative in partnership with other organisations.
- Information you distribute through trade bodies.

Your communications team will have an important role to play in making sure the information is targeted and relevant to your customers.

The ICO has written guidance on the right to be informed, which will help you with this communication task.

> Guidance on the right to be informed (GDPR)

Annexe 1: Example of building and presenting an explanation of a cancer diagnosis

Bringing together our guidance, the following example shows how a healthcare organisation could use the steps we have outlined to help them structure the process of building and presenting their explanation to an affected patient.

Example: Explanations in healthcare - cancer diagnosis

Step 1: Select priority explanation types by considering the domain, use case and impact on the individuals

First, the healthcare organisation familiarises itself with the explanation types in this guidance. Based on the healthcare setting and the impact of the cancer diagnosis on the patient's life, the healthcare organisation selects the explanation types that it determines are a priority to provide to patients subject to its AI-assisted decisions. It documents its justification for these choices:

Priority explanation types:

Rationale – Justifying the reasoning behind the outcome of the AI system to maintain accountability, and useful for patients if visualisation techniques of AI explanation are available for non-experts…

Impact – Due to high impact (life/death) situation, important for patients to understand effects and next steps…

Responsibility – Non-expert audience likely to want to know who to query the AI system's output with…

Safety and performance - Given data and domain complexity, this may help reassure patients about the accuracy, safety and reliability of the AI system's output…

Other explanation types:

Data – Simple detail on input data…

Fairness – Less important due to use of biophysical data, as opposed to social or demographic data, but still relevant in areas such as data representativeness…

The healthcare organisation formalises these explanation types in the relevant part of its policy on information governance:

Information governance policy…

…Use of AI…

…Explaining AI decisions to patients…

…Types of explanations:

- Rationale…
- Impact…
- Responsibility…
- Safety and Performance…
- Data…
- Fairness…

Step 2: Collect the information you need for each explanation type

The explanation types the healthcare organisation has chosen each has a process-based and outcome-based explanation. The quality of each explanation is also influenced by how they collect and prepare the training and test data for the AI model they choose. They therefore collect the following information for each explanation type:

Rationale

Process-based explanation: information to show that the AI system has been set up in a way that enables explanations of its underlying logic to be extracted (directly or using supplementary tools); and that these explanations are meaningful for the patients concerned.

Outcome-based explanation: information on the logic behind the model's results and on how implementers have incorporated that logic into their

decision-making. This includes how the system transforms input data into outputs, how this is translated into language that is understandable to patients, and how the medical team uses the model's results in reaching a diagnosis for a particular case.

Data collection and pre-processing: information on how data has been labelled and how that shows the reasons for classifying, for example, certain images as tumours.

Responsibility

Process-based explanation: information on those responsible within the healthcare organisations for managing the design and use of the AI model, and how they ensured the model was responsibly managed throughout its design and use.

Outcome-based explanation: information on those responsible for taking the output reached by the AI system, implementing the output into diagnosis, reviewing it, and providing explanations for how the diagnosis came about (ie who the patient can go to in order to query the diagnosis).

Data collection and pre-processing: information on who or which part of the healthcare organisation is responsible for collecting and pre-processing the patient's data. Being transparent about the process can help the healthcare organisation to build trust and confidence in their use of AI.

Safety and performance

Process-based explanation: information on the measures taken to ensure the overall safety and technical performance (security, accuracy, reliability, and robustness) of the AI model—including information about the testing, verification, and validation done to certify these.

Outcome-based explanation: Information on the safety and technical performance (security, accuracy, reliability, and robustness) of the AI model in its actual operation, eg information confirming that the model operated securely and according to its intended design in the specific patient's case. This could include the safety and performance measures used.

Data collection and pre-processing: information on the accuracy rate of the model and the accuracy-related measures the healthcare organisation used.

Impact

Process-based explanation: measures taken across the AI model's design and use to ensure that it does not negatively impact the wellbeing of the patient.

Outcome-based explanation: information on the actual impacts of the AI system on the patient.

Data collection and pre-processing: the data the healthcare organisation uses has a bearing on its impact and risk assessment.

Step 3: Build your rationale explanation to provide meaningful information about the underlying logic of your AI system

The healthcare organisation decides to use an artificial neural network to sequence and extract information from radiologic images. While this model is able to predict the existence and types of tumours, the high-dimensional character of its processing makes it opaque.

The model's design team has chosen supplementary 'salience mapping' and 'class activation mapping' tools to help them visualise the critical regions of the images that are indicative of malign tumours. These tools render the trouble-areas visible by highlighting the abnormal regions. Such mapping-enhanced images then allow technicians and radiologists to gain a clearer understanding of the clinical basis of the AI model's cancer prediction.

Step 4: Translate the rationale of your system's results into useable and easily understandable reasons

The AI system the hospital uses to detect cancer produces a result, which is a prediction that a particular area on an MRI scan contains a cancerous growth. This prediction comes out as a probability, with a particular level of confidence, measured as a percentage. The supplementary mapping tools subsequently provide the radiologist with a visual representation of the cancerous region.

The radiologist shares this information with the oncologist and other doctors on the medical team along with other detailed information about the performance measures of the system and its certainty levels.

For the patient, the oncologist or other members of the medical team then put this into language, or another format, that the patient can understand. One way the doctors choose to do this is through visually showing the patient the scan and supplementary visualisation tools to help explain the model's result. Highlighting the areas that the AI system has flagged is an intuitive way to help the patient understand what is happening. The doctors also indicate how much confidence they have in the AI system's result based on its performance and uncertainty metrics.

Step 5: Prepare implementers to deploy your AI system

Because the technician and oncologist are both using the AI system in their work, the hospital decides they need training in how to use the system.

Implementer training covers:

- how they should interpret the results that the AI system generates, based on understanding how it has been designed and the data it has been trained on;
- how they should understand and weigh the performance and certainty limitations of the system (ie how they view and interpret confusion matrices, confidence intervals, error bars, etc);
- that they should use the result as one part of their decision-making, as a complement to their existing domain knowledge;
- that they should critically examine whether the AI system's result is based on appropriate logic and rationale; and
- that in each case they should prepare a plan for communicating the AI system's result to the patient, and the role that result has played in the doctor's judgement. This includes any limitations in using the system.

Step 6: Consider the contextual factors when you deliver your explanation

The healthcare organisation considers what contextual factors are likely to have an effect on what patients want to know about the AI-assisted decisions it plans to make on a cancer diagnosis. It draws up a list of the relevant factors:

Contextual factors:

Domain – regulated, safety testing…
Data – biophysical…
Urgency – if cancer, urgent…
Impact – high, safety-critical…
Audience – mostly non-expert…

Step 7: Consider how to present your explanation

The healthcare organisation develops a template for delivering their explanation of AI decisions about cancer diagnosis in a layered way:

Layer 1

- Rationale explanation
- Impact explanation
- Responsibility explanation
- Safety and Performance explanation

Delivery – eg the clinician provides explanation face to face with patient & leaflet.

Layer 2

- Data explanation
- Fairness explanation

Delivery – eg the clinician gives the patient a leaflet.

Appendix 2: Further reading

Resources for Exploring Algorithm Types

General

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85. http://thuvien.thanglong.edu.vn:8081/dspace/bitstream/DHTL_123456789/4053/1/%5BSpringer%20Series%20in%20Statistics-1.pdf

Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206. https://www.nature.com/articles/s42256-019-0048-x

Regularised regression (LASSO and Ridge)

Gaines, B. R., & Zhou, H. (2016). Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, *27*(4), 861-871. https://arxiv.org/pdf/1611.01511.pdf

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. http://beehive.cs.princeton.edu/course/read/tibshirani-jrssb-1996.pdf

Generalised linear model (GLM)

https://CRAN.R-project.org/package=glmnet

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software, 33(1), 1-22. http://www.jstatsoft.org/v33/i01/

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). *Regularization paths for Cox's proportional hazards model via coordinate descent*. Journal of Statistical Software, 39(5), 1-13. URL http://www.jstatsoft.org/v39/i05/

Generalised additive model (GAM)

https://CRAN.R-project.org/package=gam

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 150-158). ACM. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.8241&rep=rep1&type=pdf

Wood, S. N. (2006). Generalized additive models: An introduction with R. CRC Press.

Decision tree (DT)

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

Rule/decision lists and sets

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, *18*(1), 8753-8830. http://www.jmlr.org/papers/volume18/17-716/17-716.pdf

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016, August). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675-1684). ACM. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5108651/

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350-1371. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1446488742

Wang, F., & Rudin, C. (2015). Falling rule lists. In *Artificial Intelligence and Statistics* (pp. 1013-1022). http://proceedings.mlr.press/v38/wang15a.pdf

Case-based reasoning (CBR)/ Prototype and criticism

Aamodt, A. (1991). A knowledge-intensive, integrated approach to problem solving and sustained learning. *Knowledge Engineering and Image Processing Group. University of Trondheim*, 27-85.

http://www.dphu.org/uploads/attachements/books/books_4200_0.pdf

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, *7*(1), 39-59. https://www.idi.ntnu.no/emner/tdt4171/papers/AamodtPlaza94.pdf

Bichindaritz, I., & Marling, C. (2006). Case-based reasoning in the health sciences: What's next?. *Artificial intelligence in medicine*, *36*(2), 127-135. http://cs.oswego.edu/~bichinda/isc471-hci571/AIM2006.pdf

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, *5*(4), 2403-2424. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1324399600

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems* (pp. 2280-2288). http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf

MMD-critic in python: https://github.com/BeenKim/MMD-critic

Kim, B., Rudin, C., & Shah, J. A. (2014). The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems* (pp. 1952-1960). http://papers.nips.cc/paper/5313-the-bayesian-case-model-a-generative-approach-for-case-based-reasoning-and-prototype-classification.pdf

Supersparse linear integer model (SLIM)

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. *Available at SSRN 2919024*. https://arxiv.org/pdf/1702.04690.pdf

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *48*(5), 449-466. https://pdfs.semanticscholar.org/b3d8/8871ae5432c84b76bf53f7316cf5f95a3938.pdf

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349-391. https://link.springer.com/article/10.1007/s10994-015-5528-6

Optimized scoring systems for classification problems in python:
https://github.com/ustunb/slim-python

Simple customizable risk scores in python:
https://github.com/ustunb/risk-slim

Resources for exploring supplementary explanation strategies

Surrogate models (SM)

Bastani, O., Kim, C., & Bastani, H. (2017). Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*.
https://obastani.github.io/docs/fatml17.pdf

Craven, M., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems* (pp. 24-30). http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf

Van Assche, A., & Blockeel, H. (2007). Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In *European Conference on Machine Learning* (pp. 418-429). Springer, Berlin, Heidelberg.
https://link.springer.com
/content/pdf/10.1007/978-3-540-74958-5_39.pdf

Valdes, G., Luna, J. M., Eaton, E., Simone II, C. B., Ungar, L. H., & Solberg, T. D. (2016). MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Scientific reports*, *6*, 37854.
https://www.nature.com/articles/srep37854

Partial Dependence Plot (PDP)

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
https://projecteuclid.org/download/pdf_1/euclid.aos/1013203451

Greenwell, B. M. (2017). pdp: an R Package for constructing partial dependence plots. *The R Journal*, *9*(1), 421-436.
https://pdfs.semanticscholar.org/cdfb
/164f55e74d7b116ac63fc6c1c9e9cfd01cd8.pdf

For the software in R: https://cran.r-project.org/web/packages/pdp/index.html

Individual Conditional Expectations Plot (ICE)

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44-65. https://arxiv.org/pdf/1309.6392.pdf

For the software in R see:
https://cran.r-project.org/web/packages/ICEbox/index.html
https://cran.r-project.org/web/packages/ICEbox/ICEbox.pdf

Accumulated Local Effects Plots (ALE)

Apley, D. W., & Zhu, J. (2019). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*. https://arxiv.org/pdf
/1612.08468;Visualizing

https://cran.r-project.org/web/packages/ALEPlot/index.html

Global Variable Importance

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. https://link.springer.com
/content/pdf/10.1023/A:1010933404324.pdf

Casalicchio, G., Molnar, C., & Bischl, B. (2018, September). Visualizing the feature importance for black box models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 655-670). Springer, Cham. https://arxiv.org/pdf
/1804.06620.pdf

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. arXiv:1801.01489

Fisher, A., Rudin, C., & Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. *arXiv preprint arXiv:1801.01489*. https://arxiv.org/abs/1801.01489v2

Hooker, G., & Mentch, L. (2019). Please Stop Permuting Features: An Explanation and Alternatives. *arXiv preprint arXiv:1905.03151*. https://arxiv.org/pdf/1905.03151.pdf

Zhou, Z., & Hooker, G. (2019). Unbiased Measurement of Feature Importance in Tree-Based Methods. *arXiv preprint arXiv:1903.05179*. https://arxiv.org/pdf/1903.05179.pdf

Global Variable Interaction

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916-954. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1223908046

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*. https://arxiv.org/pdf/1805.04755.pdf

Hooker, G. (2004, August). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 575-580). ACM. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.7500&rep=rep1&type=pdf

Local Interpretable Model-Agnostic Explanation (LIME)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM. https://arxiv.org/pdf/1602.04938.pdf?mod=article_inline

LIME in python: https://github.com/marcotcr/lime
LIME experiments in python: https://github.com/marcotcr/lime-experiments

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.7500&rep=rep1&type=pdf

Anchors in python: https://github.com/marcotcr/anchor
Anchors experiments in python: https://github.com/marcotcr/anchor-experiments

Shapley Additive ExPlanations (SHAP)

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing*

Systems (pp. 4765-4774). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Software for SHAP and its extensions in python:

https://github.com/slundberg/shap

R wrapper for SHAP: https://modeloriented.github.io/shapper/

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307-317. http://www.library.fa.ru/files/Roth2.pdf#page=39

Counterfactual Explanation

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076). http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*(pp. 10-19). ACM. https://arxiv.org/pdf/1809.06514.pdf

Evaluate recourse in linear classification models in python:
https://github.com/ustunb/actionable-recourse

Secondary Explainers and Attention-Based Systems

Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17082/16552

Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2016). Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*. https://arxiv.org/pdf/1612.04757

Other resources for supplementary explanation

IBM's Explainability 360: http://aix360.mybluemix.net

Biecek, B., & Burzykowski, T. (2019). *Predictive Models: Explore, Explain, and Debug, Human-Centered Interpretable Machine Learning*. Retrieved from

https://pbiecek.github.io
/PM_VEE/

Accompanying software, Dalex, Descriptive mAchine Learning Explanations:
https://github.com/ModelOriented/DALEX

Przemysław Biecek, *Interesting resources related to XAI*:
https://github.com/pbiecek/xai_resources

Christoph Molnar, iml: Interpretable machine learning
https://cran.r-project.org/web/packages/iml/index.html

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [15]) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for two supervised models: One trained to detect smiling faces in images, and one trained to detect toxic comments in text. We propose model cards as a step towards the responsible democratization of machine learning and related artificial intelligence technology, increasing transparency into how well artificial intelligence technology works. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed evaluation numbers and other relevant documentation.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Social and professional topics** → *User characteristics*; • **Software and its engineering** → **Use cases**; *Documentation*; *Software evolution*; • **Human-centered computing** → *Walkthrough evaluations*;

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01.
https://doi.org/10.1145/3287560.3287596

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

1 INTRODUCTION

Currently, there are no standardized documentation procedures to communicate the performance characteristics of trained machine learning (ML) and artificial intelligence (AI) models. This lack of documentation is especially problematic when models are used in applications that have serious impacts on people's lives, such as in health care [14, 42, 44], employment [1, 13, 29], education [23, 45] and law enforcement [2, 7, 20, 34].

Researchers have discovered systematic biases in commercial machine learning models used for face detection and tracking [4, 9, 49], attribute detection [5], criminal justice [10], toxic comment detection [11], and other applications. However, these systematic errors were only exposed after models were put into use, and negatively affected users reported their experiences. For example, after MIT Media Lab graduate student Joy Buolamwini found that commercial face recognition systems failed to detect her face [4], she collaborated with other researchers to demonstrate the disproportionate errors of computer vision systems on historically marginalized groups in the United States, such as darker-skinned women [5, 41]. In spite of the potential negative effects of such reported biases, documentation accompanying trained machine learning models (if supplied) provide very little information regarding model performance characteristics, intended use cases, potential pitfalls, or other information to help users evaluate the suitability of these systems to their context. This highlights the need to have detailed documentation accompanying trained machine learning models, including metrics that capture bias, fairness and inclusion considerations.

As a step towards this goal, we propose that released machine learning models be accompanied by short (one to two page) records we call model cards. Model cards (for model reporting) are complements to "Datasheets for Datasets" [21] and similar recently proposed documentation paradigms [3, 28] that report details of the datasets used to train and test machine learning models. Model cards are also similar to the TRIPOD statement proposal in medicine [25]. We provide two example model cards in Section 5: A smiling detection model trained on the CelebA dataset [36] (Figure 2), and a public toxicity detection model [32] (Figure 3). Where Datasheets highlight characteristics of the data feeding into the model, we

focus on trained model characteristics such as the type of model, intended use cases, information about attributes for which model performance may vary, and measures of model performance.

We advocate for measures of model performance that contain quantitative evaluation results to be broken down by individual cultural, demographic, or phenotypic groups, domain-relevant conditions, and intersectional analysis combining two (or more) groups and conditions. In addition to model evaluation results, model cards should detail the motivation behind chosen performance metrics, group definitions, and other relevant factors. Each model card could be accompanied with Datasheets [21], Nutrition Labels [28], Data Statements [3], or Factsheets [27], describing datasets that the model was trained and evaluated on. Model cards provide a way to inform users about what machine learning systems can and cannot do, the types of errors they make, and additional steps that could create more fair and inclusive outcomes with the technology.

2 BACKGROUND

Many mature industries have developed standardized methods of benchmarking various systems under different conditions. For example, as noted in [21], the electronic hardware industry provides datasheets with detailed characterizations of components' performances under different test conditions. By contrast, despite the broad reach and impact of machine learning models, there are no standard stress tests that are performed on machine learning based systems, nor standardized formats to report the results of these tests. Recently, researchers have proposed standardized forms of communicating characteristics of datasets used in machine learning [3, 21, 28] to help users understand the context in which the datasets should be used. We focus on the complementary task for machine learning models, proposing a standardized method to evaluate the performance of human-centric models: Disaggregated by unitary and intersectional groups such as cultural, demographic, or phenotypic population groups. A framework that we refer to as "Model Cards" can present such evaluation supplemented with additional considerations such as intended use.

Outside of machine learning, the need for population-based reporting of outcomes as suggested here has become increasingly evident. For example, in vehicular crash tests, dummies with prototypical female characteristics were only introduced after researchers discovered that women were more likely than men to suffer serious head injuries in real-world side impacts [18]. Similarly, drugs developed based on results of clinical trials with exclusively male participants have led to overdosing in women [17, 50]. In 1998, the U.S. Food and Drug Administration mandated that clinical trial results be disaggregated by groups such as age, race and gender [16].

While population-based analyses of errors and successes can be provided for unitary groups such as "men", "women", and "non-binary" gender groups, they should also be provided intersectionally, looking at two or more characteristics such as gender and age simultaneously. Intersectional analyses are linked to intersectionality theory, which describes how discrete experiences associated with characteristics like race or gender in isolation do not accurately reflect their interaction [8]. Kimberlé Crenshaw, who pioneered intersectional research in critical race theory, discusses the story of Emma DeGraffenreid, who was part of a failed lawsuit against General Motors in 1976, claiming that the company's hiring practices discriminated against Black women. In their court opinion, the judges noted that since General Motors hired many women for secretarial positions, and many Black people for factory roles, they could not have discriminated against Black women. However, what the courts failed to see was that only White women were hired into secretarial positions and only Black men were hired into factory roles. Thus, Black women like Emma DeGraffenreid had no chance of being employed at General Motors. This example highlights the importance of intersectional analyses: empirical analyses that emphasize the interaction between various demographic categories including race, gender, and age.

Before further discussing the details of the model card, it is important to note that at least two of the three characteristics discussed so far, race and gender, are socially sensitive. Although analyzing models by race and gender may follow from intersectionality theory, how "ground truth" race or gender categories should be labeled in a dataset, and whether or not datasets should be labeled with these categories at all, is not always clear. This issue is further confounded by the complex relationship between gender and sex. When using cultural identity categories such as race and gender to subdivide analyses, and depending on the context, we recommend either using datasets with self-identified labels or with labels clearly designated as *perceived* (rather than self-identified). When this is not possible, datasets of public figures with known public identity labels may be useful. Further research is necessary to expand how groups may be defined, for example, by automatically discovering groups with similarities in the evaluation datasets.

3 MOTIVATION

As the use of machine learning technology has rapidly increased, so too have reports of errors and failures. Despite the potentially serious repercussions of these errors, those looking to use trained machine learning models in a particular context have no way of understanding the systematic impacts of these models before deploying them.

The proposal of "Model Cards" specifically aims to standardize ethical practice and reporting - allowing stakeholders to compare candidate models for deployment across not only traditional evaluation metrics but also along the axes of ethical, inclusive, and fair considerations. This goes further than current solutions to aid stakeholders in different contexts. For example, to aid policy makers and regulators on questions to ask of a model, and known benchmarks around the suitability of a model in a given setting.

Model reporting will hold different meaning to those involved in different aspects of model development, deployment, and use. Below, we outline a few use cases for different stakeholders:

- **ML and AI practitioners** can better understand how well the model might work for the intended use cases and track its performance over time.
- **Model developers** can compare the model's results to other models in the same space, and make decisions about training their own system.
- **Software developers** working on products that use the model's predictions can inform their design and implementation decisions.

- **Policymakers** can understand how a machine learning system may fail or succeed in ways that impact people.
- **Organizations** can inform decisions about adopting technology that incorporates machine learning.
- **ML-knowledgeable individuals** can be informed on different options for fine-tuning, model combination, or additional rules and constraints to help curate models for intended use cases without requiring technical expertise.
- **Impacted individuals** who may experience effects from a model can better understand how it works or use information in the card to pursue remedies.

Not only does this practice improve model understanding and help to standardize decision making processes for invested stakeholders, but it also encourages forward-looking model analysis techniques. For example, slicing the evaluation across groups functions to highlight errors that may fall disproportionately on some groups of people, and accords with many recent notions of mathematical fairness (discussed further in the example model card in Figure 2). Including group analysis as part of the reporting procedure prepares stakeholders to begin to gauge the fairness and inclusion of future outcomes of the machine learning system. Thus, in addition to supporting decision-making processes for determining the suitability of a given machine learning model in a particular context, model reporting is an approach for responsible transparent and accountable practices in machine learning.

People and organizations releasing models may be additionally incentivized to provide model card details because it helps potential users of the models to be better informed on which models are best for their specific purposes. If model card reporting becomes standard, potential users can compare and contrast different models in a well-informed way. Results on several different evaluation datasets will additionally aid potential users, although evaluation datasets suitable for disaggregated evaluation are not yet common. Future research could include creating robust evaluation datasets and protocols for the types of disaggregated evaluation we advocate for in this work, for example, by including differential privacy mechanisms [12] so that individuals in the testing set cannot be uniquely identified by their characteristics.

4 MODEL CARD SECTIONS

Model cards serve to disclose information about a trained machine learning model. This includes how it was built, what assumptions were made during its development, what type of model behavior different cultural, demographic, or phenotypic population groups may experience, and an evaluation of how well the model performs with respect to those groups. Here, we propose a set of sections that a model card should have, and details that can inform the stakeholders discussed in Section 3. A summary of all suggested sections is provided in Figure 1.

The proposed set of sections below are intended to provide relevant details to consider, but are not intended to be complete or exhaustive, and may be tailored depending on the model, context, and stakeholders. Additional details may include, for example, interpretability approaches, such as saliency maps, TCAV [33], and Path-Integrated Gradients [38, 43]); stakeholder-relevant explanations (e.g., informed by a careful consideration of philosophical, psychological, and other factors concerning what is as a good explanation in different contexts [22]); and privacy approaches used in model training and serving.

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.

4.1 Model Details

This section of the model card should serve to answer basic questions regarding the model version, type and other details.

Person or organization developing model: What person or organization developed the model? This can be used by all stakeholders to infer details pertaining to model development and potential

conflicts of interest.
Model date: When was the model developed? This is useful for all stakeholders to become further informed on what techniques and data sources were likely to be available during model development.
Model version: Which version of the model is it, and how does it differ from previous versions? This is useful for all stakeholders to track whether the model is the latest version, associate known bugs to the correct model versions, and aid in model comparisons.
Model type: What type of model is it? This includes basic model architecture details, such as whether it is a Naive Bayes classifier, a Convolutional Neural Network, etc. This is likely to be particularly relevant for software and model developers, as well as individuals knowledgeable about machine learning, to highlight what kinds of assumptions are encoded in the system.
Paper or other resource for more information: Where can resources for more information be found?
Citation details: How should the model be cited?
License: License information can be provided.
Feedback on the model: E.g., what is an email address that people may write to for further information?

There are cases where some of this information may be sensitive. For example, the amount of detail corporations choose to disclose might be different from academic research groups. This section should not be seen as a requirement to compromise private information or reveal proprietary training techniques; rather, a place to disclose basic decisions and facts about the model that the organization can share with the broader community in order to better inform on what the model represents.

4.2 Intended Use

This section should allow readers to quickly grasp what the model should and should not be used for, and why it was created. It can also help frame the statistical analysis presented in the rest of the card, including a short description of the user(s), use-case(s), and context(s) for which the model was originally developed. Possible information includes:
Primary intended uses: This section details whether the model was developed with general or specific tasks in mind (e.g., plant recognition worldwide or in the Pacific Northwest). The use cases may be as broadly or narrowly defined as the developers intend. For example, if the model was built simply to label images, then this task should be indicated as the primary intended use case.
Primary intended users: For example, was the model developed for entertainment purposes, for hobbyists, or enterprise solutions? This helps users gain insight into how robust the model may be to different kinds of inputs.
Out-of-scope uses: Here, the model card should highlight technology that the model might easily be confused with, or related contexts that users could try to apply the model to. This section may provide an opportunity to recommend a related or similar model that was designed to better meet that particular need, where possible. This section is inspired by warning labels on food and toys, and similar disclaimers presented in electronic datasheets. Examples include "not for use on text examples shorter than 100 tokens" or "for use on black-and-white images only; please consider our research group's full-color-image classifier for color images."

4.3 Factors

Model cards ideally provide a summary of model performance across a variety of relevant factors including *groups*, *instrumentation*, and *environments*. We briefly describe each of these factors and their relevance followed by the corresponding prompts in the model card.

4.3.1 Groups. "Groups" refers to distinct categories with similar characteristics that are present in the evaluation data instances. For human-centric machine learning models, "groups" are people who share one or multiple characteristics. Intersectional model analysis for human-centric models is inspired by the sociological concept of intersectionality, which explores how an individual's identity and experiences are shaped not just by unitary personal characteristics – such as race, gender, sexual orientation or health – but instead by a complex combination of many factors. These characteristics, which include but are not limited to cultural, demographic and phenotypic categories, are important to consider when evaluating machine learning models. Determining which groups to include in an intersectional analysis requires examining the intended use of the model and the context under which it may be deployed. Depending on the situation, certain groups may be more vulnerable than others to unjust or prejudicial treatment.

For human-centric computer vision models, the visual presentation of age, gender, and Fitzpatrick skin type [15] may be relevant. However, this must be balanced with the goal of preserving the privacy of individuals. As such, collaboration with policy, privacy, and legal experts is necessary in order to ascertain which groups may be responsibly inferred, and how that information should be stored and accessed (for example, using differential privacy [12]).

Details pertaining to groups, including who annotated the training and evaluation datasets, instructions and compensation given to annotators, and inter-annotator agreement, should be provided as part of the data documentation made available with the dataset. See [3, 21, 28] for more details.

4.3.2 Instrumentation. In addition to groups, the performance of a model can vary depending on what instruments were used to capture the input to the model. For example, a face detection model may perform differently depending on the camera's hardware and software, including lens, image stabilization, high dynamic range techniques, and background blurring for portrait mode. Performance may also vary across real or simulated traditional camera settings such as aperture, shutter speed and ISO. Similarly, video and audio input will be dependent on the choice of recording instruments and their settings.

4.3.3 Environment. A further factor affecting model performance is the environment in which it is deployed. For example, face detection systems are often less accurate under low lighting conditions or when the air is humid [51]. Specifications across different lighting and moisture conditions would help users understand the impacts of these environmental factors on model performance.

4.3.4 Card Prompts. We propose that the Factors section of model cards expands on two prompts:

Relevant factors: What are foreseeable salient factors for which model performance may vary, and how were these determined?
Evaluation factors: Which factors are being reported, and why were these chosen? If the relevant factors and evaluation factors are different, why? For example, while Fitzpatrick skin type is a relevant factor for face detection, an evaluation dataset annotated by skin type might not be available until reporting model performance across groups becomes standard practice.

4.4 Metrics

The appropriate metrics to feature in a model card depend on the type of model that is being tested. For example, classification systems in which the primary output is a class label differ significantly from systems whose primary output is a score. In all cases, the reported metrics should be determined based on the model's structure and intended use. Details for this section include:
Model performance measures: What measures of model performance are being reported, and why were they selected over other measures of model performance?
Decision thresholds: If decision thresholds are used, what are they, and why were those decision thresholds chosen? When the model card is presented in a digital format, a threshold slider should ideally be available to view performance parameters across various decision thresholds.
Approaches to uncertainty and variability: How are the measurements and estimations of these metrics calculated? For example, this may include standard deviation, variance, confidence intervals, or KL divergence. Details of how these values are approximated should also be included (e.g., average of 5 runs, 10-fold cross-validation).

4.4.1 Classification systems. For classification systems, the error types that can be derived from a confusion matrix are *false positive rate*, *false negative rate*, *false discovery rate*, and *false omission rate*. We note that the relative importance of each of these metrics is system, product and context dependent.

For example, in a surveillance scenario, surveillors may value a low false negative rate (or the rate at which the surveillance system fails to detect a person or an object when it should have). On the other hand, those being surveilled may value a low false positive rate (or the rate at which the surveillance system detects a person or an object when it should not have). We recommend listing all values and providing context about which were prioritized during development and why.

Equality between some of the different confusion matrix metrics is equivalent to some definitions of fairness. For example, equal false negative rates across groups is equivalent to fulfilling Equality of Opportunity, and equal false negative and false positive rates across groups is equivalent to fulfilling Equality of Odds [26].

4.4.2 Score-based analyses. For score-based systems such as pricing models and risk assessment algorithms, describing differences in the distribution of measured metrics across groups may be helpful. For example, reporting measures of central tendency such as the mode, median and mean, as well as measures of dispersion or variation such as the range, quartiles, absolute deviation, variance and standard deviation could facilitate the statistical commentary necessary to make more informed decisions about model development. A model card could even extend beyond these summary statistics to reveal other measures of differences between distributions such as cross entropy, perplexity, KL divergence and pinned area under the curve (pinned AUC) [11].

There are a number of applications that do not appear to be score-based at first glance, but can be considered as such for the purposes of intersectional analysis. For instance, a model card for a translation system could compare BLEU scores [40] across demographic groups, and a model card for a speech recognition system could compare word-error rates. Although the primary outputs of these systems are not scores, looking at the score differences between populations may yield meaningful insights since comparing raw inputs quickly grows too complex.

4.4.3 Confidence. Performance metrics that are disaggregated by various combinations of instrumentation, environments and groups makes it especially important to understand the confidence intervals for the reported metrics. Confidence intervals for metrics derived from confusion matrices can be calculated by treating the matrices as probabilistic models of system performance [24].

4.5 Evaluation Data

All referenced datasets would ideally point to any set of documents that provide visibility into the source and composition of the dataset. Evaluation datasets should include datasets that are publicly available for third-party use. These could be existing datasets or new ones provided alongside the model card analyses to enable further benchmarking. Potential details include:
Datasets: What datasets were used to evaluate the model?
Motivation: Why were these datasets chosen?
Preprocessing: How was the data preprocessed for evaluation (e.g., tokenization of sentences, cropping of images, any filtering such as dropping images without faces)?

To ensure that model cards are statistically accurate and verifiable, the evaluation datasets should not only be representative of the model's typical use cases but also anticipated test scenarios and challenging cases. For instance, if a model is intended for use in a workplace that is phenotypically and demographically homogeneous, and trained on a dataset that is representative of the expected use case, it may be valuable to evaluate that model on two evaluation sets: one that matches the workplace's population, and another set that contains individuals that might be more challenging for the model (such as children, the elderly, and people from outside the typical workplace population). This methodology can highlight pathological issues that may not be evident in more routine testing.

It is often difficult to find datasets that represent populations outside of the initial domain used in training. In some of these situations, synthetically generated datasets may provide representation for use cases that would otherwise go unevaluated [35]. Section 5.2 provides an example of including synthetic data in the model evaluation dataset.

4.6 Training Data

Ideally, the model card would contain as much information about the training data as the evaluation data. However, there might be cases where it is not feasible to provide this level of detailed information about the training data. For example, the data may be proprietary, or require a non-disclosure agreement. In these cases, we advocate for basic details about the distributions over groups in the data, as well as any other details that could inform stakeholders on the kinds of biases the model may have encoded.

4.7 Quantitative Analyses

Quantitative analyses should be *disaggregated*, that is, broken down by the chosen factors. Quantitative analyses should provide the results of evaluating the model according to the chosen metrics, providing confidence interval values when possible. Parity on the different metrics across disaggregated population subgroups corresponds to how *fairness* is often defined [37, 48]. Quantitative analyses should demonstrate the metric variation (e.g., with error bars), as discussed in Section 4.4 and visualized in Figure 2.
The disaggregated evaluation includes:
Unitary results: How did the model perform with respect to each factor?
Intersectional results: How did the model perform with respect to the intersection of evaluated factors?

4.8 Ethical Considerations

This section is intended to demonstrate the ethical considerations that went into model development, surfacing ethical challenges and solutions to stakeholders. Ethical analysis does not always lead to precise solutions, but the process of ethical contemplation is worthwhile to inform on responsible practices and next steps in future work.

While there are many frameworks for ethical decision-making in technology that can be adapted here [19, 30, 46], the following are specific questions you may want to explore in this section:
Data: Does the model use any sensitive data (e.g., protected classes)?
Human life: Is the model intended to inform decisions about matters central to human life or flourishing – e.g., health or safety? Or could it be used in such a way?
Mitigations: What risk mitigation strategies were used during model development?
Risks and harms: What risks may be present in model usage? Try to identify the potential recipients, likelihood, and magnitude of harms. If these cannot be determined, note that they were considered but remain unknown.
Use cases: Are there any known model use cases that are especially fraught? This may connect directly to the intended use section of the model card.

If possible, this section should also include any additional ethical considerations that went into model development, for example, review by an external board, or testing with a specific community.

4.9 Caveats and Recommendations

This section should list additional concerns that were not covered in the previous sections. For example, did the results suggest any further testing? Were there any relevant groups that were not represented in the evaluation dataset? Are there additional recommendations for model use? What are the ideal characteristics of an evaluation dataset for this model?

5 EXAMPLES

We present worked examples of model cards for two models: an image-based classification system and a text-based scoring system.

5.1 Smiling Classifier

To show an example of a model card for an image classification problem, we use the public CelebA dataset [36] to examine the performance of a trained "smiling" classifier across both age and gender categories. Figure 2 shows our prototype.

These results demonstrate a few potential issues. For example, the false discovery rate on older men is much higher than that for other groups. This means that many predictions incorrectly classify older men as smiling when they are not. On the other hand, men (in aggregate) have a higher false negative rate, meaning that many of the men that are in fact smiling in the photos are incorrectly classified as not smiling.

The results of these analyses give insight into contexts the model might not be best suited for. For example, it may not be advisable to apply the model on a diverse group of audiences, and it may be the most useful when detecting the presence of a smile is more important than detecting its absence (for example, in an application that automatically finds 'fun moments' in images). Additional fine-tuning, for example, with images of older men, may help create a more balanced performance across groups.

5.2 Toxicity Scoring

Our second example provides a model card for Perspective API's TOXICITY classifier built to detect 'toxicity' in text [32], and is presented in Figure 3. To evaluate the model, we use an intersectional version of the open source, synthetically created Identity Phrase Templates test set published in [11]. We show two versions of the quantitative analysis: one for TOXICITY v. 1, the initial version of the this model, and one for TOXICITY v. 5, the latest version.

This model card highlights the drastic ways that models can change over time, and the importance of having a model card that is updated with each new model release. TOXICITY v. 1 has low performance for several terms, especially "lesbian", "gay", and "homosexual". This is consistent with what some users of the initial TOXICITY model found, as reported by the team behind Perspective API in [47]. Also in [47], the Perspective API team shares the bias mitigation techniques they applied to the TOXICITY v. 1 model, in order to create the more equitable performance in TOXICITY v. 5. By making model cards a standard part of API launches, teams like the Perspective API team may be able to find and mitigate some of these biases earlier.

6 DISCUSSION & FUTURE WORK

We have proposed frameworks called model cards for reporting information about what a trained machine learning model is and how well it works. Model cards include information about the context of the model, as well as model performance results disaggregated by different unitary and intersectional population groups. Model

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

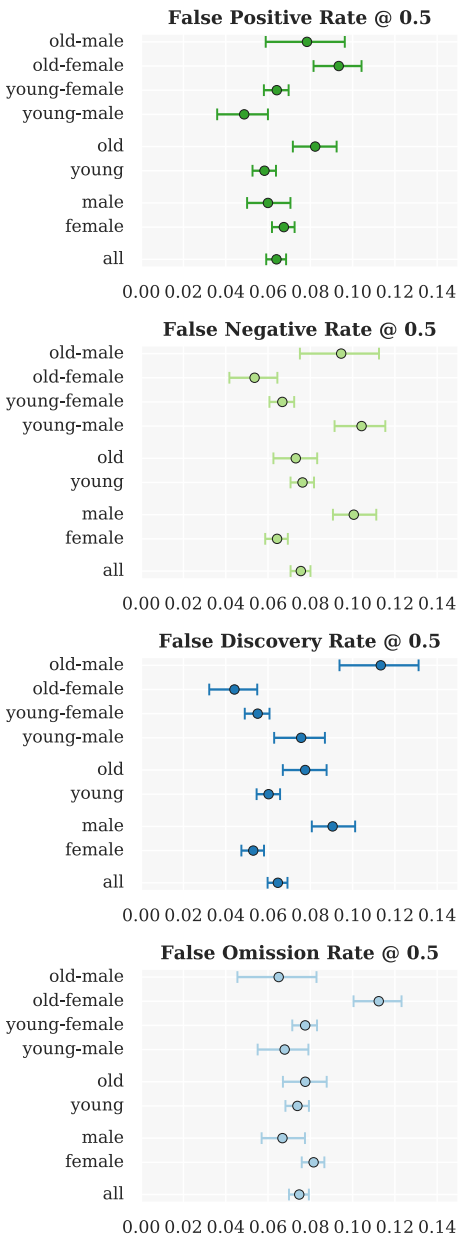
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses

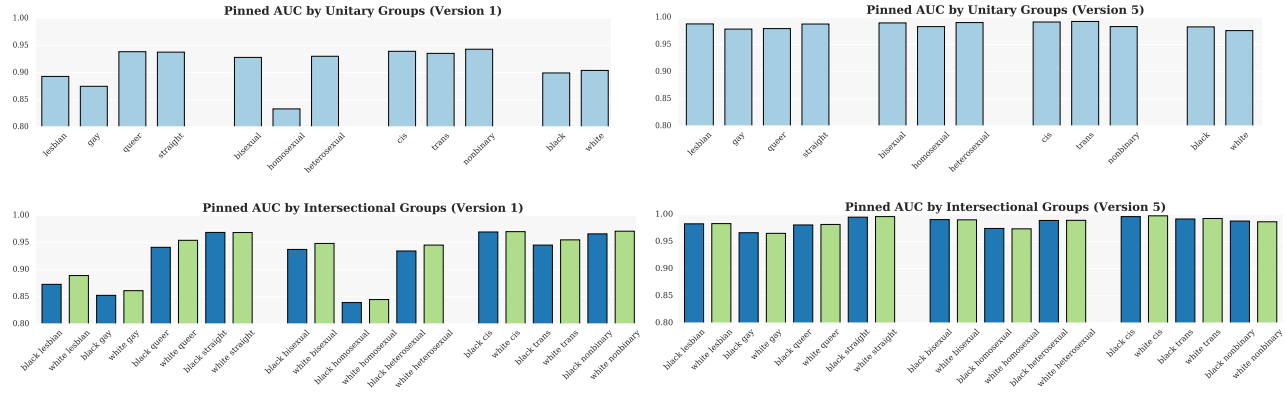


Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

cards are intended to accompany a model after careful review has determined that the foreseeable benefits outweigh the foreseeable risks in the model's use or release.

To demonstrate the use of model cards in practice, we have provided two examples: A model card for a smiling classifier tested on the CelebA dataset, and a model card for a public toxicity detector tested on the Identity Phrase Templates dataset. We report confusion matrix metrics for the smile classifier and Pinned AUC for the toxicity detector, along with model details, intended use, pointers to information about training and evaluation data, ethical considerations, and further caveats and recommendations.

The framework presented here is intended to be general enough to be applicable across different institutions, contexts, and stakeholders. It also is suitable for recently proposed requirements for analysis of algorithmic decision systems in critical social institutions, for example, for models used in determining government benefits, employment evaluations, criminal risk assessment, and criminal DNA analysis [39].

Model cards are just one approach to increasing transparency between developers, users, and stakeholders of machine learning models and systems. They are designed to be flexible in both scope and specificity in order to accommodate the wide variety of machine learning model types and potential use cases. Therefore the usefulness and accuracy of a model card relies on the integrity of the creator(s) of the card itself. It seems unlikely, at least in the near term, that model cards could be standardized or formalized to a degree needed to prevent misleading representations of model results (whether intended or unintended). It is therefore important to consider model cards as one transparency tool among many, which could include, for example, algorithmic auditing by third-parties (both quantitative and qualitative), "adversarial testing" by technical and non-technical analysts, and more inclusive user feedback mechanisms. Future work will aim to refine the methodology of creating model cards by studying how model information is interpreted and used by different stakeholders. Researchers should also explore how model cards can strengthen and complement other transparency methods

7 ACKNOWLEDGEMENTS

Thank you to Joy Buolamwini, Shalini Ananda and Shira Mitchell for invaluable conversations and insight.

REFERENCES

[1] Avrio AI. 2018. Avrio AI: AI Talent Platform. (2018). https://www.goavrio.com/
[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[3] Emily M. Bender and Batya Friedman. 2018. "Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science". *Transactions of the ACL (TACL)* (2018).
[4] Joy Buolamwini. 2016. How I'm fighting Bias in Algorithms. (2016). https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms#t-63664
[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html
[6] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[7] Federal Trade Commission. 2016. Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues. (2016). https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report
[8] Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.* (1989), 139.
[9] Black Desi. 2009. HP computers are racist. (2009). https://www.youtube.com/watch?v=t4DT3tQqgRM
[10] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. (2016). https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html
[11] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2018).
[12] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*, Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–19.
[13] Entelo. 2018. Recruitment Software | Entelo. (2018). https://www.entelo.com/
[14] Daniel Faggella. 2018. Follow the Data: Deep Learning Leads the Transformation of Enterprise - A Conversation with Naveen Rao. (2018).
[15] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.
[16] Food and Drug Administration. 1989. Guidance for the Study of Drugs Likely to Be Used in the Elderly. (1989).
[17] U.S. Food and Drug Administration. 2013. FDA Drug Safety Communication: Risk of next-morning impairment after use of insomnia drugs; FDA requires lower recommended doses for certain drugs containing zolpidem (Ambien, Ambien CR, Edluar, and Zolpimist). (2013). https://web.archive.org/web/20170428150213/https://www.fda.gov/drugs/drugsafety/ucm352085.htm
[18] IIHS (Insurance Institute for Highway Safety: Highway Loss Data Institute). 2003. Special Issue: Side Impact Crashworthiness. *Status Report* 38, 7 (2003).
[19] Institute for the Future, Omidyar Network's Tech, and Society Solutions Lab. 2018. Ethical OS. (2018). https://ethicalos.org/
[20] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. 2016. The Perpetual Line-Up. (2016). https://www.perpetuallineup.org/
[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). http://arxiv.org/abs/1803.09010
[22] Google. 2018. Responsible AI Practices. (2018). https://ai.google/education/responsible-ai-practices
[23] Gooru. 2018. Navigator for Teachers. (2018). http://gooru.org/about/teachers
[24] Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*. Springer, 345–359.
[25] Collins GS, Reitsma JB, Altman DG, and Moons KM. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. *Annals of Internal Medicine* 162, 1 (2015), 55–63. DOI: http://dx.doi.org/10.7326/M14-0697
[26] Moritz Hardt, Eric Price, , and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf
[27] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R. Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. *CoRR* abs/1808.07261 (2018).
[28] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018). http://arxiv.org/abs/1805.03677
[29] Ideal. 2018. AI For Recruiting Software | Talent Intelligence for High-Volume Hiring. (2018). https://ideal.com/
[30] DrivenData Inc. 2018. An Ethics Checklist for Data Scientists. (2018). http://deon.drivendata.org/
[31] Jigsaw. 2017. Conversation AI Research. (2017). https://conversationai.github.io/
[32] Jigsaw. 2017. Perspective API. (2017). https://www.perspectiveapi.com/
[33] B. Kim, Wattenberg M., J. Gilmer, Cai C., Wexler J., , F. Viegas, and R. Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) . *ICML* (2018).
[34] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801. DOI: http://dx.doi.org/10.1109/TIFS.2012.2214212
[35] Der-Chiang Li, Susan C Hu, Liang-Sian Lin, and Chun-Wu Yeh. 2017. Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets. *PloS one* 12, 8 (2017), e0181853.

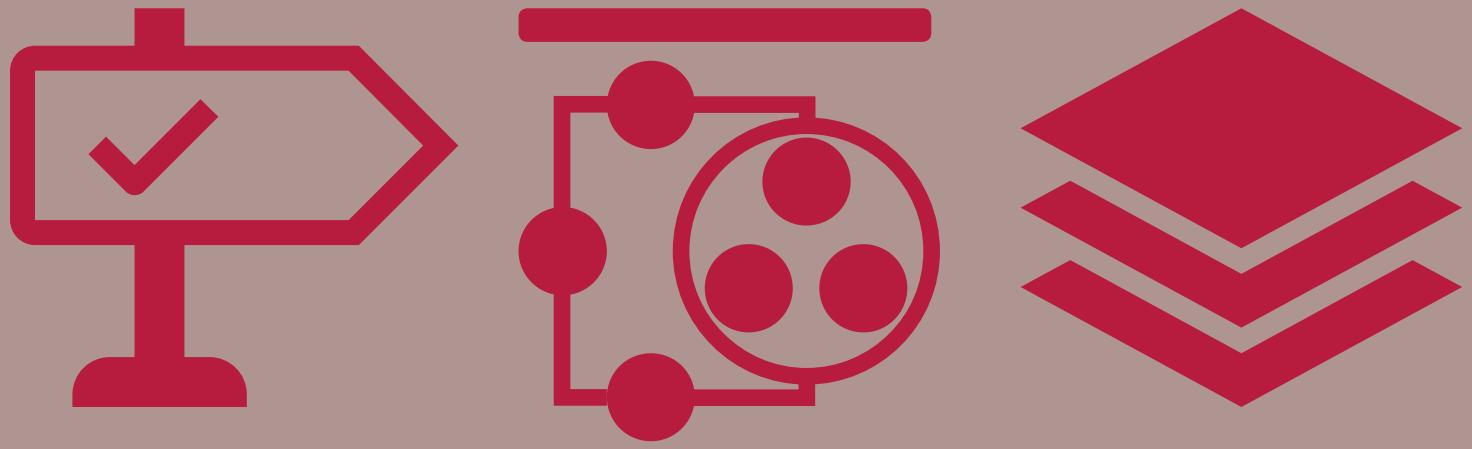
[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
[37] Shira Mitchell, Eric Potash, and Solon Barocas. 2018. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv:1811.07867* (2018).
[38] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the Model Understand the Question? *Proceedings of the Association for Computational Linguistics* (2018).
[39] AI Now. 2018. Litigating Algorithms: Challenging Government Use Of Algorithmic Decision Systems. AI Now Institute.
[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.
[41] Inioluwa Raji. 2018. Black Panther Face Scorecard: Wakandans Under the Coded Gaze of AI. (2018).
[42] Microsoft Research. 2018. Project InnerEye - Medical Imaging AI to Empower Clinicians. (2018). https://www.microsoft.com/en-us/research/project/medical-image-analysis/
[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. PMLR, Sydney, Australia.
[44] Digital Reasoning Systems. 2018. AI-Enabled Cancer Software | Healthcare AI : Digital Reasoning. (2018). https://digitalreasoning.com/solutions/healthcare/
[45] Turnitin. 2018. Revision Assistant. (2018). http://turnitin.com/en_us/what-we-offer/revision-assistant
[46] Shannon Vallor, Brian Green, and Irina Raicu. 2018. Ethics in Technology Practice: An Overview. (22 6 2018). https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/
[47] Lucy Vasserman, John Li, CJ Adams, and Lucas Dixon. 2018. Unintended bias and names of frequently targeted groups. *Medium* (2018). https://medium.com/the-false-positive/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23
[48] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. (2018).
[49] Joz Wang. 2010. Flickr Image. (2010). https://www.flickr.com/photos/jozjozjoz/3529106844
[50] Amy Westervelt. 2018. The medical research gender gap: how excluding women from clinical trials is hurting our health. (2018).
[51] Mingyuan Zhou, Haiting Lin, S Susan Young, and Jingyi Yu. 2018. Hybrid sensing face detection and registration for low-light and unconstrained conditions. *Applied optics* 57, 1 (2018), 69–78.

National Audit Office

National Audit Office

Framework to review models

Interactive



Contents

Go to previous page viewed

Introduction

Government Departments and Agencies routinely develop and use models to generate insight into business problems and business decisions. These models can vary in complexity from relatively simple spreadsheets, to detailed forecasts using specialist software. The outputs of these models and associated decisions can involve large amounts of money and resources.

Our report on forecasting in government to achieve value for money, identified weaknesses associated with forecasting in 71 NAO reports reviewed between 2010 and 2013. These weaknesses included:

- limited or poor quality data;
- unrealistic assumptions and optimism bias;
- a lack of forecasting or modelling; and
- inadequate sensitivity and scenario analysis.

This framework provides a structured approach to review models, which organisations can use to determine whether the modelling outputs they produce, are reasonable.

Evidence base

The framework to review models builds on the evidence and guidance available from:

- HM Treasury's 'review of quality assurance of government analytical models' (2013);
- HM Treasury's 'Aqua Book' (2015);
- The Department for Energy and Climate Change (DECC) 'Quality Assurance: Guidance for Models' (2014); and
- International Standard on Auditing 540 'auditing accounting estimates, including fair value accounting estimates, and related disclosures'.

Go to previous page viewed

And our experience of reviewing government models, for example:

- The Work Programme – review of a spreadsheet model projecting cost of welfare to work programme over lifetime of the contract.
- Long term public finance report projections – review of a non-transferable, internal, specialist actuarial model projecting public sector pensions over the next 50 years.
- Training new teachers – review of the published Teacher Supply Model, which estimates how many initial teacher training places are needed each year.

How to use the framework

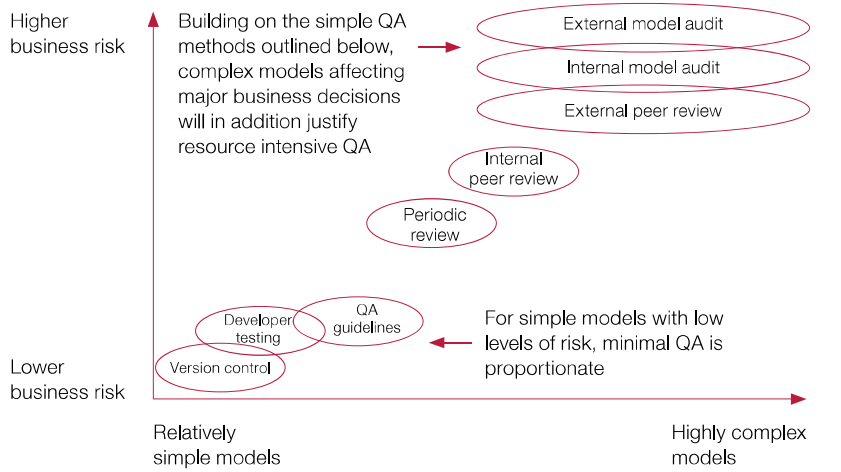
This framework is aimed at people who commission analysis, provide analytical assurance and deliver the analysis itself.

It is not intended to be a checklist, instead it is a flexible approach which can be tailored, based on:

- the amount of time and resource available;
- the complexity and risk associated with the model; and
- the level of assurance needed to reach an overall judgement.

This concept is in line with HM Treasury's 'review of quality assurance of government analytical models' (see diagram).

Schematic showing indicative types of quality assurance that might be expected given different levels of risk[1]

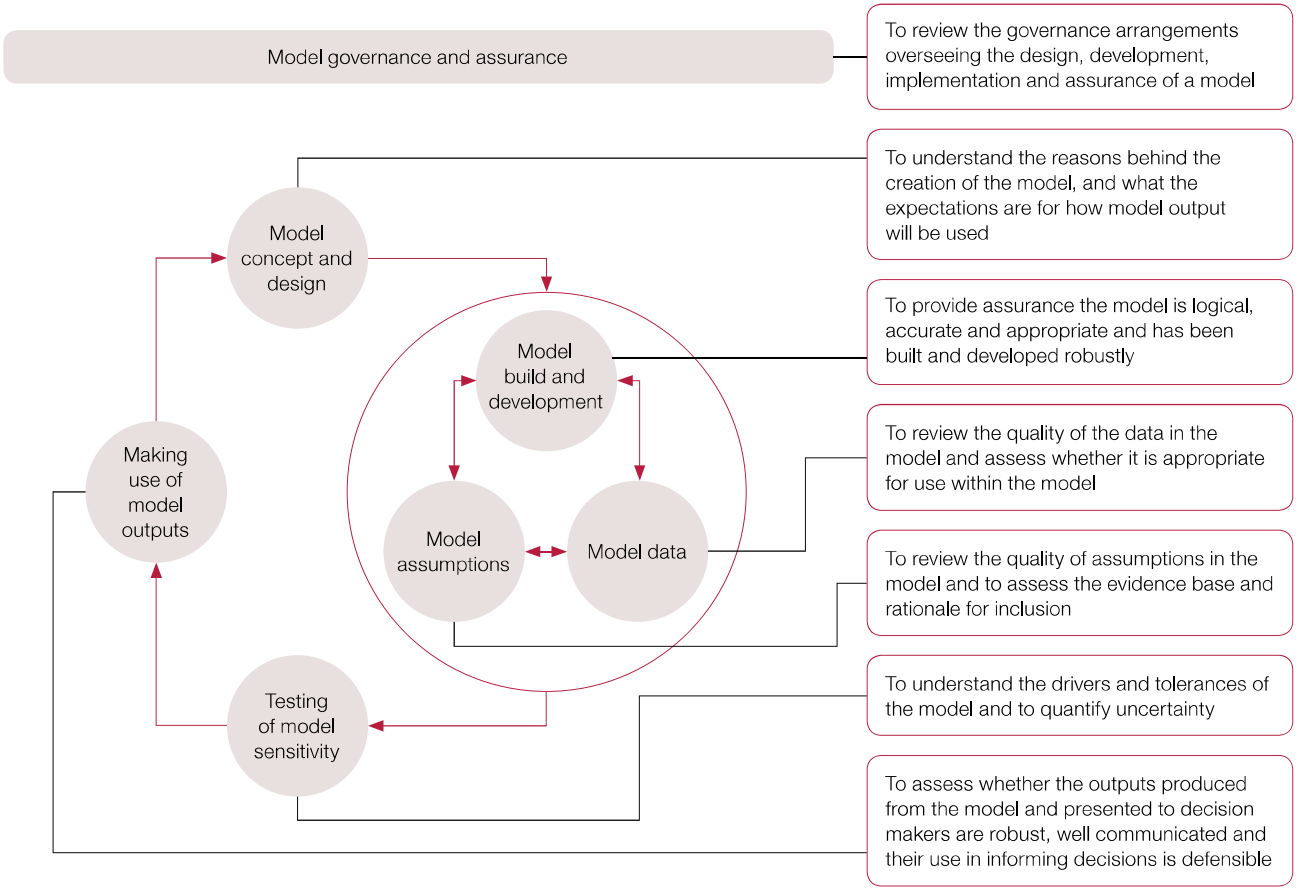


1 Chart 2.C, page 22, Review of quality assurance of Government analytical models: final report, HM Treasury (2013).

Go to previous page viewed

The framework is split into seven stages starting with the model concept and design, ending with making use of model outputs and all overseen by a governance and assurance structure[2] (see diagram).

Areas to consider when reviewing models



Go to previous page viewed

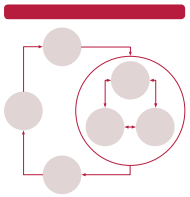
2 The questions in the framework are not exhaustive, meaning there will be other checks that can be applied.

Deciding on whether a model is robust and used appropriately to support business decisions requires a proportionate, evidence-based judgement. It will often be the case that a review will identify issues and weaknesses in some aspect of how the model was designed, built and used. Crucially, the objective of a model review is to identify, in your opinion, whether those issues had an impact on the quality of the model. And whether there is a risk it could materially impact on the outputs, how they are interpreted and used in decision making and risk management processes.

How the NAO can help

If you have any queries about this framework or suggestions for how it can be improved, please use the contact form and select Value for Money methodology.

The model

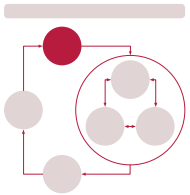


Model Governance and Assurance

To review the governance arrangements overseeing the design, development, implementation and assurance of a model.

Questions to consider	Examples of checks to make or evidence to look for
Who is the single Senior Responsible Owner (SRO) for the model?	Documentation of roles and responsibilities throughout the model development and use process.
Is the model 'business critical'?	Define was makes a model 'business critical'. Test this definition with definitions from other organisations. Evidence the Accounting Officer's governance statement (typically within the annual report) includes an appropriate quality assurance framework for business critical models. Evidence the Accounting Officer maintains an up to date list of business critical models and that this is publically available.
Does the model have good documentation on governance and assurance?	Are roles and responsibilities (i.e. commissioner, lead analyst, lead analytical assurer) documented? What processes are in place for succession planning/ handover, i.e. when a key person leaves the modelling project? Has the model been developed in collaboration with customers and/or stakeholders? For example: • Are requirements captured and documented into a specification? • Are assumptions listed and agreed? Is there an agreed quality assurance plan throughout the model development process? Is there evidence the customer of a model has influenced it to meet expectations?
How are model outputs challenged and used?	Is there a forum available for people outside the model development process to challenge the development and use of model outputs? How do model customers develop an understanding of the caveats of the model? Are model limitations and caveats reported alongside the main outputs of the model?

Go to previous page viewed



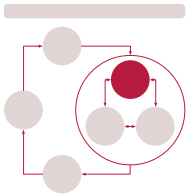
Model concept and design

To understand the reasons behind the creation of the model, and what the expectations are for how model output will be used.

Questions to consider	Examples of checks to make or evidence to look for
What is the decision the model is designed to support?	Identify who the stakeholders of the decision are. Consideration given to alternative solutions to support the decision. Was the model designed specifically to support this decision, or is an existing model being re-used? [If so, is this appropriate?]
Is there evidence of the rationale and the scoping of the model concept?	Documentation detailing the rationale, concept and structure of the model, such as: • what the model aims to replicate; • the input, output and model logic; • the model type (including options for alternative approaches which have been rejected); • the stakeholders responsible for policy and delivery; • the required precision (offset against complexity); and • identification of the limitations of the model.

Questions to consider	Examples of checks to make or evidence to look for
Is there a technical guide that demonstrates the logical flow of the model?	Compare the data flow, logic and structure in the model with the description in the technical guide. Are you able to understand the model?

Go to previous page viewed



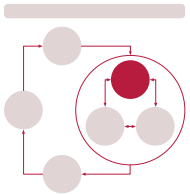
Model build and development

To provide assurance the model is logical, accurate and appropriate and has been built and developed robustly.

Questions to consider	Examples of checks to make or evidence to look for
Has the model been published?	If the model has not been published, identify the rationale for why not.
Do you understand the model?	Are you able to draw a simple picture representing the model or can you describe it in ley terms? Are inputs, calculations and outputs separate?
Does the model respond logically to basic changes being made to the model inputs?	Review how changing basic model inputs impact the model outputs, for example by: • simplifying settings to the most basic scenario; • examining the initial (starting) conditions for the model; • sensitivity analysis with realistic input variations; and • sensitivity analysis with extreme or implausible inputs variations.

Questions to consider	Examples of checks to make or evidence to look for
How accurate is the detail of the model?	Take sample checks to assess whether the model is doing what it should, for example by re-performing calculations on sections of the model. Consistency of accuracy and aggregation of the data. For Excel based models identify areas that might expose weaknesses in the model, such as: • circular reference warnings; • hard coding of values; • linking of data from other files; and • complexity of formulae. For syntax based models, review whether comments or notes explain what the element of the model is doing and whether it is understandable to someone unfamiliar with the model.

Go to previous page viewed

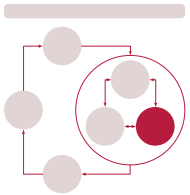


Model build and development *continued*

To provide assurance the model is logical, accurate and appropriate and has been built and developed robustly.

Questions to consider	Examples of checks to make or evidence to look for
How accurately does the model perform against historical data?	Review (or perform) checks assessing how the model predicts known history, both on data available during development and since implementation. For older models, use back casting to determine its 'forecasting' record.
Has the model been subject to external review during or after development?	Identify who has reviewed the model, and why. Review documentation produced by bodies reviewing the model. This is not limited to the building of the model and could cover any of the areas outlined in this framework. Identify whether there is an external assurance statement.
What documentation and processes are in place to ensure a corporate memory for the model exists?	Review how changes to the model, for example, detail of change, rationale and impact, are recorded. Review the adequacy of any model documentation (technical and non-technical) provided for new users, for details of what the model does and how to operate it.

Go to previous page viewed

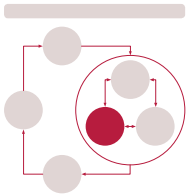


Model data

To review the quality of the data in the model and assess whether it is appropriate for use within the model.

Questions to consider	Examples of checks to make or evidence to look for
Is the data in the model of good quality?	Review the quality of data and sources, such as the extent to which data: • are up-to-date; • source is documented; • is based on a robust sample; • is consistent with other sources; and • meets the requirements it is being used for. Check data (as much as is practically feasible) in the model to the source data for accuracy. Does model documentation outline the limitations of the data? Where good quality data is lacking, what steps have been taken to work around this, for example making use of experts to provide estimates.
Is the data the model using, coming from other models?	Review whether separate models also need to be part of the scope of the model review.
What processes does the model use to handle input data?	Review how input data is included in the model, this could include considerations such as how data is cleaned or transformed from the original source, and how easily this is repeated when the model is refreshed. Check that data is applied consistently throughout the model.

Go to previous page viewed

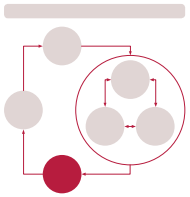


Model assumptions

To review the quality of assumptions in the model and to assess the evidence base and rationale for inclusion.

Questions to consider	Examples of checks to make or evidence to look for
Are the details of assumptions recorded and justified?	Identify and review list of assumptions, for example: • Suitability of selection based on the purpose of the model. • Underlying evidence – source and quality. • Level of simplification/complexity. • Rationale for level of accuracy and aggregation. • Distinction between data and structural assumptions. What are the main assumptions in the model?
What process is used to change/update assumptions?	Review the process for managing how assumptions are changed within the model. Review whether assumptions should have been updated in light of any changes to circumstances.
Have the status of the assumptions been critically compared to third party sources, or benchmarked against industry norms?	Check to similar models. Check to published standard assumptions.

Go to previous page viewed

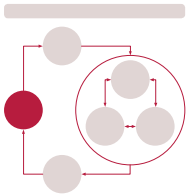


Testing of model sensitivity

To understand the drivers and tolerances of the model and to quantify uncertainty.

Questions to consider	Examples of checks to make or evidence to look for
What are the uncertainties of the model?	Review whether uncertainty has been quantified in the model (i.e. are high and low estimates provided alongside a point estimate?). Review whether the model estimates the level of confidence in the output. In the context of materiality, consider developing: • a list of modelling uncertainties; • a list of input data, evidence and intelligence used in the analysis and consider each type of uncertainty that could affect it; and • a diagram representing key parts of the model with consideration for what additional factors might act at that point and affect the analysis outcome.
Has sensitivity analysis been performed to calculate ranges or the likelihood of outcomes occurring?	Review whether levels used in sensitivity analysis are realistic and conservative based on the source data. Review or perform analysis such as Monte Carlo simulation or scenario analysis.
Do changes in the inputs/ assumptions have a material or significant impact on outputs?	Review or perform additional runs of the model to test sensitivities on outputs when the assumptions are changed. Review or perform additional runs of the model to test sensitivities on outputs when inputs are changed.
Have issues over poor quality data and assumptions and other identified risks been addressed?	Test for the impact of weak information in the model.

Go to previous page viewed



Making use of the output

To assess whether forecasts receive sufficient challenge, are integrated in to decision making and risk management systems and are compared with actual outcomes in order to inform future development.

Questions to consider	Examples of checks to make or evidence to look for
Are you able to validate model outputs?	Review appropriateness of model output by comparing to: • previous runs of the model; • other models such as parallel systems; and • independent sources.
What is the process for the routine review of outputs?	Review process for circulating outputs internally and externally, checks could involve different roles, for example: • Technical staff not directly involved with the model. • Senior staff responsible for the model. • External expertise.
Are the limitations and uncertainty of the model output communicated to decision makers?	Review how model outputs are presented to decision makers, for example how findings are presented in a business case.
Are decisions based on the model output proportionate to the robustness of the model?	Review whether decisions are appropriate and proportionate to the robustness of the model, for example considering monetary impact of decision given constraints of the model.
Are the outputs from the model responsive to the ongoing needs of the organisation?	Review whether the model is being used to track on-going performance as a monitoring tool.
Is the output from the model adjusted outside of the model?	Review whether any additional procedures or adjustments that are made to the model output are justified and how they impact on the robustness of decisions made.
Does the model output meet the requirements and aims of the model as outlined in the model concept?	Compare the actual outputs of the model with the aims of the concept model.
Are forecasts compared with actual outputs in order to validate the results and inform future development?	Compare the actual outputs with reality to check accuracy and check if this is used to update future iterations.

Go to previous page viewed

Contact

How the NAO can help

If you have any queries about this framework or suggestions for how it can be improved, please use the contact form and select Value for Money methodology.

The National Audit Office scrutinises public spending for Parliament and is independent of government. The Comptroller and Auditor General (C&AG), Sir Amyas Morse KCB, is an Officer of the House of Commons and leads the NAO, which employs some 810 people. The C&AG certifies the accounts of all government departments and many other public sector bodies. He has statutory authority to examine and report to Parliament on whether departments and the bodies they fund have used their resources efficiently, effectively, and with economy. Our studies evaluate the value for money of public spending, nationally and locally. Our recommendations and reports on good practice help government improve public services, and our work led to audited savings of £1.15 billion in 2014.

Authors
Elliott White and Thomas Jordan.

Go to previous page viewed

National Audit Office

Design and production by NAO Communications | DP Ref: 11018-002 |

Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing

Inioluwa Deborah Raji*
Partnership on AI
deb@partnershiponai.org

Andrew Smart*
Google
andrewsmart@google.com

Rebecca N. White
Google

Margaret Mitchell
Google

Timnit Gebru
Google

Ben Hutchinson
Google

Jamila Smith-Loud
Google

Daniel Theron
Google

Parker Barnes
Google

ABSTRACT

Rising concern for the societal implications of artificial intelligence systems has inspired a wave of academic and journalistic literature in which deployed systems are audited for harm by investigators from outside the organizations deploying the algorithms. However, it remains challenging for practitioners to identify the harmful repercussions of their own systems prior to deployment, and, once deployed, emergent issues can become difficult or impossible to trace back to their source.

In this paper, we introduce a framework for algorithmic auditing that supports artificial intelligence system development end-to-end, to be applied throughout the internal organization development lifecycle. Each stage of the audit yields a set of documents that together form an overall audit report, drawing on an organization's values or principles to assess the fit of decisions made throughout the process. The proposed auditing framework is intended to contribute to closing the *accountability gap* in the development and deployment of large-scale artificial intelligence systems by embedding a robust process to ensure audit integrity.

CCS CONCEPTS

• **Social and professional topics** → **System management**; **Technology audits**; • **Software and its engineering** → **Software development process management**.

KEYWORDS

Algorithmic audits, machine learning, accountability, responsible innovation

*Both authors contributed equally to this paper. This work was done by Inioluwa Deborah Raji as a fellow at Partnership on AI (PAI), of which Google, Inc. is a partner. This should not be interpreted as reflecting the official position of PAI as a whole, or any of its partner organizations.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02.
https://doi.org/10.1145/3351095.3372873

ACM Reference Format:
Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3351095.3372873

1 INTRODUCTION

With the increased access to artificial intelligence (AI) development tools and Internet-sourced datasets, corporations, nonprofits and governments are deploying AI systems at an unprecedented pace, often in massive-scale production systems impacting millions if not billions of users [1]. In the midst of this widespread deployment, however, come valid concerns about the effectiveness of these automated systems for the full scope of users, and especially a critique of systems that have the propensity to replicate, reinforce or amplify harmful existing social biases [8, 37, 62]. External audits are designed to identify these risks from outside the system and serve as accountability measures for these deployed models. However, such audits tend to be conducted after model deployment, when the system has already negatively impacted users [26, 51].

In this paper, we present internal algorithmic audits as a mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards, such as organizational AI principles. The audit process is necessarily boring, slow, meticulous and methodical—antithetical to the typical rapid development pace for AI technology. However, it is critical to slow down as algorithms continue to be deployed in increasingly high-stakes domains. By considering historical examples across industries, we make the case that such audits can be leveraged to anticipate potential negative consequences before they occur, in addition to providing decision support to design mitigations, more clearly defining and monitoring potentially adverse outcomes, and anticipating harmful feedback loops and system-level risks [20]. Executed by a dedicated team of organization employees, internal audits operate within the product development context and can inform the ultimate decision to abandon the development of AI technology when the risks outweigh the benefits (see Figure 1).

Inspired from the practices and artifacts of several disciplines, we go further to develop SMACTR, a defined internal audit framework meant to guide practical implementations. Our framework strives to establish interdisciplinarity as a default in audit and engineering processes while providing the much needed structure to support the conscious development of AI systems.

2 GOVERNANCE, ACCOUNTABILITY AND AUDITS

We use *accountability* to mean the state of being responsible or answerable for a system, its behavior and its potential impacts [38]. Although algorithms themselves cannot be held accountable as they are not moral or legal agents [7], the organizations designing and deploying algorithms can through *governance* structures. Proposed standard ISO 37000 defines this structure as "the system by which the whole organization is directed, controlled and held accountable to achieve its core purpose over the long term."[1] If the responsible development of artificial intelligence is a core purpose of organizations creating AI, then a governance system by which the whole organization is held accountable should be established.

[1]https://committee.iso.org/sites/tc309/home/projects/ongoing/ongoing-1.html

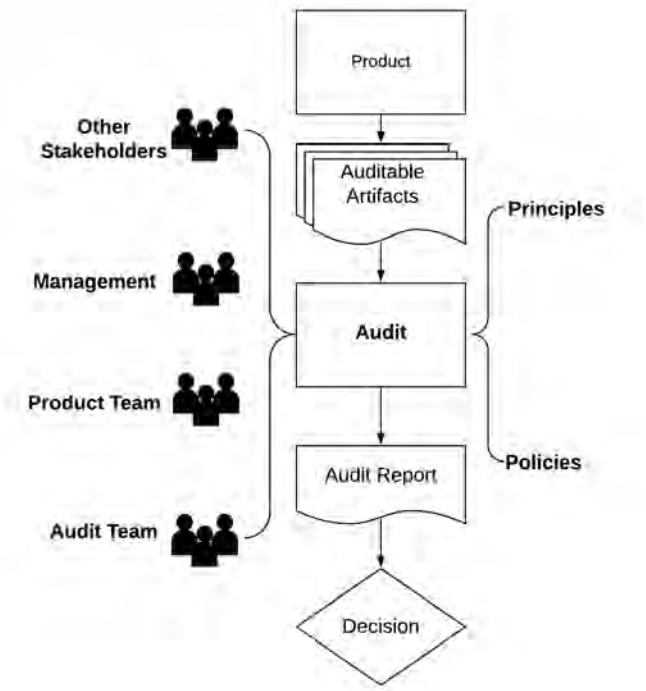


Figure 1: High-level overview of the context of an internal algorithmic audit. The audit is conducted during product development and prior to launch. The audit team leads the product team, management and other stakeholders in contributing to the audit. Policies and principles, including internal and external ethical expectations, also feed into the audit to set the standard for performance.

In environmental studies, Lynch and Veland [45] introduced the concept of *urgent governance*, distinguishing between *auditing* for system reliability vs societal harm. For example, a power plant can be consistently productive while causing harm to the environment through pollution [42]. Similarly, an AI system can be found technically reliable and functional through a traditional engineering quality assurance pipeline without meeting declared ethical expectations. A separate governance structure is necessary for the evaluation of these systems for ethical compliance. This evaluation can be embedded in the established quality assurance workflow but serves a different purpose, evaluating and optimizing for a different goal centered on social benefits and values rather than typical performance metrics such as accuracy or profit [39]. Although concerns about reliability are related, and although practices for testing production AI systems are established for industry practitioners [4], issues involving social impact, downstream effects in critical domains, and ethics and fairness concerns are not typically covered by concepts such as technical debt and reliability engineering.

2.1 What is an audit?

Audits are tools for interrogating complex processes, often to determine whether they comply with company policy, industry standards or regulations [43]. The IEEE standard for software development defines an audit as "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures" [32]. Building from methods of external auditing in investigative journalism and research [17, 62, 65], algorithmic auditing has started to become similar in spirit to the well-established practice of bug bounties, where external hackers are paid for finding vulnerabilities and bugs in released software [46]. These audits, modeled after intervention strategies in information security and finance [62], have significantly increased public awareness of algorithmic accountability.

An external audit of automated facial analysis systems exposed high disparities in error rates among darker-skinned women and lighter-skinned men [8], showing how structural racism and sexism can be encoded and reinforced through AI systems. [8] reveals *interaction failures*, in which the production and deployment of an AI system interacts with unjust social structures to contribute to biased predictions, as Safiya Noble has described [54]. Such findings demonstrate the need for companies to understand the social and power dynamics of their deployed systems' environments, and record such insights to manage their products' impact.

2.2 AI Principles as Customized Ethical Standards

According to Mittelstadt [49], at least 63 public-private initiatives have produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI. Important values such as ensuring AI technologies are subject to human direction and control, and avoiding the creation or reinforcement of unfair bias, have been included in many organizations' ethical charters. However, the AI industry lacks proven methods to translate principles into practice [49], and AI principles have been criticized for being vague and providing

little to no means of accountability [27, 82]. Nevertheless, such principles are becoming common methods to define the ethical priorities of an organization and thus the operational goals for which to aim [34, 83]. Thus, in the absence of more formalized and universal standards, they can be used as a North Star to guide the evaluation of the development lifecycle, and internal audits can investigate alignment with declared AI principles prior to model deployment. We propose a framing of risk analyses centered on the failure to achieve AI principle objectives, outlining an audit practice that can begin translating ethical principles into practice.

2.3 Audit Integrity and Procedural Justice

Audit results are at times approached with skepticism since they are reliant on and vulnerable to human judgment. To establish the integrity of the audit itself as an independently valid result, the audit must adhere to the proper execution of an established audit process. This is a repeatedly observed phenomenon in tax compliance auditing, where several international surveys of tax compliance demonstrate that a fixed and vetted tax audit methodology is one of the most effective strategies to convince companies to respect audit results and pay their full taxes [22, 53].

Procedural justice implies the legitimacy of an outcome due to the admission of a fair and thorough process. Establishing procedural justice to increase compliance is thus a motivating factor for establishing common and robust frameworks through which independent audits can demonstrate adherence to standards. In addition, audit integrity is best established when auditors themselves live up to an ethical standard, vetted by adherence to an expected code of conduct or norm in how the audit is to be conducted. In finance, for example, it became clear that any sense of dishonesty or non-transparency in audit methodology would lead audit targets to dismiss rather than act on results [66].

2.4 The Internal Audit

External auditing, in which companies are accountable to a third party [62], are fundamentally limited by lack of access to internal processes at the audited organizations. Although external audits conducted by credible experts are less affected by organization-internal considerations, external auditors can only access model outputs, for example by using an API [65]. Auditors do not have access to intermediate models or training data, which are often protected as trade secrets [9]. Internal auditors' direct access to systems can thus help extend traditional external auditing paradigms by incorporating additional information typically unavailable for external evaluations to reveal previously unidentifiable risks.

The goals of an internal audit are similar to quality assurance, with the objective to enrich, update or validate the risk analysis for product deployment. Internal audits aim to evaluate how well the product candidate, once in real-world operation, will fit the expected system behaviour encoded in standards.

A modification in objective from a post-deployment audit to pre-deployment audit applied throughout the development process enables proactive ethical intervention methods, rather than simply informing reactive measures only implementable after deployment, as is the case with a purely external approach. Because there is an increased level of system access in an internal audit, identified gaps in performance or processes can be mapped to sociotechnical considerations that should be addressed through joint efforts with product teams. As the audit results can lead to ambiguous conclusions, it is critical to identify key stakeholders and decision makers who can drive appropriate responses to audit outcomes.

Additionally, with an internal audit, because auditors are employees of the organization and communicate their findings primarily to an internal audience, there is opportunity to leverage these audit outcomes for recommendations of structural organizational changes needed to make the entire engineering development process auditable and aligned with ethical standards. Ultimately, internal audits complement external accountability, generating artifacts or transparent information [70] that third parties can use for external auditing, or even end-user communication. Internal audits can thus enable review and scrutiny from additional stakeholders, by enforcing transparency through stricter reporting requirements.

3 LESSONS FROM AUDITING PRACTICES IN OTHER INDUSTRIES

Improving the governance of artificial intelligence development is intended to reduce the risks posed by new technology. While not without faults, safety-critical and regulated industries such as aerospace and medicine have long traditions of auditable processes and design controls that have dramatically improved safety [77, 81].

3.1 Aerospace

Globally, there is one commercial airline accident per two million flights [63]. This remarkable safety record is the result of a joint and concerted effort over many years by aircraft and engine manufacturers, airlines, governments, regulatory bodies, and other industry stakeholders [63]. As modern avionic systems have increased in size and complexity (for example, the Boeing 787 software is estimated at 13 million lines of code [35]), the standard 1-in-1,000,000,000 per use hour maximum failure probability for critical aerospace systems remains an underappreciated engineering marvel [19].

However, as the recent Boeing 737 MAX accidents indicate, safety is never finished, and the qualitative impact of failures cannot be ignored—even one accident can impact the lives of many and is rightfully acknowledged as a catastrophic tragedy. Complex systems tend to drift toward unsafe conditions unless constant vigilance is maintained [42]. It is the sum of the tiny probabilities of individual events that matters in complex systems—if this grows without bound, the probability of catastrophe goes to one. The *Borel-Cantelli* Lemmas are formalizations of this statistical phenomenon [13], which means that we can never be satisfied with safety standards. Additionally, standards can be compromised if competing business interests take precedence. Because the non-zero risk of failure grows over time, without continuous active measures being developed to mitigate risk, disaster becomes inevitable [29].

3.1.1 Design checklists. Checklists are simple tools for assisting designers in having a more informed view of important questions, edge cases and failures [30]. Checklists are widely used in aerospace for their proven ability to improve safety and designs. There are several cautions about using checklists during the development of complex software, such as the risk of blind application, the broader

context and nuanced interrelated concerns are not considered. However, a checklist can be beneficial. It is good practice to avoid yes/no questions to reduce the risk that the checklist becomes a box-ticking activity, for example by asking designers and engineers to describe their processes for assessing ethical risk. Checklist use should also be related to real-world failures and higher-level system hazards.

3.1.2 Traceability. Another key concept from aerospace and safety-critical software engineering is *traceability*—which is concerned with the relationships between product requirements, their sources and system design. This practice is familiar to the software industry in requirements engineering [2]. However, in AI research, it can often be difficult to trace the provenance of large datasets or to interpret the meaning of model weights—to say nothing of the challenge of understanding how these might relate to system requirements. Additionally, as the complexity of sociotechnical systems is rapidly increasing, and as the speed and complexity of large-scale artificial intelligence systems increase, new approaches are necessary to understand risk [42].

3.1.3 Failure Modes and Effects Analysis. Finally, a standard tool in safety engineering is a *Failure Modes and Effects Analysis* (FMEA), methodical and systematic risk management approach that examines a proposed design or technology for foreseeable failures [72]. The main purpose of a FMEA is to define, identify and eliminate potential failures or problems in different products, designs, systems and services. Prior to conducting a FMEA, known issues with a proposed technology should be thoroughly mapped through a literature review and by collecting and documenting the experiences of the product designers, engineers and managers. Further, the risk exercise is based on known issues with relevant datasets and models, information that can be gathered from interviews and from extant technical documentation.

FMEAs can help designers improve or upgrade their products to reduce risk of failure. They can also help decision makers formulate corresponding preventive measures or improve reactive strategies in the event of post-launch failure. FMEAs are widely used in many fields including aerospace, chemical engineering, design, mechanical engineering and medical devices. To our knowledge, however, the FMEA method has not been applied to examine ethical risks in production-scale artificial intelligence models or products.

3.2 Medical devices

Internal and external quality assurance audits are a daily occurrence in the pharmaceutical and medical device industry. Audit document trails are as important as the drug products and devices themselves. The history of quality assurance audits in medical devices dates from several medical disasters in which devices, such as infusion pumps and autoinjectors, failed or were used improperly [80].

3.2.1 Design Controls. For medical devices, the stages of product development are strictly defined. In fact, federal law (Code of Federal Regulations Title 21) mandates that medical-device makers establish and maintain "design control" procedures to ensure that design requirements are met and designs and development processes are auditable. Practically speaking, design controls are a documented method of ensuring that the end product matches the intended use, and that potential risks from using the technology have been anticipated and mitigated [77]. The purpose is to ensure that anticipated risks related to the use of technology are driven down to the lowest degree that is reasonably practicable.

3.2.2 Intended Use. Medical-device makers must maintain procedures to ensure that design requirements meet the "intended use" of the device. The intended use of a "device" (or, increasingly in medicine, an algorithm—see [60] for more) determines the level of design control required: for example, a tongue depressor (a simple piece of wood) is the lowest class of risk (Class I), while a deep brain implant would be the highest (Class III). The intended use of a tongue depressor could be "to displace the tongue to facilitate examination of the surrounding organs and tissues", differentiating a tongue depressor from a Popsicle stick. This may be important when considering an algorithm that can be used to identify cats or to identify tumors; depending on its intended use, the same algorithm might have drastically different risk profiles, and additional risks arise from unintended uses of the technology.

3.2.3 Design History File. For products classified as medical devices, at every stage of the development process, device makers must document the design input, output, review, verification, validation, transfer and changes—the design control process (section 3.2.1). Evidence that medical device designers and manufacturers have followed design controls must be kept in a design history file (DHF), which must be an accurate representation and documentation of the product and its development process. Included in the DHF is an extensive risk assessment and hazard analysis, which must be continuously updated as new risks are discovered. Companies also proactively maintain "post-market surveillance" for any issues that may arise with safety of a medical device.

3.2.4 Structural Vulnerability. In medicine there is a deep acknowledgement of socially determinant factors in healthcare access and effectiveness, and an awareness of the social biases influencing the dynamic of prescriptions and treatments. This widespread acknowledgement led to the framework of operationalizing structural vulnerability in healthcare contexts, and effectively the design of an assessment tool to record the anticipated social conditions surrounding a particular remedy or medical recommendation [61]. Artificial intelligence models are equally subject to social influence and social impact, and undergoing such assessments on more holistic and population- or environment-based considerations is relevant to algorithmic auditing.

3.3 Finance

As automated accounting systems started to appear in the 1950s, corporate auditors continued to rely on manual procedures to audit "around the computer". In the 1970s, the Equity Funding Corporation scandal and the passage of the Foreign Corrupt Practices Act spurred companies to more thoroughly integrate internal controls throughout their accounting systems. This heightened the need to audit these systems directly. The 2002 Sarbanes-Oxley Act introduced sweeping changes to the profession in demanding greater focus on financial reporting and fraud detection [10].

Financial auditing had to play catch-up as the complexity and automation of financial business practices became too unwieldy to manage manually. Stakeholders in large companies and government

regulators desired a way to hold companies accountable. Concerns among regulators and shareholders that the managers in large financial firms would squander profits from newly created financial instruments prompted the development of financial audits [74].

Additionally, as financial transactions and markets became more automated, abstract and opaque, threats to social and economic values were answered increasingly with audits. But financial auditing lagged behind the process of technology-enabled financialization of markets and firms.

3.3.1 Audit Infrastructure. In general, internal financial audits seek assurance that the organization has a formal governance process that is operating as intended: values and goals are established and communicated, the accomplishment of goals is monitored, accountability is ensured and values are preserved. Further, internal audits seek to find out whether significant risks within the organization are being managed and controlled to an acceptable level [71].

Internal financial auditors typically have unfettered access to necessary information, people, records and outsourced operations across the organization. IIA Performance Standard 2300, Performing the Engagement [55], states that internal auditors should identify, analyze, evaluate and record sufficient information to achieve the audit objectives. The head of internal audit determines how internal auditors carry out their work and the level of evidence required to support their conclusions.

3.4 Discussion and Challenges

The lessons from other industries above are a useful guide toward building internal accountability to society as a stakeholder. Yet, there are many novel and unique aspects of artificial intelligence development that present urgent research challenges to overcome.

Current software development practice in general, and artificial intelligence development in particular, does not typically follow the *waterfall* or verification-and-validation approach [16]. These approaches are still used, in combination with agile methods, in the above-mentioned industries because they are much more documentation-oriented, auditable and requirements-driven. Agile artificial intelligence development is much faster and iterative, and thus presents a challenge to auditability. However, applying agile methodologies to internal audits themselves is a current topic of research in the internal audit profession.[2]

Most internal audit functions outside of heavily regulated industries tend to take a risk-based approach. They work with product teams to ask "what could go wrong" at each step of a process and use that to build a risk register [59]. This allows risks to rise to the surface in a way that is informed by the people who know these processes and systems the best. Internal audits can also leverage relevant experts from within the company to facilitate such discussions and provide additional insight on potential risks [3].

Large-scale production AI systems are extraordinarily complex, and a critical line of future research relates to addressing the interaction of highly complex coupled sociotechnical systems. Moreover, there is a dynamic complex interaction between users as sources of data, data collection, and model training and updating. Additionally, governance processes based solely on risk have been criticized for being unable to anticipate the most profound impacts from technological innovation, such as the financial crisis in 2008, in which big data and algorithms played a large role [52, 54, 57].

With artificial intelligence systems it can be difficult to trace model output back to requirements because these may not be explicitly documented, and issues may only become apparent once systems are released. However, from an ethical and moral perspective it is incumbent on producers of artificial intelligence systems to anticipate ethics-related failures before launch. However, as [58] and [31] point out, the design, prototyping and maintenance of AI systems raises many unique challenges not commonly faced with other kinds of intelligent systems or computing systems more broadly. For example, *data entanglement* results from the fact that artificial intelligence is a tool that mixes data sources together. As Scully et al. point out, artificial intelligence models create entanglement and make the isolation of improvements effectively impossible [67], which they call *Change Anything Change Everything*. We suggest that by having explicit documentation about the purpose, data, and model space, potential hazards could be identified earlier in the development process.

Selbst and Barocas argue that "one must seek explanations of the process behind a model's development, not just explanations of the model itself" [68]. As a relatively young community focused on fairness, accountability, and transparency in AI, we have some indication of the system culture requirements needed to normalize, for example, an adequately thorough documentation procedure and guidelines [24, 48]. Still, we lack the formalization of a standard model development template or practice, or process guidelines for when and in which contexts it is appropriate to implement certain recommendations. In these cases, internal auditors can work with engineering teams to construct the missing documentation to assess practices against the scope of the audit. Improving documentation can then be a remediation for future work.

Also, as AI is at times considered a "general purpose technology" with multiple and dual uses [78], the lack of reliable standardization poses significant challenges to governance efforts. This challenge is compounded by increasing customization and variability of what an AI product development lifecycle looks like depending on the anticipated context of deployment or industry.

We thus combine learnings from prior practice in adjacent industries while recognizing the uniqueness of the commercial AI industry to identify key opportunities for internal auditing in our specific context. We do so in a way that is appropriate to the requirements of an AI system.

4 SMACTR: AN INTERNAL AUDIT FRAMEWORK

We now outline the components of an initial internal audit framework, which can be framed as encompassing five distinct stages—Scoping, Mapping, Artifact Collection, Testing and Reflection (SMACTR)—all of which have their own set of documentation requirements and account for a different level of the analysis of a system. Figure 2 illustrates the full set of artifacts recommended for each stage.

To illustrate the utility of this framework, we contextualize our descriptions with the hypothetical example of Company X Inc.,

[2] https://deloitte.wsj.com/riskandcompliance/2018/08/06/mind-over-matter-implementing-agile-internal-audit/

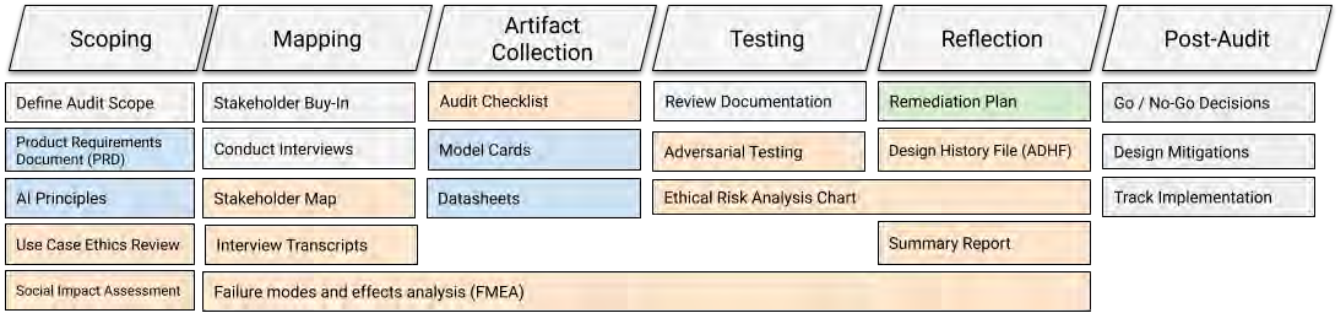


Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

a large multinational software engineering consulting firm, specializing in developing custom AI solutions for a diverse range of clients. We imagine this company has designated five AI principles, paraphrased from the most commonly identified AI principles in a current online English survey [34]–"Transparency", "Justice, Fariness & Non-Discrimination", "Safety & Non-Maleficence", "Responsibility & Accountability" and "Privacy". We also assume that the corporate structure of Company X is typical of any technical consultancy, and design our stakeholder map by this assumption.

Company X has decided to pilot the SMACTR internal audit framework to fulfill a corporate mandate towards responsible innovation practice, accommodate external accountability and operationalize internal consistency with respect to its identified AI principles. The fictional company thus pilots the audit framework on two hypothetical client projects.

The first (hypothetical) client wishes to develop a child abuse screening tool similar to that of the real cases extensively studied and reported on [11, 14, 15, 21, 25, 36]. This complex case intersects heavily with applications in high-risk scenarios with dire consequences. This scenario demonstrates how, for algorithms interfacing with high-risk contexts, a structured framework can allow for the careful consideration of all the possibilities and risks with taking on the project, and the extent of its understood social impact.

The second invented client is Happy-Go-Lucky, Inc., an imagined photo service company looking for a smile detection algorithm to automatically trigger the cameras in their installed physical photo booths. In this scenario, the worst case is a lack of customer satisfaction—the stakes are low and the situation seems relatively straightforward. This scenario demonstrates how in even seemingly simple and benign cases, ethical consideration of system deployment can reveal underlying issues to be addressed prior to deployment, especially when we contextualize the model within the setting of the product and deployment environment.

An end-to-end worked example of the audit framework is available as supplementary material to this paper for the Happy-Go-Lucky, Inc. client case. This includes demonstrative templates of all recommended documentation, with the exception of specific process files such as any experimental results, interview transcripts, a design history file and the summary report. Workable templates can also be accessed as an online resource here.

4.1 The Governance Process

To design our audit procedure, we suggest complementing formal risk assessment methodologies with ideas from responsible innovation, which stresses four key dimensions: *anticipation*, *reflexivity*, *inclusion* and *responsiveness* [73], as well as system-theoretic concepts that help grapple with increasing complexity and coupling of artificial intelligence systems with the external world [42]. Risk-based assessments can be limited in their ability to capture social and ethical stakes, and they should be complemented by anticipatory questions such as, "what if...?". The aim is to increase ethical foresight through systematic thinking about the larger sociotechnical system in which a product will be deployed [50]. There are also intersections between this framework and just effective product development theory [5], as many of the components of audit design refocus the product development process to prioritize the user and their ultimate well-being, resulting in a more effective product performance outcome.

At a minimum, the internal audit process should enable critical reflections on the potential impact of a system, serving as internal education and training on ethical awareness in addition to leaving what we refer to as a "transparency trail" of documentation at each step of the development cycle (see Figure 2). To shift the process into an actionable mechanism for accountability, we present a validated and transparently outlined procedure that auditors can commit to. The thoroughness of our described process will hopefully engage the trust of audit targets to act on and acknowledge post-audit recommendations for engineering practices in alignment with prescribed AI principles.

This process primarily addresses how to conduct internal audits, providing guidance for those that have already deemed an audit necessary but would like to further define the scope and execution details. Though not covered here, an equally important process is determining what systems to audit and why. Each industry has a way to judge what requires a full audit, but that process is discretionary and dependent on a range of contextual factors pertinent to the industry, the organization, audit team resourcing, and the case

at hand. Risk prioritization and the necessary variance in scrutiny is a separately interesting and rich research topic on its own. The process outlined below can be applied in full or in a lighter-weight formulation, depending on the level of assessment desired.

4.2 The Scoping Stage

For both clients, a product or request document is provided to specify the requirements and expectations of the product or feature. The goal of the scoping stage is to clarify the objective of the audit by reviewing the motivations and intended impact of the investigated system, and confirming the principles and values meant to guide product development. This is the stage in which the risk analysis begins by mapping out intended use cases and identifying analogous deployments either within the organization or from competitors or adjacent industries. The goal is to anticipate areas to investigate as potential sources of harm and social impact. At this stage, interaction with the system should be minimal.

In the case of the smile-triggered phone booth, a smile detection model is required, providing a simple product, with not a broad scope of considerations as the potential for harm does not go much beyond inconvenience or customer exclusion and dissatisfaction. For the child abuse detection product, there are many more approaches to solving the issue and many more options for how the model interacts with the broader system. The use case itself involves many ethical considerations, as an ineffective model may result in serious consequences like death or family separation.

The key artifacts developed by the auditors from this stage include an ethical review of the system use case and a social impact assessment. Pre-requisite documents from the product and engineering team should be a declaration or confirmation statement of ethical objectives, standards and AI principles. The product team should also provide a Product Requirements Document (PRD), or project proposal from the initial planning of the audited product.

4.2.1 Artifact: Ethical Review of System Use Case. When a potential AI system is in the development pipeline, it should be reviewed with a series of questions that first and foremost check to see, at a high level, whether the technology aligns with a set of ethical values or principles. This can take the form of an ethical review that considers the technology from a responsible innovation perspective by asking who is likely to be impacted and how.

Importantly, we stress standpoint diversity in this process. **Algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values.** Thus it is not always possible for individual technology workers to identify or assess their own biases or faulty assumptions [33]. For this reason, a critical range of viewpoints is included in the review process. The essential inclusion of independent domain experts and marginalized groups in the ethical review process "has the potential to lead to more rigorous critical reflection because their experiences will often be precisely those that are most needed in identifying problematic background assumptions and revealing limitations with research questions, models, or methodologies" [33]. Another method to elicit implicit biases or motivated cognition [40] is to ask people to reflect on their preliminary assessment and then ask whether they might have reason to regret the action later on. This can shed light on how our position in society biases our assumptions and ways of knowing [18].

An internal ethics review board that includes a diversity of voices should review proposed projects and document its views. Internal ethics review boards are common in biomedical research, and the purpose of these boards is to ensure that the rights, safety, and well-being of all human subjects involved in medical research are protected [56]. Similarly, the purpose of an ethics review board for AI systems includes safeguarding human rights, safety, and well-being of those potentially impacted.

4.2.2 Artifact: Social Impact Assessment. A social impact assessment should inform the ethical review. Social impact assessments are commonly defined as a method to analyze and mitigate the unintended social consequences, both positive and negative, that occur when a new development, program, or policy engages with human populations and communities [79]. In it, we describe how the use of an artificial intelligence system might change people's ways of life, their culture, their community, their political systems, their environment, their health and well-being, their personal and property rights, and their experiences (positive or negative) [79].

The social impact assessment includes two primary steps: an assessment of the severity of the risks, and an identification of the relevant social, economic, and cultural impacts and harms that an artificial intelligence system applied in context may create. The severity of risk is the degree to which the specific context of the use case is assessed to determine the degree in which potential harms may be amplified. The severity assessment proceeds from the analysis of impacts and harms to give a sense of the relative severity of the harms and impacts depending on the sensitivity, constraints, and context of the use case.

4.3 The Mapping Stage

The mapping stage is not a step in which testing is actively done, but rather a review of what is already in place and the perspectives involved in the audited system. This is also the time to map internal stakeholders, identify key collaborators for the execution of the audit, and orchestrate the appropriate stakeholder buy-in required for execution. At this stage, the FMEA (Section 3.1.3) should begin and risks should be prioritized for later testing.

As Company X is a consultancy, this stage mainly requires identifying the stakeholders across product and engineering teams anchored to this particular client project, and recording the nature of their involvement and contribution. This enables an internal record of individual accountability with respect to participation towards the final outcome, and enables the trace of relevant contacts for future inquiry.

For the child abuse detection algorithm, the initial identification of failure modes reveals the high stakes of the application, and immediate threats to the "Safety & Non-Maleficence" principle. False positives overwhelm staff and may lead to the separation of families that could have recovered. False negatives may result in a dead or injured child that could have been rescued. For the smile detector, failures disproportionately impact those with alternative emotional expressions—those with autism, different cultural norms on the formality of smiling, or different expectations for the photograph who are then excluded from the product by design.

The key artifacts from this stage include a stakeholder map and collaborator contact list, a system map of the product development lifecycle, and the engineering system overview, especially in cases where multiple models inform the end product. Additionally, this stage includes a design history file review of all existing documentation of the development process or historical artifacts on past versions of the product. Finally, it includes a report or interview transcripts on key findings from internal ethnographic fieldwork involving the stakeholders and engineers.

4.3.1 Artifact: Stakeholder Map. Who was involved in the system audit and collaborators in the execution of the audit should be outlined. Clarifying participant dynamics ensures a more transparent representation of the provided information, giving further context to the intended interpretation of the final audit report.

4.3.2 Artifact: Ethnographic Field Study. As Leveson points out, bottom-up decentralized decision making can lead to failures in complex sociotechnical systems [42]. Each local decision may be correct in the limited context in which it was made, but can lead to problems when these decisions and organizational behaviors interact. With modern large-scale artificial intelligence projects and API development, it can be difficult to gain a shared understanding at the right level of system description to understand how local decisions, such as the choice of dataset or model architecture, will impact final system behavior.

Therefore, ethnography-inspired fieldwork methodology based on how audits are conducted in other industries, such as finance [74] and healthcare [64] is useful to get a deeper and qualitative understanding of the engineering and product development process. As in internal financial auditing, access to key people in the organization is important. This access involves semi-structured interviews with a range of individuals close to the development process and documentation gathering to gain an understanding of possible gaps that need to be examined more closely.

Traditional metrics for artificial intelligence like loss may conceal fairness concerns, social impact risks or abstraction errors [69]. A key challenge is to assess how the numerical metrics specified in the design of an artificial intelligence system reflect or conform with these values. Metrics and measurement are important parts of the auditing process, but should not become aims and ends in themselves when weighing whether an algorithmic system under audit is ethically acceptable for release. Taking metrics measured in isolation risks recapitulating the abstraction error that [69] point out, "To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore to make a category error, or as we posit here, an abstraction error." What we consider data is already an interpretation, highly subjective and contested [23]. Metrics must be understood in relation to the engineering context in which they were developed and the social context into which they will be deployed. During the interviews, auditors should capture and pay attention to what falls outside the measurements and metrics, and to render explicit the assumptions and values the metrics apprehend [75]. For example, the decision about whether to prioritize the false positive rate over false negative rate (precision/recall) is a question about values and cannot be answered without stating the values of the organization, team or even engineer within the given development context.

4.4 The Artifact Collection Stage

Note that the collection of these artifacts advances adherence to the declared AI principles of the organization on "Responsibility & Accountability" and "Transparency".

In this stage, we identify and collect all the required documentation from the product development process, in order to prioritize opportunities for testing. Often this implies a record of data and model dynamics though application-based systems can include other product development artifacts such as design documents and reviews, in addition to systems architecture diagrams and other implementation planning documents and retrospectives.

At times documentation can be distributed across different teams and stakeholders, or is missing altogether. In certain cases, the auditor is in a position to enforce retroactive documentation requirements on the product team, or craft documents themselves.

The model card for the smile detection model is the template model card from the original paper [48]. A hypothetical datasheet for this system is filled out using studies on the CelebA dataset, with which the smile detector is built [44, 47]. In the model card, we identify potential for misuse if smiling is confused for positive affect. From the datasheet for the CelebA dataset, we see that although the provided binary gender labels seem balanced for this dataset (58.1% female, 42% male), other demographic details are quite skewed (77.8% aged 0-45, 22.1% aged over 46 and 14.2% lighter-skinned, 85.8% darker-skinned)[47].

The key artifact from auditors during this stage is the audit checklist, one method of verifying that all documentation pre-requisites are provided in order to commence the audit. Those pre-requisites can include model and data transparency documentation.

4.4.1 Artifact: Design Checklist. This checklist is a method of taking inventory of all the expected documentation to have been generated from the product development cycle. It ensures that the full scope of expected product processes and that the corresponding documentation required to be completed before the audit review can begin are finished. This is also a procedural evaluation of the development process for the system, to ensure that appropriate actions were pursued throughout system development ahead of the evaluation of the final system outcome.

4.4.2 Artifacts: Datasheets and Model Cards. Two recent standards can be leveraged to create auditable documentation, model cards and datasheets [24, 48]. Both model cards and datasheets are important tools toward making algorithmic development and the algorithms themselves more auditable, with the aim of anticipating risks and harms with using artificial intelligence systems. Ideally, these artifacts should be developed and/or collected by product stakeholders during the course of system development.

To clarify the intended use cases of artificial intelligence models and minimize their usage in contexts for which they are not well suited, Mitchell et al. recommend that released models be accompanied by documentation detailing their performance characteristics [48], called a *model card*. This should include information about how the model was built, what assumptions were made during development, and what type of model behavior might be experienced by different cultural, demographic or phenotypic groups. A

model card is also extremely useful for internal development purposes to make clear to stakeholders details about trained models that are included in larger software pipelines, which are parts of internal organizational dynamics, which are then parts of larger sociotechnical logics and processes. A robust model card is key to documenting the intended use of the model as well as information about the evaluation data, model scope and risks, and what might be affecting model performance.

Model cards are intended to complement "Datasheets for Datasets" [24]. Datasheets for machine learning datasets are derived by analogy from the electronics hardware industry, where a datasheet for an electronics component describes its operating characteristics, test results, and recommended uses. A critical part of the datasheet covers the data collection process. This set of questions are intended to provide consumers of the dataset with the information they need to make informed decisions about using the dataset: what mechanisms or procedures were used to collect the data? Was any ethical review process conducted? Does the dataset relate to people?

This documentation feeds into the auditors' assessment process.

4.5 The Testing Stage

This stage is where the majority of the auditing team's testing activity is done—when the auditors execute a series of tests to gauge the compliance of the system with the prioritized ethical values of the organization. Auditors engage with the system in various ways, and produce a series of artifacts to demonstrate the performance of the analyzed system at the time of the audit. Additionally, auditors review the documentation collected from the previous stage and begin to make assessments of the likelihood of system failures to comply with declared principles.

High variability in approach is likely during this stage, as the tests that need to be executed change dramatically depending on organizational and system context. Testing should be based on a risk prioritization from the FMEA.

For the smile detector, we might employ counterfactual adversarial examples designed to confuse the model and find problematic failure modes derived from the FMEA. For the child prediction model, we test performance on a selection of diverse user profiles. These profiles can also be treated for variables that correlate with vulnerable groups to test whether the model has learned biased associations with race or SES.

For the ethical risk analysis chart, we look at the principles and realize that there are immediate risks to the "Privacy" principle—with one case involving juvenile data, which is sensitive, and the other involving face data, a biometric. This is also when it becomes clear that in the smiling booth case, there is disproportionate performance for certain underrepresented user subgroups, thus jeopardizing the "Justice, Fariness & Non-Discrimination" principle.

The main artifacts from this stage of the auditing process are the results of tests such as adversarial probing of the system and an ethical risk analysis chart.

4.5.1 Artifact: Adversarial Testing . Adversarial testing is a common approach to finding vulnerabilities in both pre-release and post-launch technology, for example in privacy and security testing [6]. In general, adversarial testing attempts to simulate what a hostile actor might do to gain access to a system, or to push the limits of the system into edge case or unstable behavior to elicit very-low probability but high-severity failures.

In this process, direct non-statistical testing uses tailored inputs to the model to see if they result in undesirable outputs. These inputs can be motivated by an intersectional analysis, for example where an ML system might produce unfair outputs based on demographic and phenotypic groups that might combine in non-additive ways to produce harm, or over time recapitulate harmful stereotypes or reinforce unjust social dynamics (for example, in the form of opportunity denial). This is distinct from adversarially attacking a model with human-imperceptible pixel manipulations to trick the model into misidentifying previously learned outputs [28], but these approaches can be complementary. This approach is more generally defined—encompassing a range of input options to try in an active attempt to fool the system and incite identified failure modes from the FMEA.

Internal adversarial testing prior to launch can reveal unexpected product failures before they can impact the real world. Additionally, proactive adversarial testing of already-launched products can be a best practice for lifecycle management of released systems. The FMEA should be updated with these results, and the relative changes to risks assessed.

4.5.2 Artifact: Ethical Risk Analysis Chart. The ethical risk analysis chart considers the combination of the likelihood of a failure and the severity of a failure to define the importance of the risk. Highly likely and dangerous risks are considered the most high-priority threats. Each risk is assigned a severity indication of "high", "mid" and "low" depending on their combination of these features.

Failure likelihood is estimated by considering the occurrence of certain failures during the adversarial testing of the system and the severity of the risk is identified in earlier stages, from informative processes such as the social impact assessment and ethnographic interviews.

4.6 The Reflection Stage

This phase of the audit is the more reflective stage, when the results of the tests at the execution stage are analyzed in juxtaposition with the ethical expectations clarified in the audit scoping. Auditors update and formalize the final risk analysis in the context of test results, outlining specific principles that may be jeopardized by the AI system upon deployment. This phase will reflect on product decisions and design recommendations that could be made following the audit results.

Additionally, key artifacts at this stage may include a mitigation plan or action plan, jointly developed by the audit and engineering teams, that outlines prioritized risks and test failures that the engineering team is in a position to mitigate for future deployments or for a future version of the audited system.

For the smile detection algorithm, the decision could be to train a new version of the model on more diverse data before considering deployment, and add more samples of underrepresented populations in CelebA to the training data. It could be decided that the use case does not necessarily define affect, but treats smiling as a favourable photo pose. Design choices for other parts of the product outside the model should be considered—for instance, an opt-in functionality with user permissions required on the screen before

applying the model-controlled function, and the default being that the model-controlled trigger is disabled. There could also be an included disclaimer on privacy, assuring users of safe practices for face data storage and consent. Once these conditions are met, Company X could be confident to greenlight developing this product for the client.

For the child abuse detection model—this is a more complex decision. Given the ethical considerations involved, the project may be stalled or even cancelled, requiring further inquiry into the ethics of the use case, and the capability of the team to complete the mitigation plan required to deploy an algorithm in such a high risk scenario.

4.6.1 Artifact: Algorithmic Use-related Risk Analysis and FMEA. The risk analysis should be informed by the social impact assessment and known issues with similar models. Following Leveson's work on safety engineering [42], we stress that careful attention must be paid to the distinction between the *designers' mental models* of the artificial intelligence system and the *user's mental model*. The designers' mental models are an idealization of the artificial intelligence system before the model is released. Significant differences exist between this ideal model and how the actual system will behave or be used once deployed. This may be due to many factors, such as distributional drift [41] where the training and test set distributions differ from the real-world distribution, or intentional or unintentional misuse of the model for purposes other than those for which it was designed. Reasonable and foreseeable misuse of the model should be anticipated by the designer. Therefore, the *user's mental model* of the system should be anticipated and taken into consideration. Large gaps between the *intended* and *actual* uses of algorithms have been found in contexts such as criminal justice and web journalism [12].

This adds complexity to anticipated hazards and risks, nevertheless these should be documented where possible. Christin points out "the importance of studying the practices, uses, and implementations surrounding algorithmic technologies. Intellectually, this involves establishing new exchanges between literatures that may not usually interact, such as critical data studies, the sociology of work, and organizational analysis". We propose that known use-related issues with deployed systems be taken into account during the design stage. The format of the risk analysis can be variable depending on context, and there are many valuable templates to be found in *Failure Modes and Effects Analysis* (Section 3.1.3) framing and other risk analysis tools in finance and medical deployments.

4.6.2 Artifact: Remediation and Risk Mitigation Plan. After the audit is completed and findings are presented to the leadership and product teams, it is important to develop a plan for remediating these problems. The goal is to drive down the risk of ethical concerns or potential negative social impacts to the extent reasonably practicable. This plan can be reviewed by the audit team and leadership to better inform deployment decisions.

For the concerns raised in any audit against ethical values, a technical team will want to know: what is the threshold for acceptable performance? If auditors discover, for example, unequal classifier performance across subgroups, how close to parity is necessary to say the classifier is acceptable? In safety engineering, a risk threshold is usually defined under which the risk is considered tolerable. Though a challenging problem, similar standards could be established and developed in the ethics space as well.

4.6.3 Artifact: Algorithmic Design History File. Inspired by the concept of the design history file from the medical device industry [77], we propose an algorithmic design history file (ADHF) which would collect all the documentation from the activities outlined above related to the development of the algorithm. It should point to the documents necessary to demonstrate that the product or model was developed in accordance with an organization's ethical values, and that the benefits of the product outweigh any risks identified in the risk analysis process.

This design history file would form the basis of the final audit report, which is a written evaluation by the organization's audit team. The ADHF should assist with an audit trail, enabling the reconstruction of key decisions and events during the development of the product. The algorithmic report would then be a distillation and summary of the ADHF.

4.6.4 Artifact: Algorithmic Audit Summary Report. The report aggregates all key audit artifacts, technical analyses and documentation, putting this in one accessible location for review. This audit report should be compared qualitatively and quantitatively to the expectations outlined in the given ethical objectives and any corresponding engineering requirements.

5 LIMITATIONS OF INTERNAL AUDITS

Internal auditors necessarily share an organizational interest with the target of the audit. While it is important to maintain an independent and objective viewpoint during the execution of an audit, we awknowledge that this is challenging. The audit is never isolated from the practices and people conducting the audit, just as artificial intelligence systems are not independent of their developers or of the larger sociotechnical system. Audits are not unified or monolithic processes with an objective "view from nowhere", but must be understood as a "patchwork of coupled procedures, tools and calculative processes" [74]. To avoid audits becoming simply acts of reputation management for an organization, the auditors should be mindful of their own and the organizations' biases and viewpoints. Although long-standing internal auditing practices for quality assurance in the financial, aviation, chemical, food, and pharmaceutical industries have been shown to be an effective means of controlling risk in these industries [76], the regulatory dynamics in these industries suggest that internal audits are only one important aspect of a broader system of required quality checks and balances.

6 CONCLUSION

AI has the potential to benefit the whole of society, however there is currently an inequitable risk distribution such that those who already face patterns of structural vulnerability or bias disproportionately bear the costs and harms of many of these systems. Fairness, justice and ethics require that those bearing these risks are given due attention and that organizations that build and deploy artificial intelligence systems internalize and proactively address these social risks as well, being seriously held to account for system compliance to declared ethical principles.

REFERENCES

[1] Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha. 2015. Efficient machine learning for big data: A review. *Big Data Research* 2, 3 (2015), 87–93.

[2] Amel Bennaceur, Thein Than Tun, Yijun Yu, and Bashar Nuseibeh. 2019. Requirements Engineering. In *Handbook of Software Engineering*. Springer, 51–92.

[3] Li Bing, Akintola Akintoye, Peter J Edwards, and Cliff Hardcastle. 2005. The allocation of risk in PPP/PFI construction projects in the UK. *International Journal of project management* 23, 1 (2005), 25–35.

[4] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D Sculley. 2017. The ml test score: A rubric for ml production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1123–1132.

[5] Shona L Brown and Kathleen M Eisenhardt. 1995. Product development: Past research, present findings, and future directions. *Academy of management review* 20, 2 (1995), 343–378.

[6] Chad Brubaker, Suman Jana, Baishakhi Ray, Sarfraz Khurshid, and Vitaly Shmatikov. 2014. Using Frankencerts for Automated Adversarial Testing of Certificate Validation. In *in SSL/TLS Implementations,âĂİ IEEE Symposium on Security and Privacy*. Citeseer.

[7] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25, 3 (2017), 273–291.

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[9] Jenna Burrell. 2016. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[10] Paul Eric Byrnes, Abdullah Al-Awadhi, Benita Gullvist, Helen Brown-Liburd, Ryan Teeter, J Donald Warren Jr, and Miklos Vasarhelyi. 2018. Evolution of Auditing: From the Traditional Approach to the Future Audit 1. In *Continuous Auditing: Theory and Application*. Emerald Publishing Limited, 285–297.

[11] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.

[12] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.

[13] Kai Lai Chung and Paul Erdös. 1952. On the application of the Borel-Cantelli lemma. *Trans. Amer. Math. Soc.* 72, 1 (1952), 179–186.

[14] Rachel Courtland. 2018. Bias detectives: the researchers striving to make algorithms fair. *Nature* 558, 7710 (2018), 357–357.

[15] Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review* 79 (2017), 291–298.

[16] Michael A Cusumano and Stanley A Smith. 1995. Beyond the waterfall: Software development at Microsoft. (1995).

[17] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. (2014).

[18] Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. *arXiv preprint arXiv:1807.00553* (2018).

[19] Kevin Driscoll, Brendan Hall, Håkan Sivencrona, and Phil Zumsteg. 2003. Byzantine fault tolerance, from theory to reality. In *International Conference on Computer Safety, Reliability, and Security*. Springer, 235–248.

[20] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847* (2017).

[21] Virginia Eubanks. 2018. A child abuse prediction model fails poor families. *Wired Magazine* (2018).

[22] Sellywati Mohd Faizal, Mohd Rizal Palil, Ruhanita Maelah, and Rosiati Ramli. 2017. Perception on justice, trust and tax compliance behavior in Malaysia. *Kasetsart Journal of Social Sciences* 38, 3 (2017), 226–232.

[23] Jonathan Furner. 2016. "Data": The data. In *Information Cultures in the Digital Age*. Springer, 287–306.

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[25] Jeremy Goldhaber-Fiebert and Lea Prince. 2019. Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny CountyâĂŹs Child Welfare Office. *Pittsburgh: Allegheny County.[Google Scholar]* (2019).

[26] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 90–99.

[27] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[28] Shixiang Gu and Luca Rigazio. 2014. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068* (2014).

[29] John Haigh. 2012. *Probability: A very short introduction*. Vol. 310. Oxford University Press.

[30] Brendan Hall and Kevin Driscoll. 2014. Distributed System Design Checklist. (2014).

[31] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need? *arXiv preprint arXiv:1812.05239* (2018).

[32] IEEE. 2008. IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008* (Aug 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

[33] Kristen Intemann. 2010. 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia* 25, 4 (2010), 778–796.

[34] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. Artificial Intelligence: the global landscape of ethics guidelines. *arXiv preprint arXiv:1906.11668* (2019).

[35] Paul A Judas and Lorraine E Prokop. 2011. A historical compilation of software metrics with applicability to NASA's Orion spacecraft flight software sizing. *Innovations in Systems and Software Engineering* 7, 3 (2011), 161–170.

[36] Emily Keddell. 2019. Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice. *Social Sciences* 8, 10 (2019), 281.

[37] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).

[38] Nitin Kohli, Renata Barreto, and Joshua A Kroll. 2018. Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems. In *1st Conference on Fairness, Accountability, and Transparancy. New York, NY, USA*. 7.

[39] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.

[40] Arie W Kruglanski. 1996. Motivated social cognition: Principles of the interface. (1996).

[41] Joel Lehman. 2019. Evolutionary Computation and AI Safety: Research Problems Impeding Routine and Safe Real-world Application of Evolution. *arXiv preprint arXiv:1906.10189* (2019).

[42] Nancy Leveson. 2011. *Engineering a safer world: Systems thinking applied to safety*. MIT press.

[43] Jie Liu. 2012. The enterprise risk management and the risk oriented internal audit. *Ibusiness* 4, 03 (2012), 287.

[44] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[45] Amanda H Lynch and Siri Veland. 2018. *Urgency in the Anthropocene*. MIT Press.

[46] Thomas Maillart, Mingyi Zhao, Jens Grossklags, and John Chuang. 2017. Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *Journal of Cybersecurity* 3, 2 (2017), 81–90.

[47] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *arXiv preprint arXiv:1901.10436* (2019).

[48] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

[49] Brent Mittelstadt. 2019. AI Ethics: Too Principled to Fail? *SSRN* (2019).

[50] Brent Daniel Mittelstadt and Luciano Floridi. 2016. The ethics of big data: current and foreseeable issues in biomedical contexts. *Science and engineering ethics* 22, 2 (2016), 303–341.

[51] Laura Moy. 2019. How Police Technology Aggravates Racial Inequity: A Taxonomy of Problems and a Path Forward. *Available at SSRN 3340898* (2019).

[52] Fabian Muniesa, Marc Lenglet, et al. 2013. Responsible innovation in finance: directions and implications. *Responsible Innova-tion: Managing the Responsible Emergence of Science and Innovation in Society. Wiley, London* (2013), 185–198.

[53] Kristina Murphy. 2003. Procedural justice and tax compliance. *Australian Journal of Social Issues (Australian Council of Social Service)* 38, 3 (2003).

[54] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

[55] Institute of Internal Auditors. Research Foundation and Institute of Internal Auditors. 2007. *The Professional Practices Framework*. Inst of Internal Auditors.

[56] General Assembly of the World Medical Association et al. 2014. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *The Journal of the American College of Dentists* 81, 3 (2014), 14.

[57] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[58] Charles Parker. 2012. Unexpected challenges in large scale machine learning. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. ACM, 1–6.

[59] Fiona D Patterson and Kevin Neailey. 2002. A risk register database system to aid the management of project risk. *International Journal of Project Management* 20, 5 (2002), 365–374.

[60] W Price and II Nicholson. 2017. Regulating black-box medicine. *Mich. L. Rev.* 116 (2017), 421.

[61] James Quesada, Laurie Kain Hart, and Philippe Bourgois. 2011. Structural vulnerability and health: Latino migrant laborers in the United States. *Medical anthropology* 30, 4 (2011), 339–362.

[62] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*.

[63] Clarence Rodrigues and Stephen Cusick. 2011. *Commercial aviation safety 5/e*. McGraw Hill Professional.

[64] G Sirgo Rodríguez, M Olona Cabases, MC Martin Delgado, F Esteban Reboll, A Pobo Peris, M Bodí Saera, et al. 2014. Audits in real time for safety in critical care: definition and pilot study. *Medicina intensiva* 38, 8 (2014), 473–482.

[65] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014).

[66] David Satava, Cam Caldwell, and Linda Richards. 2006. Ethics and the auditing culture: Rethinking the foundation of accounting and auditing. *Journal of Business Ethics* 64, 3 (2006), 271–284.

[67] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. 2014. Machine learning: The high interest credit card of technical debt. (2014).

[68] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.

[69] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 59–68.

[70] Hetan Shah. 2018. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170362.

[71] Dominic SB Soh and Nonna Martinov-Bennie. 2011. The internal audit function: Perceptions of internal audit roles, effectiveness and evaluation. *Managerial Auditing Journal* 26, 7 (2011), 605–622.

[72] Diomidis H Stamatis. 2003. *Failure mode and effect analysis: FMEA from theory to execution*. ASQ Quality press.

[73] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (2013), 1568–1580.

[74] Alexander Styhre. 2015. *The financialization of the firm: Managerial and social implications*. Edward Elgar Publishing.

[75] Alexander Styhre. 2018. The unfinished business of governance: towards new governance regimes. In *The Unfinished Business of Governance*. Edward Elgar Publishing.

[76] JohnK Taylor. 2018. *Quality assurance of chemical measurements*. Routledge.

[77] Marie B Teixeira, Marie Teixeira, and Richard Bradley. 2013. *Design controls for the medical device industry*. CRC press.

[78] Manuel Trajtenberg. 2018. *AI as the next GPT: a Political-Economy Perspective*. Technical Report. National Bureau of Economic Research.

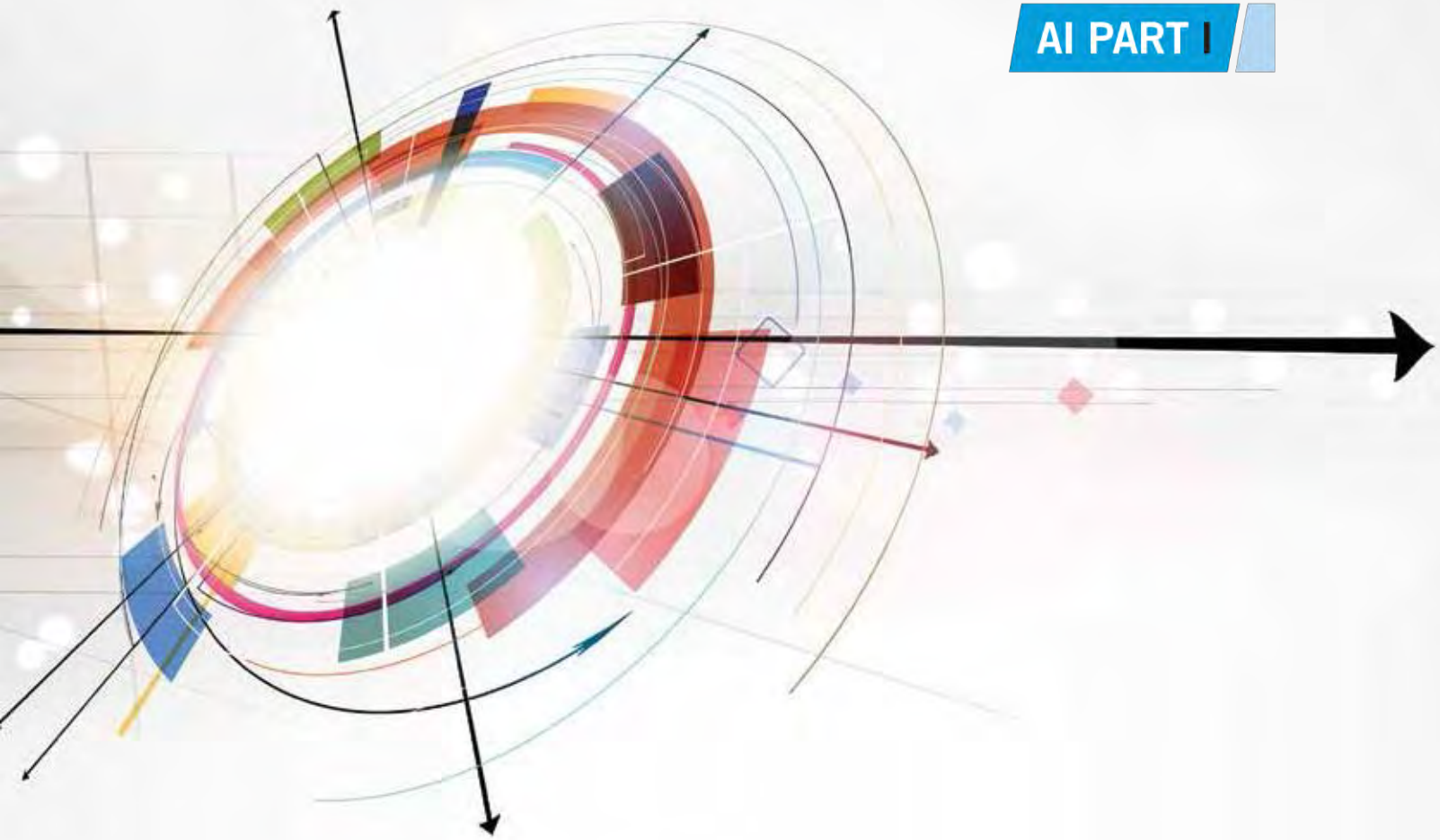
[79] Frank Vanclay. 2003. International principles for social impact assessment. *Impact assessment and project appraisal* 21, 1 (2003), 5–12.

[80] Tim Vanderveen. 2005. Averting highest-risk errors is first priority. *Patient Safety and Quality Healthcare* 2 (2005), 16–21.

[81] Ajit Kumar Verma, Srividya Ajit, Durga Rao Karanki, et al. 2010. *Reliability and safety engineering*. Vol. 43. Springer.

[82] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA*. 27–28.

[83] Yi Zeng, Enmeng Lu, and Cunqing Huangfu. 2018. Linking Artificial Intelligence Principles. *arXiv preprint arXiv:1812.04814* (2018).



AI PART I

GLOBAL PERSPECTIVES AND INSIGHTS

Artificial Intelligence – Considerations for the Profession of Internal Auditing

Special Edition

The Institute of Internal Auditors | Global

Table of Contents

Previous Issues

To access previous issues of Global Perspectives and Insights, visit www.theiia.org/gpi.

Reader Feedback

Send questions or comments to globalperspectives@theiia.org.

About The IIA

The Institute of Internal Auditors (IIA) is the internal audit profession's most widely recognized advocate, educator, and provider of standards, guidance, and certifications. Established in 1941, The IIA today serves more than 190,000 members from more than 170 countries and territories. The association's global headquarters are in Lake Mary, Fla., USA. For more information, visit www.globaliia.org.

Disclaimer

The opinions expressed in Global Perspectives and Insights are not necessarily those of the individual contributors or of the contributors' employers.

Copyright

Note

This paper, Part I of a three-part series:

- Presents an overview of AI basics.
- Explores internal audit's roles in AI.
- Discusses AI risks and opportunities.
- Introduces a framework for internal auditors (the Framework).

Parts II and III will provide information on practical applications of the AI auditing framework, including relevant engagement objectives and procedures which internal audit activities can use to customize an AI audit program to fit their organizations' risk profile and strategic objectives.

Introduction

Artificial intelligence (AI) is a broad term that refers to technologies that make machines "smart." Organizations are investing in AI research and applications to automate, augment, or replicate human intelligence — human analytical and/or decision-making — and the internal auditing profession must be prepared to fully participate in organizational AI initiatives.

There are many other terms related to AI, such as deep learning, machine learning, image recognition, natural-language processing, cognitive computing, intelligence amplification, cognitive augmentation, machine augmented intelligence, and augmented intelligence. AI, as used here, encompasses all of these concepts.

Putting AI Into Context

AI is not new. According to the McKinsey Global Institute's (MGI) discussion paper Artificial Intelligence: The Next Digital Frontier, the idea of AI dates back to 1950 when Alan Turing first proposed that a machine could communicate well enough to convince a human evaluator that it, too, was human.

While AI represents a series of significant advancements in technology, it was not the first, and likely will not be the last. Looking back over the last few decades, the advent of computers, PCs, spreadsheets, relational databases, sophisticated connectivity, and similar technological advancements have all impacted how organizations operate and accomplish their objectives. AI is poised to do the same with the potential to be as or more disruptive than many previous technological advances.

AI can be viewed as the latest significant advancement on a continuum of advancements that have occurred due to technology improvements. What *is* new is the advancement and scalability of technologies that have unleashed the practical application of AI.

This application was demonstrated publicly to a wide audience in 2011 when IBM's AI platform Watson won a "Jeopardy!" exhibition on prime time TV. According to IBM Research, IBM is "guided by the term 'augmented intelligence' rather than 'artificial intelligence,'" and focuses "on building practical AI applications that assist people with well-defined tasks." Human expertise develops technologies to make machines smart, and smart machines, in turn, augment human capabilities.

There is already widespread application of AI across diverse sectors (publicly held, privately held, government, and nonprofit) and industries. Consider, for example, that AI enables a number of new and novel capabilities that were impossible just a few years ago. But it is not only new and novel activities affected by AI. More mundane tasks that have been occurring for decades are being improved by AI such as loss modeling, credit analysis, valuations, transaction processing, and a host of others.

It is critical that internal auditors pay attention to the practical application of AI in business, and develop competencies that will enable the internal auditing profession to provide AI-related advisory and assurance services to organizations in all sectors and across all industries.

AI is dependent on big data and algorithms, and it can be intimidating, especially for internal audit activities and organizations that have yet to master big data. But internal auditors do not have to be data scientists or quantitative analysts to understand what AI can do for organizations, governments, and societies at large.

AI – The Basics

Big Data and Algorithms

AI is powered by algorithms, and algorithms are fueled by big data, so before an organization embarks on AI, it should have a strong foundation in big data. And before internal audit can think about addressing AI, it should already have a strong foundation in big data. For comprehensive guidance on understanding and auditing big data, including a discussion of opportunities and risks, and a sample work program, see The IIA's "GTAG: Understanding and Auditing Big Data," available free to IIA members and available to non-members through The IIA Bookstore (www.theiia.org).

Big data means more than just large amounts of data — big data refers to data (information) that reaches such high volume, variety, velocity, and variability that organizations invest in system architectures, tools, and practices specifically designed to handle the data. Much of this data may be generated by the organization itself, while other data may be publicly available or purchased from external sources.

To put big data to good use, organizations develop algorithms. An algorithm is a set of rules for the machine to follow. An algorithm is what enables a machine to quickly process vast amounts of data that a human cannot reasonably process, or even comprehend. The performance and accuracy of algorithms is very important. Algorithms are initially developed by humans, so human error and biases (both intentional and unintentional) will impact the performance of the algorithm. Faulty algorithms can produce minor undesirable glitches in an organization's operations, or major catastrophic outcomes. It is generally recognized that flawed algorithms, at least in part, fueled the 2008 global financial crisis.

Types of AI

In *The Conversation's* "Understanding the four types of AI, from reactive robots to self-aware beings," Arend Hintze, assistant professor of Integrative Biology & Computer Science and Engineering at Michigan State University, outlines four types of AI:

Uses of AI Technology

- Automobile manufacturers to develop self-driving vehicles.
- Online search engines to deliver targeted search results.
- Social media organizations to recognize faces in photographs and filter newsfeeds.
- Media companies to recommend books or shows to subscribers.
- Retailers to create customized online experiences for shoppers.
- Logistics companies to route optimal paths for deliveries.
- Governments to predict epidemics.
- Marketing professionals to deliver hyper-personalized content to customers in real time.
- Virtual assistants to use voice-controlled natural language to interface with consumers.

Type I. Reactive machines: This is AI at its simplest. Reactive machines respond to the same situation in exactly the same way, every time. An example of this is a machine that can beat world-class chess players because it has been programmed to recognize the chess pieces, know how each moves, and can predict the next move of both players.

Type II. Limited memory: Limited memory AI machines can look to the past, but the memories are not saved. Limited memory machines cannot build memories or "learn" from past experiences. An example is a self-driving vehicle that can decide to change lanes because a moment ago it noted an obstacle in its path.

Type III. Theory of mind: Theory of mind refers to the idea that a machine could recognize that others it interacts with have thoughts, feelings, and expectations. A machine embedded with Type III AI would be able to understand others' thoughts, feelings, and expectations, and be able to adjust its own behavior accordingly.

Type IV. Self-awareness: A machine embedded with Type IV AI would be self-aware. An extension of "theory of mind," a conscious or self-aware machine would be aware of itself, know about its internal states, and be able to predict the feelings of others.

In other words, a Type II self-driving vehicle would decide to change lanes when a pedestrian is in its path, simply because it recognizes the pedestrian as an obstacle. A Type III self-driving vehicle would understand that the pedestrian would expect the vehicle to stop, and a Type IV self-driving vehicle would *know* that it should stop because that is what the self-driving vehicle would want if it (the self-driving vehicle) were in the path of another oncoming vehicle. Wow.

Most "smart machines" today are manifestations of Type I or Type II AI. Ongoing research and development initiatives will enable organizations to advance toward practical applications of Type III and Type IV AI.

AI Opportunities and Risks

The first step toward understanding the organization's AI opportunities and risks is to thoroughly understand the organization's big data opportunities and risks. Again, for comprehensive guidance on understanding and auditing big data, including a discussion of opportunities and risks, and a sample work program, see The IIA's "GTAG: Understanding and Auditing Big Data," available free to IIA members and available to non-members through The IIA Bookstore (www.theiia.org). Examples of AI opportunities and risks include:

Opportunities

- The ability to compress the data processing cycle.
- The ability to reduce errors by replacing human actions with perfectly repeatable machine actions.
- The ability to replace time-intensive activities with time-efficient activities (process automation), reducing labor time and costs.
- The ability to have robots or drones replace humans in potentially dangerous situations.
- The ability to make better predictions, for everything from predicting sales of certain goods in particular markets to predicting epidemics and natural catastrophes.
- The ability to drive revenue and grow market share through AI initiatives.

Risks

- The risk that unidentified human biases will be imbedded in the AI technology.
- The risk that human logic errors will be imbedded in the AI technology.
- The risk that inadequate testing and oversight of AI results in ethically questionable results.
- The risk that AI products and services will cause harm, resulting in financial and/or reputational damage.
- The risks that customers or other stakeholders will not accept or adopt the organization's AI initiatives.
- The risk that the organization will be left behind by competitors if it does *not* invest in AI.
- The risk that investment in AI (infrastructure, research and development, and talent acquisition) will not yield an acceptable ROI.

More in-depth information on AI risks will be presented in Parts II and III of this three-part Global Perspectives and Insights series.

Internal Audit's Role

Internal audit is adept at evaluating and understanding the risks and opportunities related to the ability of an organization to meet its objectives. Leveraging this experience, internal audit can help an organization evaluate, understand, and communicate the degree to which artificial intelligence will have an effect (negative or positive) on the organization's ability to create value in the short, medium, or long term. Internal audit can engage through at least five critical and distinct activities related to artificial intelligence:

- For all organizations, internal audit should include AI in its risk assessment and consider whether to include AI in its risk-based audit plan.

Audit Focus

IIA Standard 2120: Risk Management (Excerpt)

The internal audit activity must evaluate the effectiveness and contribute to the improvement of risk management processes.

2120.A1 – The internal audit activity must evaluate risk exposures relating to the organization's governance, operations, and information systems regarding the:

- Achievement of the organization's objectives.
- Reliability and integrity of financial and operational information.
- Effectiveness and efficiency of operations and programs.
- Safeguarding of assets.
- Compliance with laws, regulations, policies, procedures, and contracts.

AI Competencies

- Natural language processing.
- Application program interfaces (APIs) such as facial recognition, image analytics, and text analytics.
- Algorithms and advanced modeling.
- Probabilities and applied statistics.
- Data analytics.
- Software engineering.
- Programming language.
- Machine learning.
- Computer vision.
- Robotics.

- For organizations exploring AI, internal audit should be actively involved in AI projects from their beginnings, providing advice and insight contributing to successful implementation. However, to avoid the perception of or actual impairments to both independence and objectivity, internal audit should not own, nor be responsible for, the implementation of AI processes, policies, or procedures.
- For organizations that have implemented some aspect of AI, either within its operations (such as a manufacturer using robotics on a production line) or incorporated into a product or service (such as a retailer customizing product offerings based on purchase history), internal audit should provide assurance over the management of risks related to the reliability of underlying algorithms and data on which the algorithms are based.
- Internal audit should ensure the moral and ethical issues that may surround the organization's use of AI are being addressed.
- Like the use of any other major system, proper governance structures need to be established and internal audit can provide assurance in this space.

Regardless of the specific activities performed, internal audit is well-suited to be a key contributor to an organization's AI-related activities. Internal audit:

- Understands the strategic objectives of the organization, and the processes implemented to achieve those objectives.
- Is able to evaluate whether AI activities are accomplishing their objectives.
- Can provide internal assurance over management's risk management activities relevant to AI risks.
- Is perceived as a trusted advisor that can positively support the adoption of AI to improve business processes or enhance product and service offerings.

Internal auditing should approach AI as it approaches everything — with systematic, disciplined methods to evaluate and improve the effectiveness of risk management, control, and governance processes related to AI.

AI Competencies: Filling the Understanding Gap

The pool of talent for technology professionals with AI expertise is reportedly small. Organizations who want to participate in the AI revolution need to grow or acquire talent with competencies in a multitude of areas.

While a handful of organizations in the technology, automotive, manufacturing, financial services, and utilities industries seem to be leading the AI revolution, it is hard to imagine an organization that will not be impacted by AI. Just as computers, spreadsheets, and distributed processing were a focus of select industries in their early stages, ultimately all organizations adopted aspects of these technologies. As AI becomes more mainstream, it is hard to imagine any internal audit activity that will not need to be ready to provide its organization with AI-related assurance and advisory services.

How can CAEs upskill the internal audit activity to be ready for the challenge? The first step is recognizing that new skillsets are required. Collectively, the internal audit activity must have a sufficient understanding of AI, how the organization is using it, and the risks that AI represents to the organization. The CAE must be able to communicate this understanding to senior management, the board, and the audit committee. A good place to start is with The IIA's thought leadership on AI, and The IIA's supplemental guidance on topics like big data and talent management.

Reemphasizing Cyber Resilience

Cybersecurity threats continue to define our times. The adoption and evolution of AI will force organizations to reemphasize their cyber resilience capabilities. As AI becomes more powerful and more decisions are handed off to new, complicated, and opaque algorithms, using huge data sets, protecting these systems from outside, malevolent forces is critical to success. A 2014 EY report defined cyber resilience as the ability to resist, react to, and recover from cyberattacks — and modify an environment to increase security and sustainability over time. Cyber resiliency is critical for any organization relying increasingly on AI.

Among all the complexity surrounding cybersecurity, there are four key areas where internal audit can have an immediate impact:

- Provide assurance over readiness and response to cyberthreats.
- Communicate to executive management and the board the level of risk to the organization and efforts to address such risks.
- Work collaboratively with IT and other parties to ensure effective defenses and responses are in place.
- Facilitate communication and coordination among all parties in the organization regarding the risk.

The potentially disastrous effects of a cybersecurity breach involving AI cannot be overstated. If cybersecurity competencies are not already in place, CAEs need to rapidly build the capacity within their teams.

AI Auditing Framework

The Framework is comprised of three components, AI Strategy, Governance, and the Human Factor.

AI Strategy

Each organization's AI strategy will be unique based on its approach to capitalizing on the opportunities AI provides. An organization's AI strategy might be an obvious extension of the organization's overall digital or big data strategy — organizations with a well-developed and implemented digital/big-

Audit Focus

IIA Standard 1210: Proficiency
(Excerpt)

Internal auditors must possess the knowledge, skills, and other competencies needed to perform their individual responsibilities. The internal audit activity collectively must possess or obtain the knowledge, skills, and other competencies needed to perform its responsibilities.

1210.A3 – Internal auditors must have sufficient knowledge of key information technology risks and controls and available technology-based audit techniques to perform their assigned work. However, not all internal auditors are expected to have the expertise of an internal auditor whose primary responsibility is information technology auditing.

data strategy are one step ahead in AI. According to MGI, organizations that "combine strong digital capability, robust AI adoption, and a proactive strategy see outsize financial performance."

Internal audit must consider an organization's AI strategy first. Does the organization have a defined strategy toward AI? Is it investing in AI research and development? Does it have plans in place to identify and address AI threats and opportunities? AI can become a competitive advantage for organizations, and internal audit should help management and the board realize the importance of formulating a *deliberate* AI strategy consistent with the organization's objectives.

Governance

AI governance refers to the structures, processes, and procedures implemented to direct, manage, and monitor the AI activities of the organization in pursuit of achieving the organization's objectives. The level of formality and structure for an organization's AI governance will vary based on the specific characteristics of that organization. Regardless of the specific approach, however, AI governance establishes accountability and oversight, helps to ensure that those responsible have the necessary skills and expertise to effectively monitor AI, and helps to ensure the organization's values are reflected in its AI activities. This last point should not be overlooked or given little attention. AI activities must result in decisions and actions that are in line with the ethical, social, and legal responsibilities of the organization.

Data Architecture & Infrastructure

AI data architecture and infrastructure will likely be one in the same as the organization's architecture and infrastructure for handling big data. It includes considerations for:

- The way that data is accessible (metadata, taxonomy, unique identifiers, and naming conventions).
- Information privacy and security throughout the data lifecycle (data collection, use, storage, and destruction).
- Roles and responsibilities for data ownership and use throughout the data lifecycle.

Data Quality

The completeness, accuracy, and reliability of the data on which AI algorithms are built are critical. Unfortunately, it is not unusual for organizations to have a poorly defined, incoherent structure to their data. Often, systems do not communicate with each other or do so through complicated add-ons or customizations. How this data is brought together, synthesized, and validated is crucial.

Measuring Performance

As organizations integrate AI into their activities, performance metrics should be defined to tie AI activities to business objectives and clearly illustrate whether AI is effectively supporting the achievement of those objectives. Management must actively monitor the performance of its AI activities.

The Human Factor

Algorithms are developed by humans. Human error and biases (both intentional and unintentional) will impact the performance of the algorithm. The human factor component considers whether:

- The risk of unintended human biases factored into AI design is identified and managed.
- AI has been effectively tested to ensure that results reflect the original objective.
- AI technologies can be transparent given the complexity involved.
- AI output is being used legally, ethically, and responsibly.

It is widely recognized that human error is the most common cause of information privacy and security breaches. Similarly, the human factor component addresses the risk of human error compromising the ability of AI to deliver the expected results.

The Black Box

According to the *Merriam-Webster* online dictionary, a black box is "a usually complicated electronic device whose internal mechanism is usually hidden from or mysterious to the user; *broadly*: anything that has mysterious or unknown internal functions or mechanisms." As organizations advance to implementing Type III and Type IV AI technologies — utilizing machines or platforms that can learn on their own or communicate with each other — how the algorithms are operating becomes less transparent or understandable. The black box factor will become more and more of a challenge as an organization's AI activities become more sophisticated.

Closing Thoughts

The internal auditing profession cannot be left behind in what may be the next digital frontier — artificial intelligence. To prepare, internal auditors must understand AI basics, the roles that internal audit can and should play, and AI risks and opportunities. To meet these challenges, internal auditors should leverage the Framework to deliver systematic, disciplined methods to evaluate and improve the effectiveness of risk management, control, and governance processes related to AI.

Audit Focus

Key IIA Standards

The IIA's *International Standards for the Professional Practice of Internal Auditing* includes several standards that are particularly relevant to AI, including:

IIA Standard 1210: Proficiency

IIA Standard 2010: Planning

IIA Standard 2030: Resource Management

IIA Standard 2100: Nature of Work

IIA Standard 2110: Governance

IIA Standard 2130: Control

IIA Standard 2200: Engagement Planning

IIA Standard 2201: Planning Considerations

IIA Standard 2210: Engagement Objectives

IIA Standard 2220: Engagement Scope

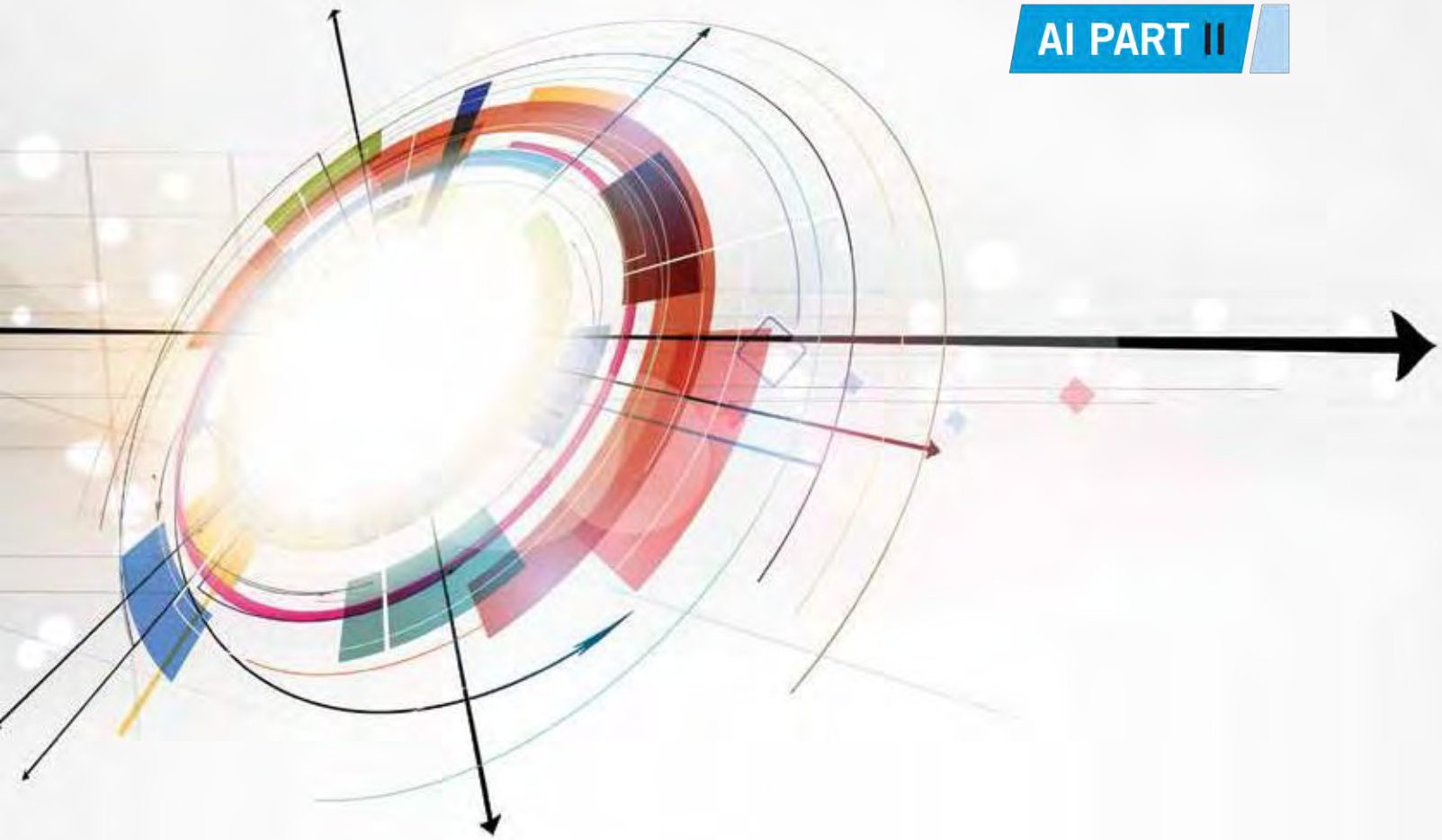
IIA Standard 2230: Engagement Resource Allocation

IIA Standard 2240: Engagement Work Program

IIA Standard 2310: Identifying Information

The Institute of Internal Auditors | Global

globaliia.org



AI PART II

GLOBAL PERSPECTIVES AND INSIGHTS

The IIA's Artificial Intelligence Auditing Framework

Practical Applications, Part A

Special Edition

The Institute of Internal Auditors | Global

Table of Contents

Previous Issues

To access previous issues of Global Perspectives and Insights, visit www.theiia.org/gpi.

Reader Feedback

Send questions or comments to globalperspectives@theiia.org.

Note: This is the second report in a three-part series. For more information, see the first report: Artificial Intelligence – Considerations for the Profession of Internal Auditing.

Introduction

A new Google project called AutoML is poised to take artificial intelligence (AI) — a broad term that refers to technologies that make machines "smart" — to another level. ML, short for machine learning, refers to computer algorithms that analyze data to learn to perform tasks. AutoML is a machine-learning algorithm that learns to build other machine-learning algorithms.

Google engineer Jeff Dean describes the project as a way for companies to build systems with AI even if they do not have extensive expertise. Only a few thousand companies today have the right talent for building AI, he estimates, but many more have the necessary data. "We want to go from thousands of organizations solving machine learning problems to millions," he told *The New York Times*.

Google is one of many organizations investing in AI research and applications to automate, augment, or replicate human intelligence — human analytical and/or decision-making. Following the creation path blazed by computer science, Microsoft recently unveiled a tool to help coders build "deep neural networks," a type of computer algorithm that eliminates "a lot of the heavy lifting," according to Joseph Sirosh, a vice president at Microsoft, in *The Times*. This focus on facilitating organizational AI initiatives means it is even more critical for the internal auditing profession to fully prepare for AI now.

There are many other terms related to AI besides machine learning, such as deep learning, image recognition, natural-language processing, cognitive computing, intelligence amplification, cognitive augmentation, machine augmented intelligence, and augmented intelligence. AI, as used in The IIA's AI Auditing Framework (Framework), encompasses all of these concepts.

The IIA's AI Auditing Framework

As explained in Artificial Intelligence – Considerations for the Profession of Internal Auditing, internal audit's role in AI is to "help an organization evaluate, understand, and communicate the degree to which artificial intelligence will have an effect (negative or positive) on the organization's ability to create value in the short, medium, or long term."

To help internal audit fulfill this role, internal auditors can leverage The IIA's AI Auditing Framework in providing AI-related advisory, assurance, or blended advisory/assurance services as appropriate to the organization. The Framework comprises three overarching components — AI Strategy, Governance, and the Human Factor — and seven elements: Cyber Resilience; AI Competencies; Data Quality; Data Architecture & Infrastructure; Measuring Performance; Ethics; and The Black Box.

Internal audit should consider numerous engagement or control objectives, and activities or procedures in implementing the Framework and providing

advisory, assurance, or blended advisory/assurance internal audit services related to the organization's AI activities. Relevant objectives and activities or procedures that address the Strategy (Cyber Resilience and AI Competencies elements) and Governance (Data Architecture & Infrastructure, and Data Quality elements) of the Framework are provided in this document. Relevant objectives and activities or procedures that address Governance (Measuring Performance element) and the Human Factor (Ethics and The Black Box elements) will be provided in Part III of this three-part series.

AI Strategy

Each organization's AI Strategy will be unique based on its approach to capitalizing on the opportunities AI provides. An organization's AI strategy might be an obvious extension of the organization's overall digital or big data strategy. The AI strategy should clearly articulate the intended result of AI activities. AI strategies should be developed collaboratively between the organization's business leaders who can articulate the intended result of AI activities and how those results relate to the organization's goals, and technology leaders who understand the organization's AI technology capabilities, constraints, and aspirations. Both business leaders and technology professionals also need to be involved in managing the execution of the AI strategy.

AI is dependent on big data, so an organization's big data strategy should be fully developed and implemented before it considers AI. Indeed, AI can help organizations capture insights from big data. As described in The IIA's Global Technology Audit Guide: Understanding and Auditing Big Data, by using these insights, "the organization can make better decisions, target new customers in creative and differentiating ways, service existing customers with a targeted and improved delivery model unique to the individual, and offer new services and capabilities that truly distinguish the company from its competitors." Organizations that capitalize on AI opportunities can develop a lasting competitive advantage, and the AI strategy should be developed and implemented against a backdrop of cyber resilience and AI competencies.

Before internal audit attempts to evaluate the organization's AI strategy, it should determine its own strategy for covering AI by including the topic in its risk assessment and considering whether AI should be included in the risk-based audit plan.

Cyber Resilience

The organization's ability to resist, react to, and recover from cyberattacks, including the intentional misuse of an organization's AI technologies for nefarious means, is becoming increasingly important (see Facebook's Corrective Actions on page 12). CAEs need to rapidly build cybersecurity competencies within their teams, continuously monitor AI/cybersecurity risks, and communicate to executive management and the board the level of risk to the organization and efforts to address such risk.

Relevant objectives and activities or procedures identified by The IIA do not comprise a prescribed audit plan, but are examples that should be useful in identifying engagement or control objectives, and in planning and performing AI audit engagements.

AI audit engagements should conform with IIA Standard 2200: Engagement Planning. AI audit plans and AI engagement objectives and procedures should always be customized to meet the needs of the organization.

AI Competencies

As noted in Artificial Intelligence – Considerations for the Profession of Internal Auditing, the pool of talent for technology professionals with AI expertise is reportedly small. Even if projects such as AutoML (see page 2) succeed, enabling organizations to build systems with AI even if they don't have extensive expertise, organizations will still need to fill a *knowledge* gap with staff who have a deep understanding of AI even if they cannot "do" AI. Staff need to:

- Know how AI works.
- Understand the risks and opportunities AI presents.
- Determine whether AI outcomes are as expected.
- Be capable of recommending or taking corrective action if needed.

Such competencies will be needed within internal audit and among the first and second lines of defense. Senior management and the board also should know how AI works and understand the risks and opportunities that AI presents.

Internal audit also should have the capability to determine if third-party providers of AI technologies are competent.

Relevant AI Strategy Objectives and Activities or Procedures

Engagement or Control Objective(s)	Activities or Procedures
Be actively involved in AI projects from their beginnings, providing advice and insight contributing to successful implementation.	**Attend** AI project team meetings.
The organization has a defined AI strategy.	**Determine** whether an AI strategy has been documented and if so, **verify** that the strategy: ■ Articulates the intended results of AI activities (strategic objectives). ■ Articulates at a high level how the AI objectives will be accomplished (strategic plan).
Provide assurance over the readiness and response to cyber threats.	**Leveraging** an established cybersecurity framework, **work** collaboratively with IT and other parties to ensure effective defenses and responses are in place.
There are sufficient resources (staff and budget) to implement the AI strategy.	**Review** process for determining staff and budget needs to support AI.
Advise on whether the strategy adequately considers AI threats and opportunities.	**Review** any existing assessments of AI threats and opportunities. If no assessments exist, **make recommendations** for moving forward (how the organization could plan to identify AI threats and opportunities).

Governance

AI governance refers to the structures, processes, and procedures implemented to direct, manage, and monitor the AI activities of the organization. Governance structure and formality will vary based on the specific characteristics of the organization. AI governance:

- Establishes accountability, responsibility, and oversight.
- Helps to ensure that those with AI responsibilities have the necessary skills and expertise.
- Helps to ensure that AI activities and AI-related decisions and actions are consistent with the organization's values, and ethical, social, and legal responsibilities.

AI policies and procedures should be established for the entire AI life cycle — from inputs to outputs. Policies and procedures also should be established for training, measuring performance, and reporting.

Accountability, Responsibility, and Oversight

AI has the potential to do great good and great harm. Ultimately, stakeholders will likely hold the board and senior executives accountable (answerable) for their organization's AI outcomes. When assessing AI governance, internal auditors can leverage the three lines of defense model. The three lines of defense, along with senior management, the governing body, external auditors, and regulators all have roles in AI. Internal auditors should understand the role of each party, and how internal audit interfaces with that role.

"The IIA's Artificial Intelligence Auditing Framework is a practical tool for helping internal audit to provide independent assurance over AI risk management, control, and governance processes."

Nur Hayati Baharuddin,
Member, IIA–Malaysia

Regulators

Regulators inform and control specific activities (such as banking, health care, or food safety) at national, regional/state, and local levels. Regulators "inform" through activities such as conducting research, participating in the development of standards and guidance, and communicating with stakeholders. Regulators "control" through activities such as supervising, and setting and enforcing regulations. As stated in The IIA's Position Paper: The Three Lines of Defense in Effective Risk Management and Control, regulators sometimes set requirements intended to strengthen controls in an organization and on other occasions perform an independent and objective function to assess the whole or some part of the first, second, or third line of defense with regard to those requirements.

To date, there are no regulations dedicated exclusively to AI. However, parts of existing regulations may be particularly relevant to AI activities, and regulators and standard setting bodies around the world have signaled their concern through research, discussion papers, recommendations, and guidance (see Regulatory Compliance on page 7).

Regulators already recognize the importance of AI audits. For example, in its guidance on Off-The-Shelf Software Use in Medical Devices, the U.S. Food and

Drug Administration recognizes the importance of auditing OTS knowledge-based software (for example, artificial intelligence, expert systems, and neural net software), stating that the manufacturer is expected to provide assurance "that the product development methodologies used by the OTS Software developer are appropriate and sufficient for the intended use…" and "recommends this include an audit of the OTS Software developer's design and development methodologies used in the construction of the OTS Software. This audit should thoroughly assess the development and qualification documentation generated for the OTS Software."

Auditors should keep apprised of the work of regulators and standard-setters in the area of AI, advise management and the board of matters of importance, and assess whether the organization's regulatory control objectives reflect emerging regulations, standards, and guidance.

"In addition to providing assurance over AI activities, internal audit should ensure audit committees and boards are equipped to understand their role in navigating the benefits and risks associated with AI in the companies they serve."

Carolyn Saint, CAE,
University of Virginia

Governing Body/Board/Audit Committee

The board is responsible for the ultimate oversight of the organization's AI activities. The board should be involved with senior management in defining the organization's AI strategy.

Internal audit must understand and be well-informed about AI generally, and the organization's AI activities specifically. In addition to providing assurance over AI activities, internal audit should offer advice and insights to help ensure that the board is prepared for its role.

Senior Management

Working with the board, senior management defines the organization's AI strategy. Senior management also sets AI objectives and develops plans to implement the AI strategy.

Internal audit should be represented on the senior management team, and should keep well-informed of senior management's AI initiatives. Regarding AI risk management, governance, and controls, internal audit should be a trusted advisor to senior management.

First Line of Defense

Operational managers should own and manage AI risks on a day-to-day basis. Internal audit should assess operational-level AI policies and procedures, verifying that control objectives are adequate and working as designed.

Second Line of Defense

Compliance, ethics, risk management, and information privacy/security are some of the second line of defense functions that likely will oversee some aspect of AI risks. Internal audit should assess second line of defense AI-related policies and procedures, verifying that control objectives are adequate and working as designed.

Third Line of Defense

Internal audit should provide independent assurance over AI risks, governance, and controls. The IIA's AI Auditing Framework can facilitate this role. Regulators and standard-setters have recognized the potential of AI in risk management and compliance. According to the Financial Stabilities Board (FSB) report Artificial intelligence and machine learning in financial services, "The use of AI and machine learning in financial services may bring key benefits for financial stability in the form of efficiencies in the provision of financial services and regulatory systemic risk surveillance… The internal (back-office) applications of AI and machine learning could improve risk management, fraud detection, and compliance with regulatory requirements, potentially at lower cost." Similarly, the most advanced internal audit departments will start to use algorithms to fuel their continuous auditing and continuous monitoring initiatives, gaining both effectiveness and efficiency.

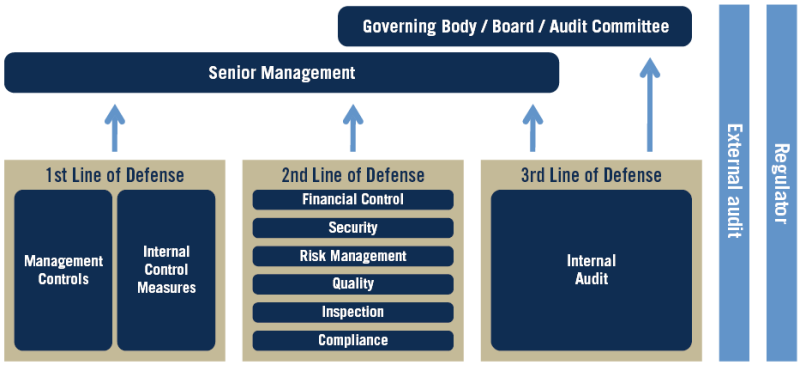
External Audit

External auditors are third parties with no vested interest in the organization, and express an opinion on whether financial statements are prepared in accordance with applicable financial reporting frameworks and/or regulations. Regarding AI, external auditors will most likely focus on outcomes — for example, the algorithms behind model risk management or valuation, and whether those algorithms have a material impact on the organization's financial statements.

"Emerging use of AI requires that audit needs specifically to address the logic used in the design of the algorithms."

Hans Nieuwlands, CEO, IIA–Netherlands

The Three Lines of Defense Model



Adapted from ECIIA/FERMA *Guidance on the 8th EU Company Law Directive, article 41*

Regulatory Compliance

Regulations typically lag technological change, and AI is no exception. However, as reported by The Hill, Tesla CEO Elon Musk warned the National Governors Association (U.S.) that regulations are needed sooner rather than later. In addition, privacy regulations such as the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the EU's General Data Protection Regulation (GDPR), effective May 2018, may complicate AI implementation. Both regulations protect personally identifiable information, which typically are inputs to AI technologies.

For example, the HIPAA Privacy Rule "set national standards for the protection of health information, as applied to the three types of covered entities: health plans, health care clearinghouses, and health care providers who conduct certain health care transactions **electronically** [emphasis added]." And according to the FSB report Artificial intelligence and machine learning in financial services, "several sections of the GDPR are particularly relevant to AI: Article 11 provides a right to 'an explanation of the decision reached after [algorithmic] assessment"; Article 9 prohibits the processing of "special [sensitive] categories of personal data"; Article 22 provides for a data subject's qualified right not to be subject to a decision with legal or significant consequences based solely on automated processing; and Article 24 provides that decisions shall not be based on special categories of personal data.

Other generally recognized regulatory concerns include compliance with anti-discrimination laws and legal liabilities, especially with regard to third parties who provide the organization with AI services. The FSB summed up concerns regarding third parties by saying "Many current providers of AI and machine learning in financial services may fall outside the regulatory perimeter or may not be familiar with applicable law and regulation. Where financial institutions rely on third-party providers of AI and machine learning services for critical functions, and rules on outsourcing may not be in place or not be understood, these servicers and providers may not be subject to supervision and oversight. Similarly, if providers of such tools begin providing financial services to institutional or retail clients, this could entail financial activities taking place outside the regulatory perimeter."

Organizations should not wait until the regulatory environment catches up to the technology environment. Even if existing regulations do not specifically address AI, the *letter* of the law, organizations should ask whether or not their AI activities are consistent with the *spirit* of existing laws. One approach is to perform scenario and "what if?" analyses to determine if AI activities could potentially be used for malicious or criminal activities, or result in unintended consequences that cause harm. Those responsible for governance also should consider that AI activities may potentially diminish internal controls if the AI learns to override established rules or if AI systems learn how to communicate with each other and "work" together without the organization's knowledge. A proactive approach in considering the spirit of existing laws will help organizations be agile as new regulations are enacted and become effective.

Relevant AI Governance Objectives and Activities or Procedures	
Engagement or Control Objective(s)	**Activities or Procedures**
Provide assurance that AI governance structures have been established, documented, and are working as designed.	**Review** business models and organizational structure; **determine** if business models and organizational structure reflect the organization's AI strategy. **Review** AI policies and procedures; **determine** whether organizational policies and procedures clearly identify AI roles and responsibilities related to AI strategy, governance, data architecture, data quality, ethical imperatives, and measuring performance.
Assess whether those with AI responsibilities have the necessary competencies to be successful. For example, those responsible for ethical imperatives should be competent in assessing the ethical behavior of those who provide human input into the AI, and should be independent of the AI activity.	**Interview** those with AI responsibilities. **Review** AI job descriptions, requisite skills, etc., and **verify** whether those responsible have their stated qualifications.
Provide assurance that AI policies and procedures have been established and documented.	**Review** AI policies and procedures and **determine** if they sufficiently address AI risks. **Determine** if policies and procedures provide for periodic "what if" analysis or scenario planning.
Provide assurance that AI activity audit trails provide sufficient information to understand what AI decisions were made, and why.	**Review** AI audit trails. **Determine** whether audit trails provide sufficient information to understand what decisions were made, and why.
Provide assurance that policies and procedures have been implemented and are working as designed, and that employees are compliant.	**Observe** employees implementing AI procedures. **Review** helpline/hotline reports and **follow up** on any reports alleging noncompliant or malicious activities related to AI. **Interview** a random sample of employees and **determine** if they are knowledgeable about AI policies and procedures. **Identify** and **review** AI access policies and procedures. **Evaluate** access policies and **test** access controls. **Assess** whether regulatory control objectives reflect emerging regulations, standards, and guidance.

"Data Infrastructure & Architecture and Data Quality are often intertwined. Relevant engagement or control objectives, and activities and procedures in one area, may overlap or impact objectives, activities, and procedures in the other area."

Lesedi Lesetedi,

Deputy Executive Director (Deputy CEO) – Strategy & Corporate Services

Botswana College of Distance & Open Learning (BOCODOL)

Data Architecture & Infrastructure

AI data architecture and infrastructure will likely be one and the same, or at least nearly the same, as the organization's architecture and infrastructure for handling big data. It includes considerations for:

- The way that data is accessible (metadata, taxonomy, unique identifiers, and naming conventions).
- Information privacy and security throughout the data lifecycle (data collection, use, storage, and destruction).
- Roles and responsibilities for data ownership and use throughout the data life cycle.

According to InfoWorld, organizations should focus on three major areas of software development to ensure the success of AI integration:

- Data integration — data from multiple sources must be integrated before AI can be incorporated into the organization's applications and systems.
- Application modernization — software updates will need to be made on a regular basis. Frequent, less intensive updates should replace infrequent, more intensive updates that slow down or disrupt systems.
- Employee education — software developers, project managers, and other technology staff need to keep up with machine learning and every aspect of the technology "stack" (the software and components that run AI).

In addition, data should be reconciled so that nuances such as rounding, demographics, and other variables are normalized before input.

Relevant Data Architecture & Infrastructure Objectives and Activities or Procedures

Engagement or Control Objective(s)	Activities or Procedures
Provide assurance that the organization is cyber resilient. Cyber resilience includes, but is broader than, cybersecurity alone. Cyber resilience encompasses security (resistance), reaction, and recovery.	**Understand** and audit big data (see The IIA's Practice Guide: Understanding and Auditing Big Data). **Assess** whether the organization is preparing for compliance with new technology regulations, such as the EU's General Data Protection Regulation (GDPR). **Assess** whether the organization's disaster recovery protocols include AI failures, including the breakdown of controls that maintain the rules set forth by AI governance.
Provide assurance that the data infrastructure has the capacity to accommodate the size and complexity of AI activity set forth in the AI strategy.	**Assess** whether the infrastructure is capable of handling structured and unstructured data.
Provide assurance that the organization has established a data taxonomy. **Evaluate** the quality, completeness, and consistency of use for the enterprisewide data taxonomy.	**Assess** whether the taxonomy is robust enough to accommodate the size and complexity of AI activities.

Data Quality

The completeness, accuracy, and reliability of the data on which AI algorithms are built are critical. For AI to be successful, organizations need access to vast amounts of high quality data — data that is well-defined and in standardized formats. Often, systems do not communicate with each other or do so through complicated add-ons or customizations. How this data is reconciled, synthesized, and validated is also critical, so systems that do not communicate with each other or do so through complicated add-ons or customizations may thwart an organization's AI activities.

In addition to data that is well-defined in standardized formats (structured data), AI technologies may be dependent on unstructured data (such as social media posts). As described in The IIA's "Global Technology Audit Guide: Understanding and Auditing Big Data," unstructured data is "typically more difficult to manage, due to its evolving and unpredictable nature, and it is usually sourced from large, disparate, and often external data sources. Consequently, new solutions have been developed to manage and analyze this data."

Ironically, organizations can turn to machine learning — a form of AI — to improve data quality. For example, there may be multiple versions of a vendor's name across an organization's many business units, data bases, and spreadsheets. A computer program could scan and reconcile all variations of the name in a matter of hours or minutes.

Internal audit also should look at how data that is used in internal audit reports has been reconciled, synthesized, and validated.

Relevant Data Quality Objectives and Activities or Procedures

Engagement or Control Objective(s)	Activities or Procedures
Provide assurance over the reliability of AI's underlying algorithms and the data on which algorithms are based.	**Obtain** a sample of the raw data that are inputs to AI. **Verify** that the organization has implemented methodologies to validate AI outcomes with actual, real-world outcomes, and that policies and procedures are in place to continuously measure, monitor, escalate, and rectify inconsistencies between the two.
Provide assurance that data input is reconciled and normalized to maximize accuracy.	**Verify** that the organization has policies and procedures in place to continuously measure, monitor, escalate, and rectify data accuracy and integrity issues. **Confirm** that the organization is consistently following and monitoring a formalized data reconciliation framework, which includes a rationale for differing methodologies and results should they exist.
Provide assurance that aggregated data is complete.	**Verify** that the organization has policies and procedures in place to limit data input bias.
Provide assurance that the completeness of data is measured and monitored and that any material exceptions that impact decision-making are identified and explained. This should be done whether the exceptions are determined by humans or AI.	**Review** AI metrics and metric reports. **Assess** whether those responsible for decision-making have received and considered explanations on material exceptions related to data quality.

Facebook's Corrective Actions

Facebook's challenges with AI have been widely reported. The behemoth social network has been under scrutiny over how its algorithm-fueled technologies have been used — or misused — for malicious means.

Timeline of Major Concerns:

- In Fall 2016, ProPublica reported that advertisers could use Facebook's ad-targeting tools to exclude certain races — a potential violation of federal housing and civil rights regulations.
- In September 2017:
    - Facebook disclosed that holders of fake accounts based in Russia purchased sizeable ads on divisive issues leading up to the 2016 presidential election.
    - ProPublica reported that Facebook's ad targeting tools enabled advertisers to target self-described ethnic "haters."
- In October 2017, concerns about fake news resurfaced when Facebook (and Google) posted false information about the mass shooting in Las Vegas.
- In testimony to a Senate judiciary subcommittee in late October, Facebook said the reach of Russia-backed ads stretched much further than they had originally known, reaching as many as 126 million Americans before and during the 2016 presidential election.

Facebook's Response

In a Sept. 20, 2017, post, Facebook Chief Operating Officer Sheryl Sandberg announced three corrective actions:

1. Facebook is clarifying its advertising policies and tightening its enforcement processes to ensure that content that goes against Facebook's community standards cannot be used to target ads. (Policies and processes relate to the Governance component of the Framework. Among other things, AI Governance should establish accountability for and oversight of enforcement.)

2. Facebook is increasing "human review and oversight" of its automated processes. (Human review and oversight relates to the Ethics component of the Framework. Among other things, Ethics addresses whether AI results reflect the original objective and whether AI output is being used legally, ethically, and responsibly.)

3. Facebook is working on a program that will encourage Facebook users to report potential abuses of its ads systems. (Reporting systems relate to the Measuring Performance component of the Framework. Reporting systems help management monitor the performance of AI activities. Measuring Performance will be covered in a future Global Perspectives and Insights report.)

By utilizing The IIA's AI Auditing Framework, internal auditors can provide assurance and advisory services to help organizations separate truth from fiction and address reporting, operations, and compliance risks associated with AI.

Using the Standards to Audit AI

Internal auditors should conform with all applicable IIA standards when planning or performing AI engagements. Key IIA standards that are particularly relevant to AI are highlighted in the sidebar, but others may apply as well.

Each standard is complemented by an Implementation Guide. Implementation guides assist internal auditors in applying the *Standards*. They collectively address internal auditing's approach, methodologies, and consideration, but do not detail processes or procedures.

Closing Thoughts

The IIA's Artificial Intelligence Auditing Framework will help internal auditors approach AI advisory and assurance services in a systematic and disciplined manner. Whether the organization's AI technologies and activities are developed in-house, through a facilitative technology such as AutoML, or by a third party, internal audit should be prepared to advise the board and senior management, coordinate with the first and second lines of defense, and provide assurance over AI risk management, governance, and controls.

This paper is Part II of a three-part series. It provides suggestions for implementing the AI Strategy and Governance components of The IIA's AI Auditing Framework. Part III will provide further suggestions for implementing the Governance component, and the Human Factor component.

Audit Focus

Key IIA Standards

The IIA's *International Standards for the Professional Practice of Internal Auditing* includes several standards that are particularly relevant to AI, including:

IIA Standard 1210: Proficiency

IIA Standard 2010: Planning

IIA Standard 2030: Resource Management

IIA Standard 2100: Nature of Work

IIA Standard 2110: Governance

IIA Standard 2130: Control

IIA Standard 2200: Engagement Planning

IIA Standard 2201: Planning Considerations

IIA Standard 2210: Engagement Objectives

IIA Standard 2220: Engagement Scope

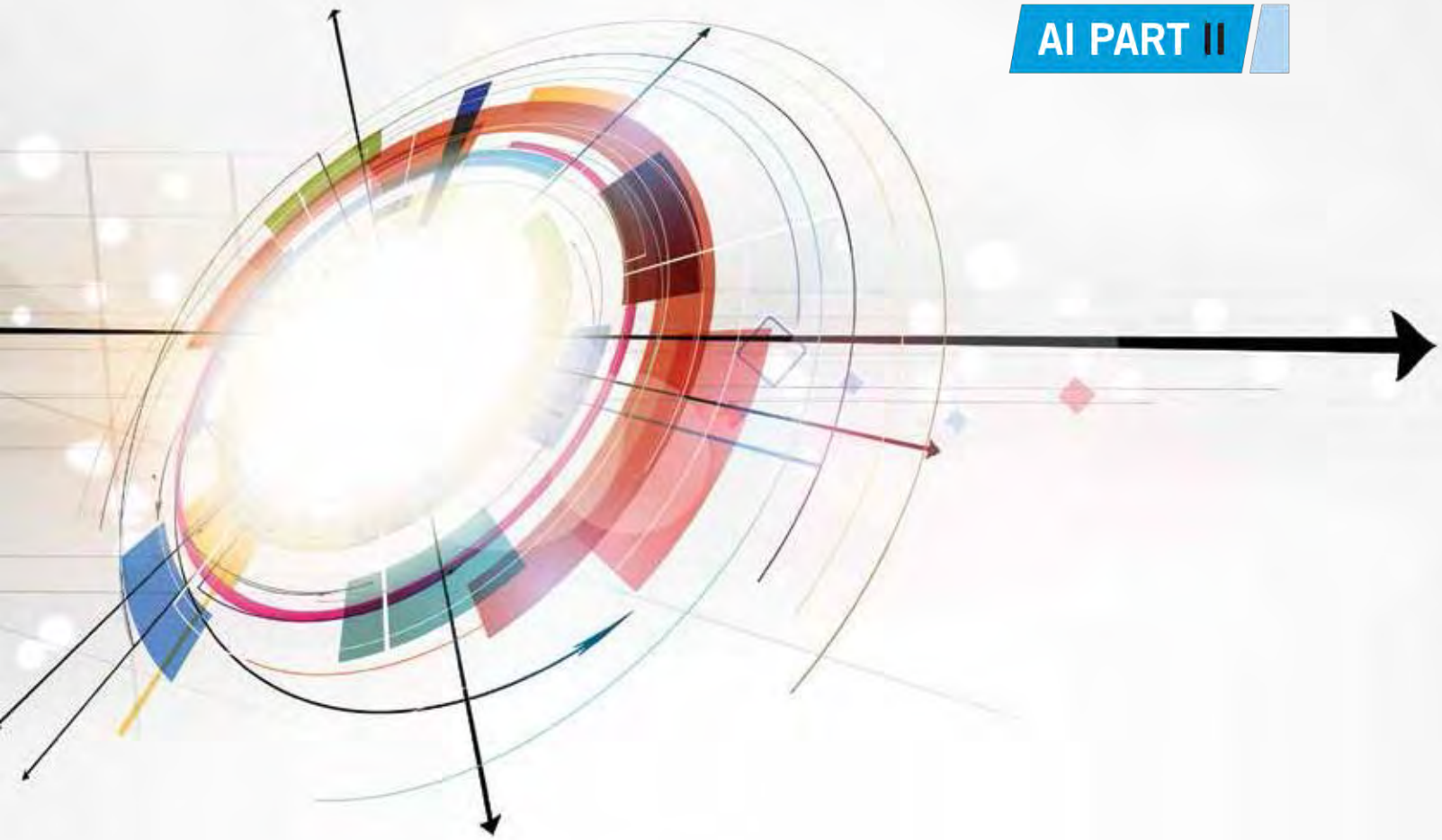
IIA Standard 2230: Engagement Resource Allocation

IIA Standard 2240: Engagement Work Program

IIA Standard 2310: Identifying Information

The Institute of Internal Auditors | Global

globaliia.org



AI PART II

GLOBAL PERSPECTIVES AND INSIGHTS

The IIA's Artificial Intelligence Auditing Framework

Practical Applications, Part A

Special Edition

The Institute of Internal Auditors | Global

Table of Contents

Previous Issues

To access previous issues of Global Perspectives and Insights, visit www.theiia.org/gpi.

Reader Feedback

Send questions or comments to globalperspectives@theiia.org.

Note: This is the second report in a three-part series. For more information, see the first report: Artificial Intelligence – Considerations for the Profession of Internal Auditing.

Introduction

A new Google project called AutoML is poised to take artificial intelligence (AI) — a broad term that refers to technologies that make machines "smart" — to another level. ML, short for machine learning, refers to computer algorithms that analyze data to learn to perform tasks. AutoML is a machine-learning algorithm that learns to build other machine-learning algorithms.

Google engineer Jeff Dean describes the project as a way for companies to build systems with AI even if they do not have extensive expertise. Only a few thousand companies today have the right talent for building AI, he estimates, but many more have the necessary data. "We want to go from thousands of organizations solving machine learning problems to millions," he told *The New York Times*.

Google is one of many organizations investing in AI research and applications to automate, augment, or replicate human intelligence — human analytical and/or decision-making. Following the creation path blazed by computer science, Microsoft recently unveiled a tool to help coders build "deep neural networks," a type of computer algorithm that eliminates "a lot of the heavy lifting," according to Joseph Sirosh, a vice president at Microsoft, in *The Times*. This focus on facilitating organizational AI initiatives means it is even more critical for the internal auditing profession to fully prepare for AI now.

There are many other terms related to AI besides machine learning, such as deep learning, image recognition, natural-language processing, cognitive computing, intelligence amplification, cognitive augmentation, machine augmented intelligence, and augmented intelligence. AI, as used in The IIA's AI Auditing Framework (Framework), encompasses all of these concepts.

The IIA's AI Auditing Framework

As explained in Artificial Intelligence – Considerations for the Profession of Internal Auditing, internal audit's role in AI is to "help an organization evaluate, understand, and communicate the degree to which artificial intelligence will have an effect (negative or positive) on the organization's ability to create value in the short, medium, or long term."

To help internal audit fulfill this role, internal auditors can leverage The IIA's AI Auditing Framework in providing AI-related advisory, assurance, or blended advisory/assurance services as appropriate to the organization. The Framework comprises three overarching components — AI Strategy, Governance, and the Human Factor — and seven elements: Cyber Resilience; AI Competencies; Data Quality; Data Architecture & Infrastructure; Measuring Performance; Ethics; and The Black Box.

Internal audit should consider numerous engagement or control objectives, and activities or procedures in implementing the Framework and providing

advisory, assurance, or blended advisory/assurance internal audit services related to the organization's AI activities. Relevant objectives and activities or procedures that address the Strategy (Cyber Resilience and AI Competencies elements) and Governance (Data Architecture & Infrastructure, and Data Quality elements) of the Framework are provided in this document. Relevant objectives and activities or procedures that address Governance (Measuring Performance element) and the Human Factor (Ethics and The Black Box elements) will be provided in Part III of this three-part series.

AI Strategy

Each organization's AI Strategy will be unique based on its approach to capitalizing on the opportunities AI provides. An organization's AI strategy might be an obvious extension of the organization's overall digital or big data strategy. The AI strategy should clearly articulate the intended result of AI activities. AI strategies should be developed collaboratively between the organization's business leaders who can articulate the intended result of AI activities and how those results relate to the organization's goals, and technology leaders who understand the organization's AI technology capabilities, constraints, and aspirations. Both business leaders and technology professionals also need to be involved in managing the execution of the AI strategy.

AI is dependent on big data, so an organization's big data strategy should be fully developed and implemented before it considers AI. Indeed, AI can help organizations capture insights from big data. As described in The IIA's Global Technology Audit Guide: Understanding and Auditing Big Data, by using these insights, "the organization can make better decisions, target new customers in creative and differentiating ways, service existing customers with a targeted and improved delivery model unique to the individual, and offer new services and capabilities that truly distinguish the company from its competitors." Organizations that capitalize on AI opportunities can develop a lasting competitive advantage, and the AI strategy should be developed and implemented against a backdrop of cyber resilience and AI competencies.

Before internal audit attempts to evaluate the organization's AI strategy, it should determine its own strategy for covering AI by including the topic in its risk assessment and considering whether AI should be included in the risk-based audit plan.

Cyber Resilience

The organization's ability to resist, react to, and recover from cyberattacks, including the intentional misuse of an organization's AI technologies for nefarious means, is becoming increasingly important (see Facebook's Corrective Actions on page 12). CAEs need to rapidly build cybersecurity competencies within their teams, continuously monitor AI/cybersecurity risks, and communicate to executive management and the board the level of risk to the organization and efforts to address such risk.

Relevant objectives and activities or procedures identified by The IIA do not comprise a prescribed audit plan, but are examples that should be useful in identifying engagement or control objectives, and in planning and performing AI audit engagements.

AI audit engagements should conform with IIA Standard 2200: Engagement Planning. AI audit plans and AI engagement objectives and procedures should always be customized to meet the needs of the organization.

AI Competencies

As noted in Artificial Intelligence – Considerations for the Profession of Internal Auditing, the pool of talent for technology professionals with AI expertise is reportedly small. Even if projects such as AutoML (see page 2) succeed, enabling organizations to build systems with AI even if they don't have extensive expertise, organizations will still need to fill a *knowledge* gap with staff who have a deep understanding of AI even if they cannot "do" AI. Staff need to:

- Know how AI works.
- Understand the risks and opportunities AI presents.
- Determine whether AI outcomes are as expected.
- Be capable of recommending or taking corrective action if needed.

Such competencies will be needed within internal audit and among the first and second lines of defense. Senior management and the board also should know how AI works and understand the risks and opportunities that AI presents.

Internal audit also should have the capability to determine if third-party providers of AI technologies are competent.

Relevant AI Strategy Objectives and Activities or Procedures

Engagement or Control Objective(s)	Activities or Procedures
Be actively involved in AI projects from their beginnings, providing advice and insight contributing to successful implementation.	**Attend** AI project team meetings.
The organization has a defined AI strategy.	**Determine** whether an AI strategy has been documented and if so, **verify** that the strategy: ■ Articulates the intended results of AI activities (strategic objectives). ■ Articulates at a high level how the AI objectives will be accomplished (strategic plan).
Provide assurance over the readiness and response to cyber threats.	**Leveraging** an established cybersecurity framework, **work** collaboratively with IT and other parties to ensure effective defenses and responses are in place.
There are sufficient resources (staff and budget) to implement the AI strategy.	**Review** process for determining staff and budget needs to support AI.
Advise on whether the strategy adequately considers AI threats and opportunities.	**Review** any existing assessments of AI threats and opportunities. If no assessments exist, **make recommendations** for moving forward (how the organization could plan to identify AI threats and opportunities).

Governance

AI governance refers to the structures, processes, and procedures implemented to direct, manage, and monitor the AI activities of the organization. Governance structure and formality will vary based on the specific characteristics of the organization. AI governance:

- Establishes accountability, responsibility, and oversight.
- Helps to ensure that those with AI responsibilities have the necessary skills and expertise.
- Helps to ensure that AI activities and AI-related decisions and actions are consistent with the organization's values, and ethical, social, and legal responsibilities.

AI policies and procedures should be established for the entire AI life cycle — from inputs to outputs. Policies and procedures also should be established for training, measuring performance, and reporting.

Accountability, Responsibility, and Oversight

AI has the potential to do great good and great harm. Ultimately, stakeholders will likely hold the board and senior executives accountable (answerable) for their organization's AI outcomes. When assessing AI governance, internal auditors can leverage the three lines of defense model. The three lines of defense, along with senior management, the governing body, external auditors, and regulators all have roles in AI. Internal auditors should understand the role of each party, and how internal audit interfaces with that role.

"The IIA's Artificial Intelligence Auditing Framework is a practical tool for helping internal audit to provide independent assurance over AI risk management, control, and governance processes."

Nur Hayati Baharuddin,
Member, IIA–Malaysia

Regulators

Regulators inform and control specific activities (such as banking, health care, or food safety) at national, regional/state, and local levels. Regulators "inform" through activities such as conducting research, participating in the development of standards and guidance, and communicating with stakeholders. Regulators "control" through activities such as supervising, and setting and enforcing regulations. As stated in The IIA's Position Paper: The Three Lines of Defense in Effective Risk Management and Control, regulators sometimes set requirements intended to strengthen controls in an organization and on other occasions perform an independent and objective function to assess the whole or some part of the first, second, or third line of defense with regard to those requirements.

To date, there are no regulations dedicated exclusively to AI. However, parts of existing regulations may be particularly relevant to AI activities, and regulators and standard setting bodies around the world have signaled their concern through research, discussion papers, recommendations, and guidance (see Regulatory Compliance on page 7).

Regulators already recognize the importance of AI audits. For example, in its guidance on Off-The-Shelf Software Use in Medical Devices, the U.S. Food and

Drug Administration recognizes the importance of auditing OTS knowledge-based software (for example, artificial intelligence, expert systems, and neural net software), stating that the manufacturer is expected to provide assurance "that the product development methodologies used by the OTS Software developer are appropriate and sufficient for the intended use…" and "recommends this include an audit of the OTS Software developer's design and development methodologies used in the construction of the OTS Software. This audit should thoroughly assess the development and qualification documentation generated for the OTS Software."

Auditors should keep apprised of the work of regulators and standard-setters in the area of AI, advise management and the board of matters of importance, and assess whether the organization's regulatory control objectives reflect emerging regulations, standards, and guidance.

"In addition to providing assurance over AI activities, internal audit should ensure audit committees and boards are equipped to understand their role in navigating the benefits and risks associated with AI in the companies they serve."

Carolyn Saint, CAE,
University of Virginia

Governing Body/Board/Audit Committee

The board is responsible for the ultimate oversight of the organization's AI activities. The board should be involved with senior management in defining the organization's AI strategy.

Internal audit must understand and be well-informed about AI generally, and the organization's AI activities specifically. In addition to providing assurance over AI activities, internal audit should offer advice and insights to help ensure that the board is prepared for its role.

Senior Management

Working with the board, senior management defines the organization's AI strategy. Senior management also sets AI objectives and develops plans to implement the AI strategy.

Internal audit should be represented on the senior management team, and should keep well-informed of senior management's AI initiatives. Regarding AI risk management, governance, and controls, internal audit should be a trusted advisor to senior management.

First Line of Defense

Operational managers should own and manage AI risks on a day-to-day basis. Internal audit should assess operational-level AI policies and procedures, verifying that control objectives are adequate and working as designed.

Second Line of Defense

Compliance, ethics, risk management, and information privacy/security are some of the second line of defense functions that likely will oversee some aspect of AI risks. Internal audit should assess second line of defense AI-related policies and procedures, verifying that control objectives are adequate and working as designed.

Third Line of Defense

Internal audit should provide independent assurance over AI risks, governance, and controls. The IIA's AI Auditing Framework can facilitate this role. Regulators and standard-setters have recognized the potential of AI in risk management and compliance. According to the Financial Stabilities Board (FSB) report Artificial intelligence and machine learning in financial services, "The use of AI and machine learning in financial services may bring key benefits for financial stability in the form of efficiencies in the provision of financial services and regulatory systemic risk surveillance… The internal (back-office) applications of AI and machine learning could improve risk management, fraud detection, and compliance with regulatory requirements, potentially at lower cost." Similarly, the most advanced internal audit departments will start to use algorithms to fuel their continuous auditing and continuous monitoring initiatives, gaining both effectiveness and efficiency.

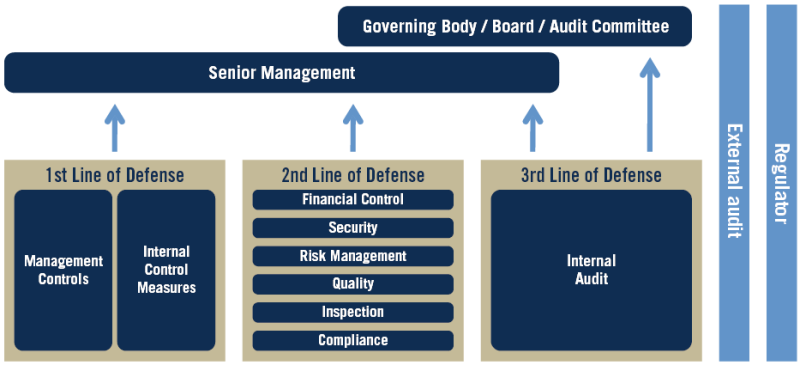
External Audit

External auditors are third parties with no vested interest in the organization, and express an opinion on whether financial statements are prepared in accordance with applicable financial reporting frameworks and/or regulations. Regarding AI, external auditors will most likely focus on outcomes — for example, the algorithms behind model risk management or valuation, and whether those algorithms have a material impact on the organization's financial statements.

"Emerging use of AI requires that audit needs specifically to address the logic used in the design of the algorithms."

Hans Nieuwlands, CEO, IIA–Netherlands

The Three Lines of Defense Model



Adapted from ECIIA/FERMA *Guidance on the 8th EU Company Law Directive, article 41*

Regulatory Compliance

Regulations typically lag technological change, and AI is no exception. However, as reported by The Hill, Tesla CEO Elon Musk warned the National Governors Association (U.S.) that regulations are needed sooner rather than later. In addition, privacy regulations such as the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the EU's General Data Protection Regulation (GDPR), effective May 2018, may complicate AI implementation. Both regulations protect personally identifiable information, which typically are inputs to AI technologies.

For example, the HIPAA Privacy Rule "set national standards for the protection of health information, as applied to the three types of covered entities: health plans, health care clearinghouses, and health care providers who conduct certain health care transactions **electronically** [emphasis added]." And according to the FSB report Artificial intelligence and machine learning in financial services, "several sections of the GDPR are particularly relevant to AI: Article 11 provides a right to 'an explanation of the decision reached after [algorithmic] assessment"; Article 9 prohibits the processing of "special [sensitive] categories of personal data"; Article 22 provides for a data subject's qualified right not to be subject to a decision with legal or significant consequences based solely on automated processing; and Article 24 provides that decisions shall not be based on special categories of personal data.

Other generally recognized regulatory concerns include compliance with anti-discrimination laws and legal liabilities, especially with regard to third parties who provide the organization with AI services. The FSB summed up concerns regarding third parties by saying "Many current providers of AI and machine learning in financial services may fall outside the regulatory perimeter or may not be familiar with applicable law and regulation. Where financial institutions rely on third-party providers of AI and machine learning services for critical functions, and rules on outsourcing may not be in place or not be understood, these servicers and providers may not be subject to supervision and oversight. Similarly, if providers of such tools begin providing financial services to institutional or retail clients, this could entail financial activities taking place outside the regulatory perimeter."

Organizations should not wait until the regulatory environment catches up to the technology environment. Even if existing regulations do not specifically address AI, the *letter* of the law, organizations should ask whether or not their AI activities are consistent with the *spirit* of existing laws. One approach is to perform scenario and "what if?" analyses to determine if AI activities could potentially be used for malicious or criminal activities, or result in unintended consequences that cause harm. Those responsible for governance also should consider that AI activities may potentially diminish internal controls if the AI learns to override established rules or if AI systems learn how to communicate with each other and "work" together without the organization's knowledge. A proactive approach in considering the spirit of existing laws will help organizations be agile as new regulations are enacted and become effective.

Relevant AI Governance Objectives and Activities or Procedures	
Engagement or Control Objective(s)	**Activities or Procedures**
Provide assurance that AI governance structures have been established, documented, and are working as designed.	**Review** business models and organizational structure; **determine** if business models and organizational structure reflect the organization's AI strategy. **Review** AI policies and procedures; **determine** whether organizational policies and procedures clearly identify AI roles and responsibilities related to AI strategy, governance, data architecture, data quality, ethical imperatives, and measuring performance.
Assess whether those with AI responsibilities have the necessary competencies to be successful. For example, those responsible for ethical imperatives should be competent in assessing the ethical behavior of those who provide human input into the AI, and should be independent of the AI activity.	**Interview** those with AI responsibilities. **Review** AI job descriptions, requisite skills, etc., and **verify** whether those responsible have their stated qualifications.
Provide assurance that AI policies and procedures have been established and documented.	**Review** AI policies and procedures and **determine** if they sufficiently address AI risks. **Determine** if policies and procedures provide for periodic "what if" analysis or scenario planning.
Provide assurance that AI activity audit trails provide sufficient information to understand what AI decisions were made, and why.	**Review** AI audit trails. **Determine** whether audit trails provide sufficient information to understand what decisions were made, and why.
Provide assurance that policies and procedures have been implemented and are working as designed, and that employees are compliant.	**Observe** employees implementing AI procedures. **Review** helpline/hotline reports and **follow up** on any reports alleging noncompliant or malicious activities related to AI. **Interview** a random sample of employees and **determine** if they are knowledgeable about AI policies and procedures. **Identify** and **review** AI access policies and procedures. **Evaluate** access policies and **test** access controls. **Assess** whether regulatory control objectives reflect emerging regulations, standards, and guidance.

"Data Infrastructure & Architecture and Data Quality are often intertwined. Relevant engagement or control objectives, and activities and procedures in one area, may overlap or impact objectives, activities, and procedures in the other area."

Lesedi Lesetedi,

Deputy Executive Director (Deputy CEO) – Strategy & Corporate Services

Botswana College of Distance & Open Learning (BOCODOL)

Data Architecture & Infrastructure

AI data architecture and infrastructure will likely be one and the same, or at least nearly the same, as the organization's architecture and infrastructure for handling big data. It includes considerations for:

- The way that data is accessible (metadata, taxonomy, unique identifiers, and naming conventions).
- Information privacy and security throughout the data lifecycle (data collection, use, storage, and destruction).
- Roles and responsibilities for data ownership and use throughout the data life cycle.

According to InfoWorld, organizations should focus on three major areas of software development to ensure the success of AI integration:

- Data integration — data from multiple sources must be integrated before AI can be incorporated into the organization's applications and systems.
- Application modernization — software updates will need to be made on a regular basis. Frequent, less intensive updates should replace infrequent, more intensive updates that slow down or disrupt systems.
- Employee education — software developers, project managers, and other technology staff need to keep up with machine learning and every aspect of the technology "stack" (the software and components that run AI).

In addition, data should be reconciled so that nuances such as rounding, demographics, and other variables are normalized before input.

Relevant Data Architecture & Infrastructure Objectives and Activities or Procedures

Engagement or Control Objective(s)	Activities or Procedures
Provide assurance that the organization is cyber resilient. Cyber resilience includes, but is broader than, cybersecurity alone. Cyber resilience encompasses security (resistance), reaction, and recovery.	**Understand** and audit big data (see The IIA's Practice Guide: Understanding and Auditing Big Data). **Assess** whether the organization is preparing for compliance with new technology regulations, such as the EU's General Data Protection Regulation (GDPR). **Assess** whether the organization's disaster recovery protocols include AI failures, including the breakdown of controls that maintain the rules set forth by AI governance.
Provide assurance that the data infrastructure has the capacity to accommodate the size and complexity of AI activity set forth in the AI strategy.	**Assess** whether the infrastructure is capable of handling structured and unstructured data.
Provide assurance that the organization has established a data taxonomy. **Evaluate** the quality, completeness, and consistency of use for the enterprisewide data taxonomy.	**Assess** whether the taxonomy is robust enough to accommodate the size and complexity of AI activities.

Data Quality

The completeness, accuracy, and reliability of the data on which AI algorithms are built are critical. For AI to be successful, organizations need access to vast amounts of high quality data — data that is well-defined and in standardized formats. Often, systems do not communicate with each other or do so through complicated add-ons or customizations. How this data is reconciled, synthesized, and validated is also critical, so systems that do not communicate with each other or do so through complicated add-ons or customizations may thwart an organization's AI activities.

In addition to data that is well-defined in standardized formats (structured data), AI technologies may be dependent on unstructured data (such as social media posts). As described in The IIA's "Global Technology Audit Guide: Understanding and Auditing Big Data," unstructured data is "typically more difficult to manage, due to its evolving and unpredictable nature, and it is usually sourced from large, disparate, and often external data sources. Consequently, new solutions have been developed to manage and analyze this data."

Ironically, organizations can turn to machine learning — a form of AI — to improve data quality. For example, there may be multiple versions of a vendor's name across an organization's many business units, data bases, and spreadsheets. A computer program could scan and reconcile all variations of the name in a matter of hours or minutes.

Internal audit also should look at how data that is used in internal audit reports has been reconciled, synthesized, and validated.

Relevant Data Quality Objectives and Activities or Procedures

Engagement or Control Objective(s)	Activities or Procedures
Provide assurance over the reliability of AI's underlying algorithms and the data on which algorithms are based.	**Obtain** a sample of the raw data that are inputs to AI. **Verify** that the organization has implemented methodologies to validate AI outcomes with actual, real-world outcomes, and that policies and procedures are in place to continuously measure, monitor, escalate, and rectify inconsistencies between the two.
Provide assurance that data input is reconciled and normalized to maximize accuracy.	**Verify** that the organization has policies and procedures in place to continuously measure, monitor, escalate, and rectify data accuracy and integrity issues. **Confirm** that the organization is consistently following and monitoring a formalized data reconciliation framework, which includes a rationale for differing methodologies and results should they exist.
Provide assurance that aggregated data is complete.	**Verify** that the organization has policies and procedures in place to limit data input bias.
Provide assurance that the completeness of data is measured and monitored and that any material exceptions that impact decision-making are identified and explained. This should be done whether the exceptions are determined by humans or AI.	**Review** AI metrics and metric reports. **Assess** whether those responsible for decision-making have received and considered explanations on material exceptions related to data quality.

Facebook's Corrective Actions

Facebook's challenges with AI have been widely reported. The behemoth social network has been under scrutiny over how its algorithm-fueled technologies have been used — or misused — for malicious means.

Timeline of Major Concerns:

- In Fall 2016, ProPublica reported that advertisers could use Facebook's ad-targeting tools to exclude certain races — a potential violation of federal housing and civil rights regulations.
- In September 2017:
    - Facebook disclosed that holders of fake accounts based in Russia purchased sizeable ads on divisive issues leading up to the 2016 presidential election.
    - ProPublica reported that Facebook's ad targeting tools enabled advertisers to target self-described ethnic "haters."
- In October 2017, concerns about fake news resurfaced when Facebook (and Google) posted false information about the mass shooting in Las Vegas.
- In testimony to a Senate judiciary subcommittee in late October, Facebook said the reach of Russia-backed ads stretched much further than they had originally known, reaching as many as 126 million Americans before and during the 2016 presidential election.

Facebook's Response

In a Sept. 20, 2017, post, Facebook Chief Operating Officer Sheryl Sandberg announced three corrective actions:

1. Facebook is clarifying its advertising policies and tightening its enforcement processes to ensure that content that goes against Facebook's community standards cannot be used to target ads. (Policies and processes relate to the Governance component of the Framework. Among other things, AI Governance should establish accountability for and oversight of enforcement.)

2. Facebook is increasing "human review and oversight" of its automated processes. (Human review and oversight relates to the Ethics component of the Framework. Among other things, Ethics addresses whether AI results reflect the original objective and whether AI output is being used legally, ethically, and responsibly.)

3. Facebook is working on a program that will encourage Facebook users to report potential abuses of its ads systems. (Reporting systems relate to the Measuring Performance component of the Framework. Reporting systems help management monitor the performance of AI activities. Measuring Performance will be covered in a future Global Perspectives and Insights report.)

By utilizing The IIA's AI Auditing Framework, internal auditors can provide assurance and advisory services to help organizations separate truth from fiction and address reporting, operations, and compliance risks associated with AI.

Using the Standards to Audit AI

Internal auditors should conform with all applicable IIA standards when planning or performing AI engagements. Key IIA standards that are particularly relevant to AI are highlighted in the sidebar, but others may apply as well.

Each standard is complemented by an Implementation Guide. Implementation guides assist internal auditors in applying the *Standards*. They collectively address internal auditing's approach, methodologies, and consideration, but do not detail processes or procedures.

Closing Thoughts

The IIA's Artificial Intelligence Auditing Framework will help internal auditors approach AI advisory and assurance services in a systematic and disciplined manner. Whether the organization's AI technologies and activities are developed in-house, through a facilitative technology such as AutoML, or by a third party, internal audit should be prepared to advise the board and senior management, coordinate with the first and second lines of defense, and provide assurance over AI risk management, governance, and controls.

This paper is Part II of a three-part series. It provides suggestions for implementing the AI Strategy and Governance components of The IIA's AI Auditing Framework. Part III will provide further suggestions for implementing the Governance component, and the Human Factor component.

Audit Focus

Key IIA Standards

The IIA's *International Standards for the Professional Practice of Internal Auditing* includes several standards that are particularly relevant to AI, including:

IIA Standard 1210: Proficiency

IIA Standard 2010: Planning

IIA Standard 2030: Resource Management

IIA Standard 2100: Nature of Work

IIA Standard 2110: Governance

IIA Standard 2130: Control

IIA Standard 2200: Engagement Planning

IIA Standard 2201: Planning Considerations

IIA Standard 2210: Engagement Objectives

IIA Standard 2220: Engagement Scope

IIA Standard 2230: Engagement Resource Allocation

IIA Standard 2240: Engagement Work Program

IIA Standard 2310: Identifying Information

The Institute of Internal Auditors | Global

globaliia.org

DOOR VIJAY GANGADIN
EN JOERI VAN ALPHEN

IORP II: Versterking van risicomanagement

Risk Governance en Risk Profile zijn twee kernbegrippen in IORP II

De invoering van de aanpassing van de Institutions for Occupational Retirement Provision Directive, IORP II, komt snel dichterbij. Het ministerie van Sociale Zaken en Werkgelegenheid (SZW) heeft in april een wetsvoorstel gepubliceerd en De Nederlandsche Bank (DNB) heeft aangekondigd dat zij dit jaar de voorbereidingen bij pensioenfondsen gaat monitoren. Ons inziens is de impact van de nieuwe eisen op het gebied van Risk Governance en Risk Profile niet te onderschatten. In deze bijdrage bespreken wij de aanpassing en geven wij handvatten hoe pensioenfondsen zich kunnen voorbereiden op deze nieuwe wetgeving.

De nieuwe IORP II regelgeving uit Europa moet uiterlijk 13 januari 2019 zijn geïmplementeerd in de Nederlandse wet- en regelgeving en dan moeten pensioenfondsen er aan voldoen. De regelgeving stelt dat er onafhankelijke functies moeten worden ingericht voor risico-management, interne audit en actuariaat. Dit zijn de zogeheten 'sleutelfuncties'. Veel van de werkzaamheden van deze functies vinden nu al plaats, maar de nodige taken zullen nader moeten worden geformaliseerd en aangepast in reglementen, processen, beleid en functionele invulling. Een nieuwe verplichting is dat er elke drie jaar een eigen-risicobeoordeling (ERB) dient te worden uitgevoerd en ingediend bij DNB. De ERB is een belangrijk instrument bij de strategische besluitvorming en het bepalen van het risicoprofiel van het fonds. Ten slotte zijn er aangescherpte eisen op het gebied van ESG-beleid en beloningsbeleid.

WIJ SPREKEN BIJ VOORKEUR OVER DE 'SIX LINES OF DEFENCE'

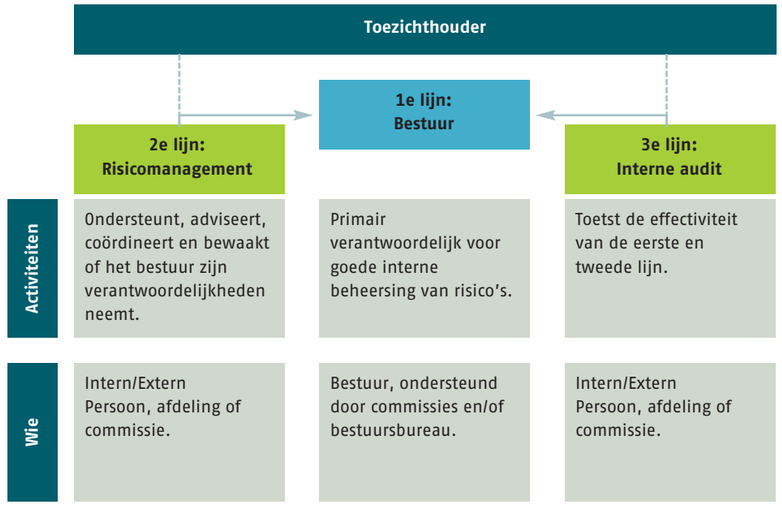
RISK GOVERNANCE

Naar onze mening kan de introductie van onafhankelijke sleutel-functies gezien worden als de introductie van het zogenaamde 'three lines of defence' model. Het model wordt weliswaar niet expliciet genoemd door de wetgever, maar het is een manier om de beoogde functiescheiding te realiseren, de onafhankelijkheid te waarborgen en de bestuurlijke besluitvorming verder te professionaliseren. Bij gereguleerde financiële instellingen is de risicomanagementfunctie de tweede lijn en de interne audit functie de derde lijn. De eerste lijn is 'de business' met aan het hoofd het management. Het management is verantwoordelijk voor de operationele beheersing van risico's en de besluitvorming. Bij pensioenfondsen vervult het bestuur de taak van het management.

Als wij de hele keten van assurance rollen zouden meenemen, spreken wij bij voorkeur over de 'six lines of defence'. De vierde, vijfde en zesde lijn vallen buiten de verantwoordelijkheid van het bestuur, maar vervullen wel een rol bij de bescherming van de deelnemers en pensioengerechtigden. De 'six lines of defence' zijn als volgt:

Lines of Defence	
Eerste lijn	Bestuur
Tweede lijn	Risicomanagement & compliance- en actuariële functie
Derde lijn	Interne audit functie
Vierde lijn	Raad van Toezicht
Vijfde lijn	Externe accountant en certificerend actuaris
Zesde lijn	Externe toezichthouders (DNB/AFM/AP)

In overeenstemming met de IORP II regelgeving, focussen wij op de eerste drie lijnen; in de vierde, vijfde en zesde lijn leidt de nieuwe regelgeving ook niet tot wijzigingen. In onze visie zouden de 'three lines of defence' als volgt ingericht kunnen worden:



Eerste lijn – Bestuur
Het bestuur vervult de functie van de eerste lijn. Hier worden de besluiten genomen die het risicoprofiel van het fonds bepalen. Ook in de commissies vindt veelal eerstelijns risicomanagement plaats. Bijvoorbeeld, in de beleggingscommissie wordt namens het bestuur gemonitord dat de risico's binnen de afgesproken grenzen blijven, of dat bij een belegging in een nieuwe categorie is stilgestaan bij alle risico's.

Tweede lijn – Risicomanagementfunctie
Op dit moment is de invulling van een onafhankelijke risico-managementfunctie alleen voor beleggingsrisico's verplicht (artikel 18 FTK). Onder IORP II wordt de scope van het risicomanagement breder getrokken, namelijk naar alle relevante risico's voor het pensioenfonds, dus inclusief operationele en niet-financiële risico's. De functie zal moeten zorgen voor een onafhankelijk oordeel bij besluitvorming van de eerste lijn (het bestuur), moeten kunnen escaleren naar DNB en andere toezichthouders indien het bestuur risico's veronachtzaamt, alsmede de eerste lijn '*challengen*' bij de identificatie en mitigering van risico's.

Wij zien verschillende mogelijkheden voor een praktische invulling van de risicomanagementfunctie. Twee voorbeelden:

1. Het bestuurslid met risicomanagement in zijn 'portefeuille' is functiehouder van de risicomanagementfunctie en de invulling van de functie wordt uitbesteed;

2. De functiehouder is een onafhankelijke risicomanager in dienst van het pensioenfonds. De invulling van de functie wordt uitbesteed, of uitgevoerd door medewerkers van het pensioenfonds als deze beschikbaar zijn en voldoende kennis hebben. De regelgeving lijkt niet uit te sluiten dat de functiehouder tevens de functie kan vervullen.

Proportionaliteit
De sleutelfuncties mogen proportioneel worden ingevuld en mogen niet tot teveel extra lasten leiden voor het pensioenfonds. 'Proportioneel' wil zeggen dat de governance stringenter dient te worden doorgevoerd (lees: striktere onafhankelijkheden) naarmate het pensioenfonds groter en/of complexer is. Het is op dit moment nog niet duidelijk hoe DNB daar in de praktijk mee zal omgaan.

Operationeel risicomanagement
Het 'beheer van het operationele risico' is formeel nieuw in de scope van de risicomanagementfunctie. Dit is een breed gebied; denk hierbij aan integriteitsrisico's, IT- & cyberrisico's en omgevingsrisico's. Deze uitbreiding van de scope van de risicomanagement functie zal daarom de nodige (extra) kennis vereisen van degene die de risico-managementfunctie vervult. Op dit moment hanteren veel pensioen-fondsen de FIRM-methodiek voor de identificatie en evaluatie van risico's. De operationele risico's zullen daarom veelal al 'in beeld' zijn. Echter, het beheersen van het risico en het *challengen* van het bestuur op dit gebied zal voor veel risicomanagers nieuw zijn. Als het bestuur bijvoorbeeld voornemens is de pensioenadministratie te gaan uit-besteden, zal de rol van de risicomanagementfunctie bestaan uit het goed reviewen of aan alle hiermee gepaard gaande operationele risico's is gedacht en het bewaken dat deze adequaat worden beheerst. Er is daartoe onder meer een goed inzicht nodig in de IT-risico's van de uitbesteding en van de processen rondom de uitbesteding.

Dr. V. Gangadin MBA CRO (links) is Partner en Boardroom Advisor bij het strategisch adviesbureau Sprenkels & Verschuren. Hij is thought leader op het snijvlak van strategische en risicomanagement vraagstukken en gepromoveerd op het onderwerp Risk Management – Risk Appetite. Hij is associate professor aan diverse (internationale) Business Scholen en motivational speaker op (inter)nationale congressen.

Drs. J. van Alphen FRM RBA MBA, is Senior Consultant bij het strategisch adviesbureau Sprenkels & Verschuren. Hij is expert op de IORP II en Solvency II wetgeving en heeft diverse advies- en managementfuncties vervuld bij onder meer DNB en Delta Lloyd.

Tweede lijn – Actuariële functie
Ook de actuariële functie is een tweedelijns sleutelfunctie in IORP II. De rol van de actuariële functie zoals beschreven in de Europese Richtlijn is zowel coördinerend, beoordelend als adviserend. Deze is niet gelijk aan de huidige rol van de externe certificerend actuaris; zijn rol is immers alleen controlerend/beoordelend. Ook op de inhoud zijn er verschillen tussen de actuariële functie en de certificerend actuaris: de actuariële functie gaat over de verplichtingen, terwijl de certificerend actuaris bijvoorbeeld ook zijn oordeel moet geven over het vereist eigen vermogen en het prudent person beginsel. Toch is de Nederlandse wetgever, om zoveel mogelijk aan te sluiten bij de praktijk, voornemens de mogelijkheid te bieden dat de certificerend actuaris de actuariële functie mag vervullen. Hier is echter nog discussie over. Vanwege het controlerende karakter van de functie van de certificerend actuaris, wijzigt de huidige rol van de adviserend actuaris niet.

Derde lijn: Interne audit functie
De taak van de interne audit functie is het evalueren van de effectiviteit van de eerste en de tweede lijn. De term 'audit' refereert hierbij dus niet aan werkzaamheden rond het jaarwerkproces. Voor de audit functie verwachten wij de grootste impact omdat deze veelal niet is ingericht, althans niet volledig onafhankelijk van het bestuur en het risicomanagement. Een eventueel aanwezige auditcommissie (die in principe verplicht is voor pensioenfondsen met een omgekeerd gemengd bestuursmodel) zal deze functie wellicht niet kunnen vervullen vanwege de personele overlap met het bestuur. Een werkbare invulling kan wellicht zijn dat de auditcommissie maar één bestuurslid telt, en opdracht geeft aan voldoende onafhankelijke partijen (bijvoorbeeld een externe accountant, auditor of audit afdeling van de sponsor in geval van een ondernemingspensioenfonds) om audits uit te voeren.

RISICOPROFIEL – RISK PROFILE

Volledig nieuw voor Nederlandse pensioenfondsen wordt de formele eigen-risicobeoordeling (ERB). Deze dient minimaal eens in de drie jaar te worden uitgevoerd, of zoveel eerder als een significante wijziging van het risicoprofiel of pensioenregeling dat vereist. De ERB dient aan DNB te worden verstrekt. Tevens dient de ERB betrokken te worden bij strategische besluitvorming door het bestuur. Met andere woorden, het bestuur moet zich rekenschap geven van de risico's bij strategische besluiten. De risicomanagementfunctie heeft hierbij een essentiële taak. Het impliceert ook dat de strategische risico's (bijvoorbeeld het risico dat het fonds niet zelfstandig kan voortbestaan, of het risico dat het op termijn niet kan indexeren) door het pensioenfonds zijn onderkend, beschreven en worden geëvalueerd in de ERB. De ERB noemt de 'strategische risico's' niet expliciet. Echter, voor een volledig risicoprofiel van het fonds is een goede identificatie van deze risico's onontbeerlijk. Pas als de ERB het volledige risicoprofiel omvat, kan de ERB bij strategische besluitvorming goed tot zijn recht komen.

De ERB is zowel kwalitatief als kwantitatief van aard. Hoewel het nog niet bekend is of er richtlijnen worden opgesteld voor de wijze waarop de kwantitatieve evaluaties moeten worden uitgevoerd, naar verwachting zal het niet sterk gaan afwijken van ALM-studies en haalbaarheidstoetsen. De ERB noemt niet expliciet een evaluatie van de solvabiliteit, maar vanwege de evaluatie van de financieringsbehoefte en de impact van maatregelen als indexatie en kortingen, wordt in wezen een evaluatie van het balansmanagement en de financiële opzet gevraagd. In die zin lijkt de ERB op de zogenaamde Own Risk & Solvency Assessment (ORSA) uit Solvency II voor verzekeraars en de Internal Capital Adequacy Assessment Process (ICAAP) uit Basel III voor banken. Met dien verstande dat de ORSA jaarlijks uitgevoerd wordt of, net als de ERB, bij grote veranderingen van het risicoprofiel.

Scenariodenken
Hoewel niet verplicht in het kader van de ERB, kunnen scenario-analyses ons inziens nuttig zijn bij het nadenken over en in kaart brengen van het risicoprofiel. Denk hierbij bijvoorbeeld aan:

- Een stijging van de levensverwachting, met gevolgen voor de voorzieningen en dus solvabiliteit;
- Een operationeel risico: cyberrisico, waarbij gegevens van deelnemers zijn ontvreemd. Wat is de impact op de organisatie? Welke controls hebben niet gewerkt/zijn er niet? Is er een financiële- en of reputationele impact?
- Reverse stresstest: welke rentedaling moet er optreden om terug te vallen naar een dekkingsgraad van 105%?

Veel pensioenfondsen beschikken over projectiemodellen (ALM, haalbaarheidstoetsen), zodat het doorrekenen van scenario's relatief goed werkbaar is. De grootste toegevoegde waarde van een ERB is naar onze mening een verbeterd inzicht in het risicoprofiel van het fonds en hoe het pensioenfonds is voorbereid op de risico's. Met betrekking tot operationele risico's kan de ERB ertoe leiden dat er additionele, risicomitigerende maatregelen moeten worden genomen.

TOT SLOT

Voor pensioenfondsen is het nu zaak om op de bestuurlijke agenda ruimte te maken voor IORP II en geen besluiten te nemen die tegenstrijdig zijn met IORP II. Met name de concrete invulling en bemensing van de sleutelfuncties (Risk Governance) en de exacte invulling van het Risk Profile, kan nog de nodige voorbereidingstijd vergen. ■

DOOR VIJAY GANGADIN
EN JOERI VAN ALPHEN

IORP II: Versterking van risicomanagement

Risk Governance en Risk Profile zijn twee kernbegrippen in IORP II

De invoering van de aanpassing van de Institutions for Occupational Retirement Provision Directive, IORP II, komt snel dichterbij. Het ministerie van Sociale Zaken en Werkgelegenheid (SZW) heeft in april een wetsvoorstel gepubliceerd en De Nederlandsche Bank (DNB) heeft aangekondigd dat zij dit jaar de voorbereidingen bij pensioenfondsen gaat monitoren. Ons inziens is de impact van de nieuwe eisen op het gebied van Risk Governance en Risk Profile niet te onderschatten. In deze bijdrage bespreken wij de aanpassing en geven wij handvatten hoe pensioenfondsen zich kunnen voorbereiden op deze nieuwe wetgeving.

De nieuwe IORP II regelgeving uit Europa moet uiterlijk 13 januari 2019 zijn geïmplementeerd in de Nederlandse wet- en regelgeving en dan moeten pensioenfondsen er aan voldoen. De regelgeving stelt dat er onafhankelijke functies moeten worden ingericht voor risico-management, interne audit en actuariaat. Dit zijn de zogeheten 'sleutelfuncties'. Veel van de werkzaamheden van deze functies vinden nu al plaats, maar de nodige taken zullen nader moeten worden geformaliseerd en aangepast in reglementen, processen, beleid en functionele invulling. Een nieuwe verplichting is dat er elke drie jaar een eigen-risicobeoordeling (ERB) dient te worden uitgevoerd en ingediend bij DNB. De ERB is een belangrijk instrument bij de strategische besluitvorming en het bepalen van het risicoprofiel van het fonds. Ten slotte zijn er aangescherpte eisen op het gebied van ESG-beleid en beloningsbeleid.

WIJ SPREKEN BIJ VOORKEUR OVER DE 'SIX LINES OF DEFENCE'

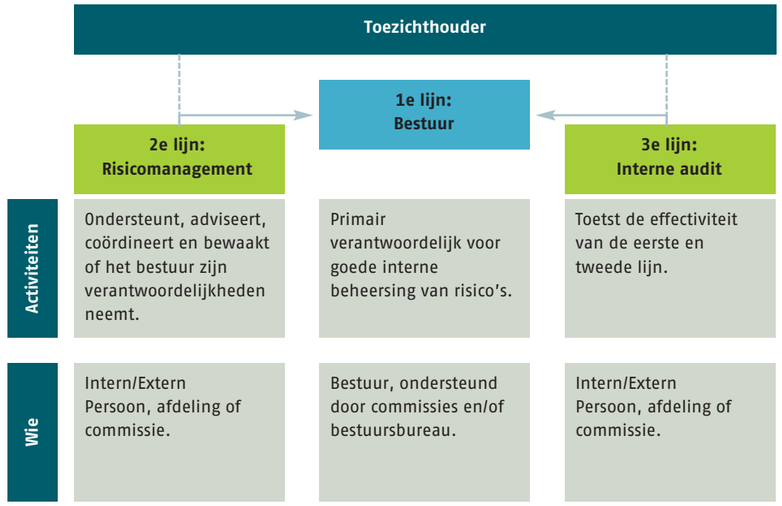
RISK GOVERNANCE

Naar onze mening kan de introductie van onafhankelijke sleutel-functies gezien worden als de introductie van het zogenaamde 'three lines of defence' model. Het model wordt weliswaar niet expliciet genoemd door de wetgever, maar het is een manier om de beoogde functiescheiding te realiseren, de onafhankelijkheid te waarborgen en de bestuurlijke besluitvorming verder te professionaliseren. Bij gereguleerde financiële instellingen is de risicomanagementfunctie de tweede lijn en de interne audit functie de derde lijn. De eerste lijn is 'de business' met aan het hoofd het management. Het management is verantwoordelijk voor de operationele beheersing van risico's en de besluitvorming. Bij pensioenfondsen vervult het bestuur de taak van het management.

Als wij de hele keten van assurance rollen zouden meenemen, spreken wij bij voorkeur over de 'six lines of defence'. De vierde, vijfde en zesde lijn vallen buiten de verantwoordelijkheid van het bestuur, maar vervullen wel een rol bij de bescherming van de deelnemers en pensioengerechtigden. De 'six lines of defence' zijn als volgt:

Lines of Defence	
Eerste lijn	Bestuur
Tweede lijn	Risicomanagement & compliance- en actuariële functie
Derde lijn	Interne audit functie
Vierde lijn	Raad van Toezicht
Vijfde lijn	Externe accountant en certificerend actuaris
Zesde lijn	Externe toezichthouders (DNB/AFM/AP)

In overeenstemming met de IORP II regelgeving, focussen wij op de eerste drie lijnen; in de vierde, vijfde en zesde lijn leidt de nieuwe regelgeving ook niet tot wijzigingen. In onze visie zouden de 'three lines of defence' als volgt ingericht kunnen worden:



Eerste lijn – Bestuur
Het bestuur vervult de functie van de eerste lijn. Hier worden de besluiten genomen die het risicoprofiel van het fonds bepalen. Ook in de commissies vindt veelal eerstelijns risicomanagement plaats. Bijvoorbeeld, in de beleggingscommissie wordt namens het bestuur gemonitord dat de risico's binnen de afgesproken grenzen blijven, of dat bij een belegging in een nieuwe categorie is stilgestaan bij alle risico's.

Tweede lijn – Risicomanagementfunctie
Op dit moment is de invulling van een onafhankelijke risico-managementfunctie alleen voor beleggingsrisico's verplicht (artikel 18 FTK). Onder IORP II wordt de scope van het risicomanagement breder getrokken, namelijk naar alle relevante risico's voor het pensioenfonds, dus inclusief operationele en niet-financiële risico's. De functie zal moeten zorgen voor een onafhankelijk oordeel bij besluitvorming van de eerste lijn (het bestuur), moeten kunnen escaleren naar DNB en andere toezichthouders indien het bestuur risico's veronachtzaamt, alsmede de eerste lijn '*challengen*' bij de identificatie en mitigering van risico's.

Wij zien verschillende mogelijkheden voor een praktische invulling van de risicomanagementfunctie. Twee voorbeelden:

1. Het bestuurslid met risicomanagement in zijn 'portefeuille' is functiehouder van de risicomanagementfunctie en de invulling van de functie wordt uitbesteed;

2. De functiehouder is een onafhankelijke risicomanager in dienst van het pensioenfonds. De invulling van de functie wordt uitbesteed, of uitgevoerd door medewerkers van het pensioenfonds als deze beschikbaar zijn en voldoende kennis hebben. De regelgeving lijkt niet uit te sluiten dat de functiehouder tevens de functie kan vervullen.

Proportionaliteit
De sleutelfuncties mogen proportioneel worden ingevuld en mogen niet tot teveel extra lasten leiden voor het pensioenfonds. 'Proportioneel' wil zeggen dat de governance stringenter dient te worden doorgevoerd (lees: striktere onafhankelijkheden) naarmate het pensioenfonds groter en/of complexer is. Het is op dit moment nog niet duidelijk hoe DNB daar in de praktijk mee zal omgaan.

Operationeel risicomanagement
Het 'beheer van het operationele risico' is formeel nieuw in de scope van de risicomanagementfunctie. Dit is een breed gebied; denk hierbij aan integriteitsrisico's, IT- & cyberrisico's en omgevingsrisico's. Deze uitbreiding van de scope van de risicomanagement functie zal daarom de nodige (extra) kennis vereisen van degene die de risico-managementfunctie vervult. Op dit moment hanteren veel pensioen-fondsen de FIRM-methodiek voor de identificatie en evaluatie van risico's. De operationele risico's zullen daarom veelal al 'in beeld' zijn. Echter, het beheersen van het risico en het *challengen* van het bestuur op dit gebied zal voor veel risicomanagers nieuw zijn. Als het bestuur bijvoorbeeld voornemens is de pensioenadministratie te gaan uit-besteden, zal de rol van de risicomanagementfunctie bestaan uit het goed reviewen of aan alle hiermee gepaard gaande operationele risico's is gedacht en het bewaken dat deze adequaat worden beheerst. Er is daartoe onder meer een goed inzicht nodig in de IT-risico's van de uitbesteding en van de processen rondom de uitbesteding.

Dr. V. Gangadin MBA CRO (links) is Partner en Boardroom Advisor bij het strategisch adviesbureau Sprenkels & Verschuren. Hij is thought leader op het snijvlak van strategische en risicomanagement vraagstukken en gepromoveerd op het onderwerp Risk Management – Risk Appetite. Hij is associate professor aan diverse (internationale) Business Scholen en motivational speaker op (inter)nationale congressen.

Drs. J. van Alphen FRM RBA MBA, is Senior Consultant bij het strategisch adviesbureau Sprenkels & Verschuren. Hij is expert op de IORP II en Solvency II wetgeving en heeft diverse advies- en managementfuncties vervuld bij onder meer DNB en Delta Lloyd.

Tweede lijn – Actuariële functie
Ook de actuariële functie is een tweedelijns sleutelfunctie in IORP II. De rol van de actuariële functie zoals beschreven in de Europese Richtlijn is zowel coördinerend, beoordelend als adviserend. Deze is niet gelijk aan de huidige rol van de externe certificerend actuaris; zijn rol is immers alleen controlerend/beoordelend. Ook op de inhoud zijn er verschillen tussen de actuariële functie en de certificerend actuaris: de actuariële functie gaat over de verplichtingen, terwijl de certificerend actuaris bijvoorbeeld ook zijn oordeel moet geven over het vereist eigen vermogen en het prudent person beginsel. Toch is de Nederlandse wetgever, om zoveel mogelijk aan te sluiten bij de praktijk, voornemens de mogelijkheid te bieden dat de certificerend actuaris de actuariële functie mag vervullen. Hier is echter nog discussie over. Vanwege het controlerende karakter van de functie van de certificerend actuaris, wijzigt de huidige rol van de adviserend actuaris niet.

Derde lijn: Interne audit functie
De taak van de interne audit functie is het evalueren van de effectiviteit van de eerste en de tweede lijn. De term 'audit' refereert hierbij dus niet aan werkzaamheden rond het jaarwerkproces. Voor de audit functie verwachten wij de grootste impact omdat deze veelal niet is ingericht, althans niet volledig onafhankelijk van het bestuur en het risicomanagement. Een eventueel aanwezige auditcommissie (die in principe verplicht is voor pensioenfondsen met een omgekeerd gemengd bestuursmodel) zal deze functie wellicht niet kunnen vervullen vanwege de personele overlap met het bestuur. Een werkbare invulling kan wellicht zijn dat de auditcommissie maar één bestuurslid telt, en opdracht geeft aan voldoende onafhankelijke partijen (bijvoorbeeld een externe accountant, auditor of audit afdeling van de sponsor in geval van een ondernemingspensioenfonds) om audits uit te voeren.

RISICOPROFIEL – RISK PROFILE

Volledig nieuw voor Nederlandse pensioenfondsen wordt de formele eigen-risicobeoordeling (ERB). Deze dient minimaal eens in de drie jaar te worden uitgevoerd, of zoveel eerder als een significante wijziging van het risicoprofiel of pensioenregeling dat vereist. De ERB dient aan DNB te worden verstrekt. Tevens dient de ERB betrokken te worden bij strategische besluitvorming door het bestuur. Met andere woorden, het bestuur moet zich rekenschap geven van de risico's bij strategische besluiten. De risicomanagementfunctie heeft hierbij een essentiële taak. Het impliceert ook dat de strategische risico's (bijvoorbeeld het risico dat het fonds niet zelfstandig kan voortbestaan, of het risico dat het op termijn niet kan indexeren) door het pensioenfonds zijn onderkend, beschreven en worden geëvalueerd in de ERB. De ERB noemt de 'strategische risico's' niet expliciet. Echter, voor een volledig risicoprofiel van het fonds is een goede identificatie van deze risico's onontbeerlijk. Pas als de ERB het volledige risicoprofiel omvat, kan de ERB bij strategische besluitvorming goed tot zijn recht komen.

De ERB is zowel kwalitatief als kwantitatief van aard. Hoewel het nog niet bekend is of er richtlijnen worden opgesteld voor de wijze waarop de kwantitatieve evaluaties moeten worden uitgevoerd, naar verwachting zal het niet sterk gaan afwijken van ALM-studies en haalbaarheidstoetsen. De ERB noemt niet expliciet een evaluatie van de solvabiliteit, maar vanwege de evaluatie van de financieringsbehoefte en de impact van maatregelen als indexatie en kortingen, wordt in wezen een evaluatie van het balansmanagement en de financiële opzet gevraagd. In die zin lijkt de ERB op de zogenaamde Own Risk & Solvency Assessment (ORSA) uit Solvency II voor verzekeraars en de Internal Capital Adequacy Assessment Process (ICAAP) uit Basel III voor banken. Met dien verstande dat de ORSA jaarlijks uitgevoerd wordt of, net als de ERB, bij grote veranderingen van het risicoprofiel.

Scenariodenken
Hoewel niet verplicht in het kader van de ERB, kunnen scenario-analyses ons inziens nuttig zijn bij het nadenken over en in kaart brengen van het risicoprofiel. Denk hierbij bijvoorbeeld aan:

- Een stijging van de levensverwachting, met gevolgen voor de voorzieningen en dus solvabiliteit;
- Een operationeel risico: cyberrisico, waarbij gegevens van deelnemers zijn ontvreemd. Wat is de impact op de organisatie? Welke controls hebben niet gewerkt/zijn er niet? Is er een financiële- en of reputationele impact?
- Reverse stresstest: welke rentedaling moet er optreden om terug te vallen naar een dekkingsgraad van 105%?

Veel pensioenfondsen beschikken over projectiemodellen (ALM, haalbaarheidstoetsen), zodat het doorrekenen van scenario's relatief goed werkbaar is. De grootste toegevoegde waarde van een ERB is naar onze mening een verbeterd inzicht in het risicoprofiel van het fonds en hoe het pensioenfonds is voorbereid op de risico's. Met betrekking tot operationele risico's kan de ERB ertoe leiden dat er additionele, risicomitigerende maatregelen moeten worden genomen.

TOT SLOT

Voor pensioenfondsen is het nu zaak om op de bestuurlijke agenda ruimte te maken voor IORP II en geen besluiten te nemen die tegenstrijdig zijn met IORP II. Met name de concrete invulling en bemensing van de sleutelfuncties (Risk Governance) en de exacte invulling van het Risk Profile, kan nog de nodige voorbereidingstijd vergen. ■

Process Maturity Model

Process Capability Matrix*

*Adapted from The Processes Capabilities Matrix, published in "Auditor's Risk Management Guide," by Paul J. Sobel (2007).

Maturity Stage	Observable Process Characteristics			
	Procedures	Controls	Metrics	Improvement Mechanisms
5 Optimized	Processes and controls are continuously reviewed and improved.	Preventive and detective controls are highly automated to reduce human error and cost of operation.	Comprehensive, defined performance metrics exist, with extensive automated performance monitoring.	Extensive use of best practices, benchmarking, and/or self-assessment to continuously improve process.
4 Managed	Procedures and controls are well documented and kept current.	Preventive and detective controls are employed, with greater use of automation to reduce human error.	Many metrics are used with a blend of automated and manual performance monitoring.	Best practices and/or benchmarking are used to improve process.
3 Defined	Procedures are well documented, but not kept current to reflect changing business needs.	Preventive and detective controls are employed, still reliant on manual activities.	Some metrics are used, but performance monitoring is still manual and/or infrequent.	Generally occurs during periodic (e.g., annual) policy and procedure renewal.
2 Repeatable	Some standard procedures exist, relies on "tribal knowledge."	Mostly detective controls are in place, minimal preventive controls, and highly manual.	Few performance metrics exist, thus performance monitoring is inconsistent or informal.	Most likely in reaction to audits or service disruptions.
1 Ad Hoc	No formal procedures exist.	Controls are non-existent or primarily in reaction to a "surprise."	There are no metrics or performance monitoring.	None

Process Maturity Model

Process Capability Matrix*

*Adapted from The Processes Capabilities Matrix, published in "Auditor's Risk Management Guide," by Paul J. Sobel (2007).

Maturity Stage	Observable Process Characteristics			
	Procedures	Controls	Metrics	Improvement Mechanisms
5 Optimized	Processes and controls are continuously reviewed and improved.	Preventive and detective controls are highly automated to reduce human error and cost of operation.	Comprehensive, defined performance metrics exist, with extensive automated performance monitoring.	Extensive use of best practices, benchmarking, and/or self-assessment to continuously improve process.
4 Managed	Procedures and controls are well documented and kept current.	Preventive and detective controls are employed, with greater use of automation to reduce human error.	Many metrics are used with a blend of automated and manual performance monitoring.	Best practices and/or benchmarking are used to improve process.
3 Defined	Procedures are well documented, but not kept current to reflect changing business needs.	Preventive and detective controls are employed, still reliant on manual activities.	Some metrics are used, but performance monitoring is still manual and/or infrequent.	Generally occurs during periodic (e.g., annual) policy and procedure renewal.
2 Repeatable	Some standard procedures exist, relies on "tribal knowledge."	Mostly detective controls are in place, minimal preventive controls, and highly manual.	Few performance metrics exist, thus performance monitoring is inconsistent or informal.	Most likely in reaction to audits or service disruptions.
1 Ad Hoc	No formal procedures exist.	Controls are non-existent or primarily in reaction to a "surprise."	There are no metrics or performance monitoring.	None

by Paul de Beus and Maarten Koning

AUDITING THE FUTURE

Drs. P.W.P. de Beus (above) is partner at EY Actuaries. Drs. M. Koning RA is partner at EY Accountants. This article is written on personal title.

In June 2013 the European Insurance and Occupational Pensions Authority (EIOPA) issued its second 'unofficial' pre-consultation paper concerning 'Proposals on Guidelines on External Audit'. The proposed guidelines require the statutory auditor to perform an external audit on quantitative and qualitative elements of the Solvency Financial Condition Report. In particular: the Solvency II balance sheet, own funds, and minimum and solvency capital requirements. The opinion arising from the Solvency II external audit is expected to be similar in terms of the degree of assurance to the opinion provided by auditors of financial statements. I.e. reasonable assurance. The guidelines represent minimum requirements. For instance, the guidelines do not include materiality considerations. So, when is good, 'good enough' in the situation where so much information is based on modeled cash flow projections and balance sheets into the future?

Solvency II Audit

In the audit of a Solvency II economic balance sheet and risk based capital requirements, the auditor should evaluate the outcome of complex models and complicated processes. The model outputs often are the result of various model choices, (risk) data vendor selection, and many assumptions. Using methods and assumptions that are equally valid could lead to significantly different results. This emphasizes that an outcome, or in audit terminology 'the accounting estimate', is highly sensitive to assumptions used and therefore subject to high *estimation uncertainty*. As a consequence, the range of possible reasonable outcomes can be quite broad. This range could be much broader than the materiality threshold calculated for auditing the Solvency II balance sheet, taken as a whole. Therefore, when (actuarial) auditors evaluate their findings, two questions should be answered:

- Is it acceptable to apply a range of reasonable outcomes which is higher than the traditional audit materiality; and
- Can we develop a method to define reasonable ranges?

With this article we aim to instigate an industry discussion on this topic, and to advocate for a closer working relationship between auditors and actuaries.

Background on materiality

Consideration should be given to what is written in the current International Standards on Auditing (ISAs). ISA 320[1] provides general guidance. Misstatements are considered to be material if they, individually or in the aggregate, could reasonably be expected to influence the economic decisions of users taken on the basis of the financial statements. In addition, ISA 320 indicates that a percentage is often applied to a chosen benchmark in determining materiality.

In applying this concept to a Solvency II balance sheet, we expect to see materiality set at a percentage of Own Funds. Based on this threshold, a 'Tolerable Error' (TE) is defined. The TE is the amount or amounts set by the auditor at less than materiality for the financial statements as a whole. This is to reduce to an appropriate low level the probability that the aggregate of uncorrected and undetected misstatements exceeds materiality.

ISA 540[2] recognizes that developing 'a range' to evaluate management's estimates may be an appropriate audit response. This range is required to encompass all 'reasonable outcomes'. The standard continues in saying that, particularly in certain industries, it may not be possible to narrow the range to below the TE amount, and that this does not necessarily preclude recognition of the accounting estimate. It may indicate, however, that the estimation uncertainty associated with the accounting estimate is such that it gives rise to a significant risk. This means that the auditor's further substantive procedures are focused on the evaluation of how management has assessed the effect of estimation uncertainty and the adequacy of the related disclosures.

This is further elaborated on in ISA 545[3]. Management may evaluate alternative assumptions or outcomes of the accounting estimates through a number of methods, for instance undertaking a sensitivity analysis.

A sensitivity analyses may demonstrate that the accounting estimate is sensitive to one or more assumptions that then become the focus of the auditor's attention.

Based on the guidance obtained from the ISAs above, we conclude that it is feasible to evaluate the (audit) findings from model outputs against a reasonable range. And in specific situations, this could exceed the traditional TE.

1 – ISA 320 'Materiality in planning and performing an audit'

2 – ISA 540 'Auditing accounting estimates, including fair value estimates, and related disclosures'

3 – ISA 545 'Auditing fair value measurements and disclosures'

The challenge we face now is: how to translate the qualitative assessment of a reasonable range and make it quantifiable and objective? We believe we can do so by defining what we call a 'Tolerable Range' (TR). We believe we need to make this distinction between TE and TR because audit materiality and TE measures are mainly focusing on 'errors'. However, when using model outputs in which key assumptions have had to be made, and cash flows have been projected far away into the future, it is quite often not simply 'right or wrong'.

Assessing models

When assessing a valuation or risk model, the auditor should evaluate the model's theoretical soundness (representativeness), mathematical integrity and appropriateness of the model parameters (ISA 540).

In assessing the *representativeness of input and methodology* the verification focuses on whether input used is 'fit for purpose' (e.g. interest curve used, basis spread correction) and whether the methodology is widely used and market practice. The arithmetical correctness can be confirmed by parallel modeling using same input as used for the assessment value or even parallel modeling using underlying contractual input and market information obtained independently. The *objective measurability of input parameters* can be concluded on by reconciliation of static input data with underlying contracts/systems and reconciliation of market input data with objective systems (data providers, such as Bloomberg, Reuters). By performing various specifically selected audit procedures, assessments of one of these criteria often implicitly contains a (partial) assessment of another criteria.

Assessment measure

In the (actuarial) assessment of the *arithmetical correctness and representativeness* of a valuation or risk model, an analysis has to be made on the difference between the model output under the parallel replicated model, alternative method(s), and the assessment value. We appreciate that parallel modeling is not always feasible in full, but model points and/or benchmark portfolios can provide sufficient insight for audit purposes. A conclusion can be drawn based on an objective measure that depends on the goal of the audit procedures performed. E.g. in audit procedures on derivatives valuation, we generally use the DV01 method. A DV01 is the sensitivity of the value of the financial derivative to a change in the underlying interest curve with 1 basis point. Mathematically DV01 estimates the first derivative to the interest rate of the valuation function.

Looking at a DV01 for interest rate derivatives makes sense: many derivatives are based on a fictive notional, and are structured on big (implicit or explicit) long and short positions. E.g. an Interest Rate Swap (IRS) has a value of nil at issue date because the (big) notional/value of the floating and fixed leg are equal. After the issue date the fair value of the floating and fixed leg change due to remaining maturity and changes in underlying market interest rates. The fair value of the IRS is only the net value of both legs. Therefore, this value is quite sensitive to the underlying interest rate curve. Hence, differences with other models and/or parameters can occur quite easily, *but are not necessarily an error*, and the auditor wants to know when good is good enough for audit considerations.

Differences can be caused by definitions of day-count conventions, inter- and extrapolation, choice of data vendor to extract interest rate curves et cetera. When auditing the value of interest rate derivatives[4], 3 times DV01 could be an acceptable range (TR) for setting as a materiality threshold. This '3' is chosen for European Interest Rate Swaps for the following reasons:

- Movements of 3 to 5 bps during a day is normal for EUR interest rate swap rates;
- Differences of 2 to 4 bps at a point in time is reasonable for EUR interest rate swap rate between different data vendors.

With this, we are able to set a materiality threshold objectively, and (or some would say 'but') dependent on the underlying assessment value. Indeed, in the new world, we believe that materiality is an important consideration in model risk management, and a proportionate risk-based approach is sensible in practice. This means the auditor should use different materiality measures *per model, per risk driver, per usage, per disclosure*.

Let's consider an IRS example by assessing the value of a €100 million notional receive 2.5% fixed/pay floating swap.
The assessment value equals €1,492,594.
We determined a value using a parallel model of €1,255,579, resulting in a difference of €237.015.
The auditor has a tolerable error threshold (TE) of €150,000. Remember, this is the 'traditional' threshold in terms of error for reporting purposes. Is the difference material?
In this example we probably would not conclude so.
The DV01 is determined as €81,405, and 3 times DV01 at €244,215 (TR).

Inherently there is uncertainty around the parameters chosen. In setting a parameter, there is expert judgment involved. As long as the parameter is within a reasonable range defined by the expert, then there might not be a misstatement issue. As part of the audit procedures it should be assessed whether the expert is informed with all relevant data and should be able to substantiate his/her choice.

Can we apply this concept to assess acceptable ranges for insurance liabilities as well?

DVOC-method

The concept underlying DV01 can be generalized for other market risks and actuarial parameters (risk drivers) like volatility, lapses, mortality et cetera. Hence – looking at the set of (mathematically spoken) 'first partial derivatives of the valuation (or risk metric)

4 – In this example we assume Euro denominated derivatives; for other currencies and/or other complex derivatives with e.g. steepener features, the DV01 and/or TR might not be appropriate.

function'. We have defined these sensitivities based on inherent variability as DVOC – Delta Value of Change. Insight in the DVOC has clear advantages. One is to conclude on the key parameters driving the majority (e.g. > 95%) of the model output. Another advantage is to use these key risk drivers to analyze their (inherent) variability; this can and should be used to determine the acceptable range of resulting outcomes – the Tolerable Range (TR) for audit purposes by setting objectively a multiplier of DVOC as TR (like the '3' for DV01).
We believe there are more advantages of using DVOC – one being for 'Analyses of Changes' (or 'Movement Analyses'). But this is out of scope of this article.

Like for swaps, quite often the value of an insurance liability is also a net value: difference between discounted value future premiums and discounted value future benefits. And even so for risk metrics (like the Solvency Capital Requirement as a Surplus-at-Risk measure: the surplus is the net difference between all assets and liabilities). Therefore we see the value added by generalizing the concept of DV01 to our DVOC.

A 'fair value' based on a model will be sensitive to various key (assumed) parameters. The key parameter resulting in the highest DVOC could be used to set the TR in evaluating the audit/model validation findings for that particular model.
The next example shows some (arbitrary chosen) traditional life insurance contracts to assess sensitivities to two risk drivers – their DVOCs, where the shocks are chosen and have to be benchmarked against 'inherent variability' (like we did for the interest rate to determine the '3').

	Pure endowment	Death Benefit (whole life)	Immediate annuity
Best Estimate Liability value	€ 133.669	€ 73.143	€ 235.910
	DVOC (% change relative to BE liability)		
Interest Rate shock (-/- 1bps)	0,33%	0,53%	0,09%
Longevity shock (10% relative)	1,35%	-8,16%	3,74%

If the inherent variability of mortality tables/ probabilities would be around 15% in total (relatively; e.g. to other 'equal valid' mortality tables, or 'age adjustments'), then the TR for the pure endowment could be set at around 2% of BE liability as an example(2% equals app. 1,35% * 15%/10%).
However, for the whole life product, the TR could as a result be set at around 12% of BE – which is 6 times higher than the pure endowment. This would imply that differences compared to the assessment value within this TR are still acceptable, even when caused by assumption deviations in interest rates or other risk drivers.

Considering another example: a typical Dutch disability product ('Individual AOV' for self-employed individuals). Disability products in general pay a fixed annuity when an insured person becomes disabled, until retirement age or until rehabilitation. The key drivers for the claim provision are recovery rates, disability duration, interest rates, retirement age, and, of course, the annuity amount.
In practice we see quite often that in the modeling of the provision, various approximations are used:
- Current age, retirement age and disability inception date *set to full months or even full years*;
- Recovery rates assumed *constant per full year*;
- *Rounding or approximations* of yield curves.

Using DVOC shows that the approximations in disability duration have, relatively speaking, by far the biggest impact. Therefore, Tolerable Range could be based on this risk driver.

Now having gained insight in more key sensitivities, it could be relevant to additionally disclose those around the key parameters. By using a DVOC approach the company does have this type of information available. And with this, also the objective measurability can be assessed.

It should be clear that simply setting a Tolerable Range is not sufficient for drawing conclusions and that actuaries and auditors should look in sufficient detail at the underlying process on *why* certain choices have been made. And by looking at those elements that really do matter, by using DVOC, this assessment can be made much more efficient, effective, and transparent.

Closing remarks

We believe that the concept of using DVOC is a concrete -and objective- solution in line with Solvency II standards.

There is need to evaluate the (expert) judgment and define an acceptable or 'Tolerable Range' to be able to assess whether projections can be considered acceptable. In contrast, an error (when found) in the model itself, the source data, the model parameters (key risk drivers) or underlying process is a 'hard' error (factual misstatement) and should be evaluated by the 'traditional' materiality considerations.

Using the concept of DVOC helps making the TR objective, based on those key risk drivers that really do matter.

We see an increased need of many stakeholders to look beyond realized (ex-post) cash flows and to gain comfort on forward looking figures like economic values and risk metrics. These figures are nothing more than an estimate within a range of possible outcomes. To assess reasonableness, auditors and actuaries need to work side by side to translate qualitative considerations into quantifiable measures.

With this article we hope to have set a step in this direction.

@ Reacties op dit artikel graag naar **redactie.actuaris@ag-ai.nl**

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

Box 1 | Terms and Abbreviations

BD2K—Big Data 2 Knowledge, is a trans-NIH initiative established to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge, and to maximise community engagement.

DOI—Digital Object Identifier; a code used to permanently and stably identify (usually digital) objects. DOIs provide a standard mechanism for retrieval of metadata about the object, and generally a means to access the data object itself.

FAIR—Findable, Accessible, Interoperable, Reusable.

FORCE11—The Future of Research Communications and e-Scholarship; a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing, initiated in 2011.

Interoperability—the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort.

JDDCP—Joint Declaration of Data Citation Principles; Acknowledging data as a first-class research output, and to support good research practices around data re-use, JDDCP proposes a set of guiding principles for citation of data within scholarly literature, another dataset, or any other research object.

RDF—Resource Description Framework; a globally-accepted framework for data and knowledge representation that is intended to be read and interpreted by machines.

concerned with long-term data stewardship; and a data science community mining, integrating and analysing new and existing data to advance discovery. To facilitate the reading of this manuscript by these diverse stakeholders, we provide definitions for common abbreviations in Box 1. Humans, however, are not the only critical stakeholders in the milieu of scientific data. Similar problems are encountered by the applications and computational agents that we task to undertake data retrieval and analysis on our behalf. These 'computational stakeholders' are increasingly relevant, and demand as much, or more, attention as their importance grows. One of the grand challenges of data-intensive science, therefore, is to improve knowledge discovery through assisting both humans, and their computational agents, in the discovery of, access to, and integration and analysis of, task-appropriate scientific data and other scholarly digital objects.

For certain types of important digital objects, there are well-curated, deeply-integrated, special-purpose repositories such as Genbank[3], Worldwide Protein Data Bank (wwPDB[4]), and UniProt[5] in the life sciences; Space Physics Data Facility (SPDF; http://spdf.gsfc.nasa.gov/) and Set of Identifications, Measurements and Bibliography for Astronomical Data (SIMBAD[6]) in the space sciences. These foundational and critical core resources are continuously curating and capturing high-value reference datasets and fine-tuning them to enhance scholarly output, provide support for both human and mechanical users, and provide extensive tooling to access their content in rich, dynamic ways. However, not all datasets or even data types can be captured by, or submitted to, these repositories. Many important datasets emerging from traditional, low-throughput bench science don't fit in the data models of these special-purpose repositories, yet these datasets are no less important with respect to integrative research, reproducibility, and reuse in general. Apparently in response to this, we see the emergence of numerous general-purpose data repositories, at scales ranging from institutional (for example, a single university), to open globally-scoped repositories such as Dataverse[7], FigShare (http://figshare.com), Dryad[8], Mendeley Data (https://data.mendeley.com/), Zenodo (http://zenodo.org/), DataHub (http://datahub.io), DANS (http://www.dans.knaw.nl/), and EUDat[9]. Such repositories accept a wide range of data types in a wide variety of formats, generally do not attempt to integrate or harmonize the deposited data, and place few restrictions (or requirements) on the descriptors of the data deposition. The resulting data ecosystem, therefore, appears to be moving away from centralization, is becoming more diverse, and less integrated, thereby exacerbating the discovery and re-usability problem for both human and computational stakeholders.

A specific example of these obstacles could be imagined in the domain of gene regulation and expression analysis. Suppose a researcher has generated a dataset of differentially-selected polyadenylation sites in a non-model pathogenic organism grown under a variety of environmental conditions that stimulate its pathogenic state. The researcher is interested in comparing the alternatively-polyadenylated genes in this local dataset, to other examples of alternative-polyadenylation, and the expression levels of these genes—both in this organism and related model organisms—during the infection process. Given that there is no special-purpose archive for differential polyadenylation data, and no model organism database for this pathogen, where does the researcher begin?

We will consider the current approach to this problem from a variety of data discovery and integration perspectives. If the desired datasets existed, where might they have been published, and how would one begin to search for them, using what search tools? The desired search would need to filter based on specific species, specific tissues, specific types of data (Poly-A, microarray, NGS), specific conditions (infection), and specific genes—is that information ('metadata') captured by the repositories, and if so, what formats is it in, is it searchable, and how? Once the data is discovered, can it be downloaded? In what format(s)? Can that format be easily integrated with private in-house data (the local dataset of alternative polyadenylation sites) as well as other data publications from third-parties and with the community's core gene/protein data repositories? Can this integration be

done automatically to save time and avoid copy/paste errors? Does the researcher have permission to use the data from these third-party researchers, under what license conditions, and who should be cited if a data-point is re-used?

Questions such as these highlight some of the barriers to data discovery and reuse, not only for humans, but even more so for machines; yet it is precisely these kinds of deeply and broadly integrative analyses that constitute the bulk of contemporary e-Science. The reason that we often need several weeks (or months) of specialist technical effort to gather the data necessary to answer such research questions is not the lack of appropriate technology; the reason is, that we do not pay our valuable digital objects the careful attention they deserve when we create and preserve them. Overcoming these barriers, therefore, necessitates that all stakeholders—including researchers, special-purpose, and general-purpose repositories—evolve to meet the emergent challenges described above. The goal is for scholarly digital objects of all kinds to become 'first class citizens' in the scientific publication ecosystem, where the quality of the publication—and more importantly, the impact of the publication—is a function of its ability to be accurately and appropriately found, re-used, and cited over time, by all stakeholders, both human and mechanical.

With this goal in-mind, a workshop was held in Leiden, Netherlands, in 2014, named 'Jointly Designing a Data Fairport'. This workshop brought together a wide group of academic and private stakeholders all of whom had an interest in overcoming data discovery and reuse obstacles. From the deliberations at the workshop the notion emerged that, through the definition of, and widespread support for, a minimal set of community-agreed guiding principles and practices, all stakeholders could more easily discover, access, appropriately integrate and re-use, and adequately cite, the vast quantities of information being generated by contemporary data-intensive science. The meeting concluded with a draft formulation of a set of foundational principles that were subsequently elaborated in greater detail—namely, that all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people. These are now referred to as the FAIR Guiding Principles. Subsequently, a dedicated FAIR working group, established by several members of the FORCE11 community[10] fine-tuned and improved the Principles. The results of these efforts are reported here.

The significance of machines in data-rich research environments

The emphasis placed on FAIRness being applied to both human-driven and machine-driven activities, is a specific focus of the FAIR Guiding Principles that distinguishes them from many peer initiatives (discussed in the subsequent section). Humans and machines often face distinct barriers when attempting to find and process data on the Web. Humans have an intuitive sense of 'semantics' (the meaning or intent of a digital object) because we are capable of identifying and interpreting a wide variety of contextual cues, whether those take the form of structural/visual/iconic cues in the layout of a Web page, or the content of narrative notes. As such, we are less likely to make errors in the selection of appropriate data or other digital objects, although humans will face similar difficulties if sufficient contextual metadata is lacking. The primary limitation of humans, however, is that we are unable to operate at the scope, scale, and speed necessitated by the scale of contemporary scientific data and complexity of e-Science. It is for this reason that humans increasingly rely on computational agents to undertake discovery and integration tasks on their behalf. This necessitates machines to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols that will be encountered during their self-guided exploration of the global data ecosystem. It also necessitates that the machines keep an exquisite record of provenance such that the data they are collecting can be accurately and adequately cited. Assisting these agents, therefore, is a critical consideration for all participants in the data management and stewardship process—from researchers and data producers to data repository hosts.

Throughout this paper, we use the phrase 'machine actionable' to indicate a continuum of possible states wherein a digital object provides increasingly more detailed information to an autonomously-acting, computational data explorer. This information enables the agent—to a degree dependent on the amount of detail provided—to have the capacity, when faced with a digital object never encountered before, to: a) identify the type of object (with respect to both structure and intent), b) determine if it is useful within the context of the agent's current task by interrogating metadata and/or data elements, c) determine if it is usable, with respect to license, consent, or other accessibility or use constraints, and d) take appropriate action, in much the same manner that a human would.

For example, a machine may be capable of determining the data-type of a discovered digital object, but not capable of parsing it due to it being in an unknown format; or it may be capable of processing the contained data, but not capable of determining the licensing requirements related to the retrieval and/or use of that data. The optimal state—where machines fully 'understand' and can autonomously and correctly operate-on a digital object—may rarely be achieved. Nevertheless, the FAIR principles provide 'steps along a path' toward machine-actionability; adopting, in whole or in part, the FAIR

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

principles, leads the resource along the continuum towards this optimal state. In addition, the idea of being machine-actionable applies in two contexts—first, when referring to the contextual metadata surrounding a digital object ('what is it?'), and second, when referring to the content of the digital object itself ('how do I process it/integrate it?'). Either, or both of these may be machine-actionable, and each forms its own continuum of actionability.

Finally, we wish to draw a distinction between data that is machine-actionable as a result of specific investment in software supporting that data-type, for example, bespoke parsers that understand life science wwPDB files or space science Space Physics Archive Search and Extract (SPASE) files, and data that is machine-actionable exclusively through the utilization of general-purpose, open technologies. To reiterate the earlier point—ultimate machine-actionability occurs when a machine can make a useful decision regarding data that it has not encountered before. This distinction is important when considering both (a) the rapidly growing and evolving data environment, with new technologies and new, more complex data-types continuously being developed, and (b) the growth of general-purpose repositories, where the data-types likely to be encountered by an agent are unpredictable. Creating bespoke parsers, in all computer languages, for all data-types and all analytical tools that require those data-types, is not a sustainable activity. As such, the focus on assisting machines in their discovery and exploration of data through application of more generalized interoperability technologies and standards at the data/repository level, becomes a first-priority for good data stewardship.

The FAIR Guiding Principles in detail

Representatives of the interested stakeholder-groups, discussed above, coalesced around four core desiderata—the FAIR Guiding Principles—and limited elaboration of these, which have been refined (Box 2) from the meeting's original draft, available at (https://www.force11.org/node/6062). A separate document that dynamically addresses community discussion relating to clarifications and explanations of the principles, and detailed guidelines for and examples of FAIR implementations, is currently being constructed (http://datafairport.org/fair-principles-living-document-menu). The FAIR Guiding Principles describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing, and reuse. While there have been a number of recent, often domain-focused publications advocating for specific improvements in practices relating to data management and archival[1,11,12], FAIR differs in that it describes concise, domain-independent, high-level principles that can be applied to a wide range of scholarly outputs. Throughout the Principles, we use the phrase '(meta)data' in cases where the Principle should be applied to both metadata and data.

The elements of the FAIR Principles are related, but independent and separable. The Principles define characteristics that contemporary data resources, tools, vocabularies and infrastructures should exhibit to assist discovery and reuse by third-parties. By minimally defining each guiding principle, the barrier-to-entry for data producers, publishers and stewards who wish to make their data holdings FAIR is purposely maintained as low as possible. The Principles may be adhered to in any combination and incrementally, as data providers' publishing environments evolve to increasing degrees of 'FAIRness'. Moreover, the modularity of the Principles, and their distinction between data and metadata, explicitly support a wide range of special circumstances. One such example is highly sensitive or personally-identifiable data, where publication of rich metadata to facilitate discovery, including clear rules regarding the process for accessing the data, provides a high degree of 'FAIRness' even in the absence of FAIR publication of the data itself. A second example involves the publication

of non-data research objects. Analytical workflows, for example, are a critical component of the scholarly ecosystem, and their formal publication is necessary to achieve both transparency and scientific reproducibility. The FAIR principles can equally be applied to these non-data assets, which need to be identified, described, discovered, and reused in much the same manner as data.

Specific exemplar efforts that provide varying levels of FAIRness are detailed later in this document. Additional issues, however, remain to be addressed. First, when community-endorsed vocabularies or other (meta)data standards do not include the attributes necessary to achieve rich annotation, there are two possible solutions: either publish an extension of an existing, closely related vocabulary, or—in the extreme case—create and explicitly publish a new vocabulary resource, following FAIR principles ('I2'). Second, to explicitly identify the standard chosen when more than one vocabulary or other (meta)data standard is available, and given that for instance in the life sciences there are over 600 content standards, the BioSharing registry (https://biosharing.org/) can be of use as it describes the standards in detail, including versions where applicable.

The Principles precede implementation

These high-level FAIR Guiding Principles precede implementation choices, and do not suggest any specific technology, standard, or implementation-solution; moreover, the Principles are not, themselves, a standard or a specification. They act as a guide to data publishers and stewards to assist them in evaluating whether their particular implementation choices are rendering their digital research artefacts Findable, Accessible, Interoperable, and Reusable. We anticipate that these high level principles will enable a broad range of integrative and exploratory behaviours, based on a wide range of technology choices and implementations. Indeed, many repositories are already implementing various aspects of FAIR using a variety of technology choices and several examples are detailed in the next section; examples include *Scientific Data* itself and how narrative data articles are anchored to a progressively FAIR structured metadata.

Examples of FAIRness, and the resulting value-added

Dataverse[7]: Dataverse is an open-source data repository software installed in dozens of institutions globally to support public community repositories or institutional research data repositories. Harvard Dataverse, with more than 60,000 datasets, is the largest of the current Dataverse repositories, and is open to all researchers from all research fields. Dataverse generates a formal citation for each deposit, following the standard defined by Altman and King[13]. Dataverse makes the Digital Object Identifier (DOI), or other persistent identifiers (Handles), public when the dataset is published ('F'). This resolves to a landing page, providing access to metadata, data files, dataset terms, waivers or licenses, and version information, all of which is indexed and searchable ('F', 'A', and 'R'). Deposits include metadata, data files, and any complementary files (such as documentation or code) needed to understand the data and analysis ('R'). Metadata is always public, even if the data are restricted or removed for privacy issues ('F', 'A'). This metadata is offered at three levels, extensively supporting the 'I' and 'R' FAIR principles: 1) data citation metadata, which maps to DataCite schema or Dublin Core Terms, 2) domain-specific metadata, which when possible maps to metadata standards used within a scientific domain, and 3) file-level metadata, which can be deep and extensive for tabular data files (including column-level metadata). Finally, Dataverse provides public machine-accessible interfaces to search the data, access the metadata and download the data files, using a token to grant access when data files are restricted ('A').

FAIRDOM (http://fair-dom.org/about): integrates the SEEK[14] and openBIS[15] platforms to produce a FAIR data and model management facility for Systems Biology. Individual research assets (or aggregates of data and models) are identified with unique and persistent HTTP URLs, which can be registered with DOIs for publication ('F'). Assets can be accessed over the Web in a variety of formats appropriate for individuals and/or their computers (RDF, XML) ('I'). Research assets are annotated with rich metadata, using community standards, formats and ontologies ('I'). The metadata is stored as RDF to enable interoperability and assets can be downloaded for reuse ('R').

ISA[16]: is a community-driven metadata tracking framework to facilitate standards-compliant collection, curation, management and reuse of life science datasets. ISA provides progressively FAIR structured metadata to Nature Scientific Data's Data Descriptor articles, and many GigaScience data papers, and underpins the EBI MetaboLights database among other data resources. At the heart is a general-purpose, extensible ISA model, originally only available as a tabular representation but subsequently enhanced as an RDF-based representation[17], and JSON serializations to enable the 'I' and 'R', becoming 'FAIR' when published as linked data (http://elixir-uk.org/node-events/201cisa-as-a-fair-research-object201d-hack-the-spec-event-1) and complementing other research objects[18].

Open PHACTS[19]: Open PHACTS is a data integration platform for information pertaining to drug discovery. Access to the platform is mediated through a machine-accessible interface[20] which provides multiple representations that are both human (HTML) and machine readable (RDF, JSON,

Box 3 | Emergent community/collaborative initiatives with FAIR as a core focus or activity

bioCADDIE (https://biocaddie.org): The NIH BD2K biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) consortium works to develop a Data Discovery Index (DDI) prototype, which is set to be as transformative and impactful for data as PubMed for the biomedical literature[30]. The DDI focuses on finding ('F') and accessing ('A') the datasets stored across different sources, and progressively works to identify relevant metadata[31] ('I') and maps them to community standards ('R'), linking to BioSharing.

CEDAR[32]: The Center for Expanded Data Annotation and Retrieval (CEDAR) is an NIH BD2K funded center of excellence to develop tools and technologies that reduce the burden of authoring and enhancing metadata that meet community-based standards. CEDAR will enable the creation of metadata templates that implement community based standards for experimental metadata, from BioSharing (https://biosharing.org), and that will be uniquely identifiable and retrievable with HTTP URIs, and annotated with vocabularies and ontologies drawn from BioPortal (http://bioportal.bioontology.org) ('F','A','I','R'). These templates will guide users to create rich metadata with unique and stable HTTP identifiers ('F') that can be retrieved using HTTP ('A') and accessible in a variety of formats (JSON-LD, TURTLE, RDF/XML, CSV, etc) ('I'). These metadata will use community standards, as defined by the template, and include provenance and data usage ('R').

These two projects, among others, provide tools and or collaborative opportunities for those who wish to improve the FAIRness of their data.

XML, CSV, etc), providing the 'A' facet of FAIRness. The interface allows multiple URLs to be used to access information about a particular entity through a mappings service ('F' and 'A'). Thus, a user can provide a ChEMBL URL to retrieve information sourced from, for example, Chemspider or DrugBank. Each call provides a canonical URL in its response ('A' and 'I'). All data sources used are described using standardized dataset descriptions, following the global VoID standard, with rich provenance ('R' and 'I'). All interface features are described using RDF following the Linked Data API specification ('A'). Finally, a majority of the datasets are described using community agreed upon ontologies ('I').

wwPDB[4,21]: wwPDB is a special-purpose, intensively-curated data archive that hosts information about experimentally-determined 3D structures of proteins and nucleic acids. All wwPDB entries are stably hosted on an FTP server ('A') and represented in machine-readable formats (text and XML); the latter are machine-actionable using the metadata provided by the wwPDB conforming to the Macromolecular Information Framework (mmCIF[22]), a data standard of the International Union of Crystallography (IUCr) ('F','I' for humans, 'F','I' for IUCr-aware machines). The wwPDB metadata contains cross-references to common identifiers such as PubMed and NCBI Taxonomy, and their wwPDB metadata are described in data dictionaries and schema documents (http://mmcif.wwpdb.org and http://pdbml.wwpdb.org) which conform to the IUCr data standard for the chemical and structural biology domains ('R'). A variety of software tools are available to interpret both wwPDB data and meta-data ('I','R' for humans, 'I','R' for machines with this software). Each entry is represented by a DOI ('F', 'A' for humans and machines). The DOI resolves to a zipped file which requires special software for further interrogation/interpretation. Other wwPDB access points[23–25] provide access to wwPDB records through URLs that are likely to be stable in the long-term ('F'), and all data and metadata is searchable through one or more of the wwPDB-affiliated websites ('F')

UniProt[26]: UniProt is a comprehensive resource for protein sequence and annotation data. All entries are uniquely identified by a stable URL, that provides access to the record in a variety of formats including a web page, plain-text, and RDF ('F' and 'A'). The record contains rich metadata ('F') that is both human-readable (HTML) and machine-readable (text and RDF), where the RDF formatted response utilizes shared vocabularies and ontologies such as UniProt Core, FALDO, and ECO ('I'). Interlinking with more than 150 different databases, every UniProt record has extensive links into, for example, PubMed, enabling rich citation. These links are machine-actionable in the RDF representation ('R'). Finally, in the RDF representation, the UniProt Core Ontology explicitly types all records, leaving no ambiguity—neither for humans nor machines—about what the data represents ('R'), enabling fully-automated retrieval of records and cross-referencing information.

In addition to, and in support of, communities and resources that are already pursuing FAIR objectives, the Data Citation Implementation Group of Force11 has published specific technical recommendations for how to implement many of the principles[27], with a particular focus on identifiers and their resolution, persistence, and metadata accessibility especially related to citation. In addition, the 'Skunkworks' group that emerged from the Lorentz Workshop has been creating software supporting infrastructures[28] that are, end-to-end, compatible with FAIR principles, and can be implemented over existing repositories. These code modules have a particular focus on metadata publication and searchability, compatibility in cases of strict privacy considerations, and the extremely difficult problem of data and metadata interoperability (manuscript in preparation). Finally, there are several emergent projects, some listed in Box 3, for which FAIR is a key objective. These projects may provide valuable advice and guidance for those wishing to become more FAIR.

FAIRness is a prerequisite for proper data management and data stewardship

The ideas within the FAIR Guiding Principles reflect, combine, build upon and extend previous work by both the Concept Web Alliance (https://conceptweblog.wordpress.com/) partners, who focused on machine-actionability and harmonization of data structures and semantics, and by the scientific and scholarly organizations that developed the Joint Declaration of Data Citation Principles (JDDCP[29]),

who focused on primary scholarly data being made citable, discoverable and available for reuse, so as to be capable of supporting more rigorous scholarship. An attempt to define the similarities and overlaps between the FAIR Principles and the JDDCP is provided at (https://www.force11.org/node/6062). The FAIR Principles are also complementary to the 'Data Seal of Approval' (DSA) (http://datasealofapproval.org/media/filer_public/2013/09/27/guidelines_2014-2015.pdf) in that they share the general aim to render data re-usable for users other than those who originally generated them. While the DSA focuses primarily on the responsibilities and conduct of data producers and repositories, FAIR focuses primarily on the data itself. Clearly, the broader community of stakeholders is coalescing around a set of common, dovetailed visions spanning all facets of the scholarly data publishing ecosystem.

The end result, when implemented, will be more rigorous management and stewardship of these valuable digital resources, to the benefit of the entire academic community. As stated at the outset, good data management and stewardship is not a goal in itself, but rather a pre-condition supporting knowledge discovery and innovation. Contemporary e-Science requires data to be Findable, Accessible, Interoperable, and Reusable in the long-term, and these objectives are rapidly becoming expectations of agencies and publishers. We demonstrate, therefore, that the FAIR Data Principles provide a set of mileposts for data producers and publishers. They guide the implementation of the most basic levels of good Data Management and Stewardship practice, thus helping researchers adhere to the expectations and requirements of their funding agencies. We call on all data producers and publishers to examine and implement these principles, and actively participate with the FAIR initiative by joining the Force11 working group. By working together towards shared, common goals, the valuable data produced by our community will gradually achieve the critical goals of FAIRness.

References

1. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biol.* **13,** e1002295 (2015).
2. Bechhofer, S. *et al.* Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nat. Preced.* doi:10.1038/npre.2010.4626.1 (2010).
3. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41,** D36–D42 (2013).
4. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10,** 980–980 (2003).
5. The Uniprot Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).
6. Wenger, M. *et al.* The SIMBAD astronomical database-The CDS reference database for astronomical objects. *Astron. Astrophys. Suppl. Ser.* **143,** 9–22 (2000).
7. Crosas, M. "The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data". *D-Lib Mag* **17** (1), p2 (2011).
8. White, H. C., Carrier, S., Thompson, A., Greenberg, J. & Scherle, R. The Dryad data repository: A Singapore framework metadata architecture in a DSpace environment. *Univ. Göttingen*, p157 (2008).
9. Lecarpentier, D. *et al.* EUDAT: A New Cross-Disciplinary Data Infrastructure for Science. *Int. J. Digit. Curation* **8,** 279–287 (2013).
10. Martone, M. E. FORCE11: Building the Future for Research Communications and e-Scholarship. *Bioscience* **65,** 635 (2015).
11. White, E. *et al.* Nine simple ways to make it easier to (re)use your data. *Ideas Ecol. Evol.* **6** (2013).
12. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput. Biol.* **9,** e1003285 (2013).
13. Altman, M. & King, G. in *D-Lib Magazine* **13,** no. 3/4 (2007).
14. Wolstencroft, K. *et al.* SEEK: a systems biology data and model management platform. *BMC Syst. Biol.* **9,** 33 (2015).
15. Bauch, A. *et al.* openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics* **12,** 468 (2011).
16. Sansone, S.-A. *et al.* Toward interoperable bioscience data. *Nat. Genet.* **44,** 121–126 (2012).
17. González-Beltrán, A., Maguire, E., Sansone, S.-A. & Rocca-Serra, P. linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics* **15,** S4 (2014).
18. González-Beltrán, A. *et al.* From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics. *PLoS ONE* **10,** e0127612 (2015).
19. Harland, L. Open PHACTS: A Semantic Knowledge Infrastructure for Public and Commercial Drug Discovery Research. *Knowl. Eng. Knowl. Manag. Lect. Notes Comput. Sci.* **7603/2012,** 1–7 (2012).
20. Groth, P. *et al.* API-centric Linked Data integration: The Open PHACTS Discovery Platform case study. *Web Semant. Sci. Serv. Agents World Wide Web* **29,** 12–18 (2014).
21. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
22. Bourne, P. E., Berman, H. M., Watenpaugh, K., Westbrook, J. D. & Fitzgerald, P. M. D. The macromolecular crystallographic information file (mmCIF). *Meth. Enzym* **277,** 571–590 (1997).
23. Rose, P. W. *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **43,** D345–D356 (2015).
24. Kinjo, A. R. *et al.* Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res.* **40,** D453–D460 (2012).
25. Gutmanas, A. *et al.* PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* **42,** D285–D291 (2014).
26. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).
27. Starr, J. *et al.* Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Comput. Sci.* **1,** e1 (2015).
28. Wilkinson, M., Dumontier, M. & Durbin, P. DataFairPort: The Perl libraries version 0.231 doi:10.5281/zenodo.33584 (2015).
29. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. San Diego CA: FORCE11. https://www.force11.org/datacitation (2014).
30. Ohno-machado, L. *et al.* NIH BD2K bioCADDIE white paper—Data Discovery Index. http://dx.doi.org/10.6084/m9.figshare.1362572 (2015).

31. NIH BD2K bioCADDIE WG3 Members. WG3-MetadataSpecifications: NIH BD2K bioCADDIE Data Discovery Index WG3 Metadata Specification v1 doi:10.5281/zenodo.28019 (2015).
32. Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J. Am. Med. Informatics Assoc.* **22,** 1148–1152 (2015).

Acknowledgements

The original Lorentz Workshop 'Jointly Designing a Data FAIRport' was organized by Barend Mons in collaboration with and co-sponsored by the Lorentz center, The Dutch Techcenter for the Life Sciences and the Netherlands eScience Center. The principles and themes described in this manuscript represent the significant voluntary contributions and participation of the authors at, and/or subsequent to, this workshop and from the wider Force11, BD2K and ELIXIR communities. We also acknowledge and thank the organizers and backers of the NBDC/DBCLS BioHackathon 2015, where several of the authors made significant revisions to the FAIR Principles.

Author Contributions

M.W. was the primary author of the manuscript, and participated extensively in the drafting and editing of the FAIR Principles. M.D. was significantly involved in the drafting of the FAIR Principles. B.M. conceived of the FAIR Data Initiative, contributed extensively to the drafting of the principles, and to this manuscript text. All other authors are listed alphabetically, and contributed to the manuscript either by their participation in the initial workshop and/or by editing or commenting on the manuscript text.

Additional Information

Competing financial interests: M.A. is the *Nature Genetics*' Editor in Chief; S.A.S. is *Scientific Data*'s Honorary Academic Editor and consultant.

How to cite this article: Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Mark D. Wilkinson[1], Michel Dumontier[2], IJsbrand Jan Aalbersberg[3], Gabrielle Appleton[3], Myles Axton[4], Arie Baak[5], Niklas Blomberg[6], Jan-Willem Boiten[7], Luiz Bonino da Silva Santos[8], Philip E. Bourne[9], Jildau Bouwman[10], Anthony J. Brookes[11], Tim Clark[12], Mercè Crosas[13], Ingrid Dillo[14], Olivier Dumon[3], Scott Edmunds[15], Chris T. Evelo[16], Richard Finkers[17], Alejandra Gonzalez-Beltran[18], Alasdair J.G. Gray[19], Paul Groth[3], Carole Goble[20], Jeffrey S. Grethe[21], Jaap Heringa[22], Peter A.C. 't Hoen[23], Rob Hooft[24], Tobias Kuhn[25], Ruben Kok[22], Joost Kok[26], Scott J. Lusher[27], Maryann E. Martone[28], Albert Mons[29], Abel L. Packer[30], Bengt Persson[31], Philippe Rocca-Serra[18], Marco Roos[32], Rene van Schaik[33], Susanna-Assunta Sansone[18], Erik Schultes[34], Thierry Sengstag[35], Ted Slater[36], George Strawn[37], Morris A. Swertz[38], Mark Thompson[32], Johan van der Lei[39], Erik van Mulligen[39], Jan Velterop[40], Andra Waagmeester[41], Peter Wittenburg[42], Katherine Wolstencroft[43], Jun Zhao[44] & Barend Mons[45,46,47]

[1]Center for Plant Biotechnology and Genomics, Universidad Politécnica de Madrid, Madrid 28223, Spain. [2]Stanford University, Stanford 94305-5411, USA. [3]Elsevier, Amsterdam 1043 NX, The Netherlands. [4]Nature Genetics, New York 10004-1562, USA. [5]Euretos and Phortos Consultants, Rotterdam 2741 CA, The Netherlands. [6]ELIXIR, Wellcome Genome Campus, Hinxton CB10 1SA, UK. [7]Lygature, Eindhoven 5656 AG, The Netherlands. [8]Vrije Universiteit Amsterdam, Dutch Techcenter for Life Sciences, Amsterdam 1081 HV, The Netherlands. [9]Office of the Director, National Institutes of Health, Rockville 20892, USA. [10]TNO, Zeist 3700 AJ, The Netherlands. [11]Department of Genetics, University of Leicester, Leicester LE1 7RH, UK. [12]Harvard Medical School, Boston, Massachusetts MA 02115, USA. [13]Harvard University, Cambridge, Massachusetts MA 02138, USA. [14]Data Archiving and Networked Services (DANS), The Hague 2593 HW, The Netherlands. [15]GigaScience, Beijing Genomics Institute, Shenzhen 518083, China. [16]Department of Bioinformatics, Maastricht University, Maastricht 6200 MD, The Netherlands. [17]Wageningen UR Plant Breeding, Wageningen 6708 PB, The Netherlands. [18]Oxford e-Research Center, University of Oxford, Oxford OX1 3QG, UK. [19]Heriot-Watt University, Edinburgh EH14 4AS, UK. [20]School of Computer Science, University of Manchester, Manchester M13 9PL, UK. [21]Center for Research in Biological Systems, School of Medicine, University of California San Diego, La Jolla, California 92093-0446, USA. [22]Dutch Techcenter for the Life Sciences, Utrecht 3501 DE, The Netherlands. [23]Department of Human Genetics, Leiden University Medical Center, Dutch Techcenter for the Life Sciences, Leiden 2300 RC, The Netherlands. [24]Dutch TechCenter for Life Sciences and ELIXIR-NL, Utrecht 3501 DE, The

Netherlands. [25]VU University Amsterdam, Amsterdam 1081 HV, The Netherlands. [26]Leiden Center of Data Science, Leiden University, Leiden 2300 RA, The Netherlands. [27]Netherlands eScience Center, Amsterdam 1098 XG, The Netherlands. [28]National Center for Microscopy and Imaging Research, UCSD, San Diego 92103, USA. [29]Phortos Consultants, San Diego 92011, USA. [30]SciELO/FAPESP Program, UNIFESP Foundation, São Paulo 05468-901, Brazil. [31]Bioinformatics Infrastructure for Life Sciences (BILS), Science for Life Laboratory, Dept of Cell and Molecular Biology, Uppsala University, S-751 24, Uppsala, Sweden. [32]Leiden University Medical Center, Leiden 2333 ZA, The Netherlands. [33]Bayer CropScience, Gent Area 1831, Belgium. [34]Leiden Institute for Advanced Computer Science, Leiden University Medical Center, Leiden 2300 RA, The Netherlands. [35]Swiss Institute of Bioinformatics and University of Basel, Basel 4056, Switzerland. [36]Cray, Inc., Seattle 98164, USA. [37]Unaffiliated. [38]University Medical Center Groningen (UMCG), University of Groningen, Groningen 9713 GZ, The Netherlands. [39]Erasmus MC, Rotterdam 3015 CE, The Netherlands. [40]Independent Open Access and Open Science Advocate, Guildford GU1 3PW, UK. [41]Micelio, Antwerp 2180, Belgium. [42]Max Planck Compute and Data Facility, MPS, Garching 85748, Germany. [43]Leiden Institute of Advanced Computer Science, Leiden University, Leiden 2333 CA, The Netherlands. [44]Department of Computer Science, Oxford University, Oxford OX1 3QD, UK. [45]Leiden University Medical Center, Leiden and Dutch TechCenter for Life Sciences, Utrecht 2333 ZA, The Netherlands. [46]Netherlands eScience Center, Amsterdam 1098 XG, The Netherlands. [47]Erasmus MC, Rotterdam 3015 CE, The Netherlands.

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases. Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice [1, 13, 24], hiring [19], critical infrastructure [11, 21], and finance [18]. Even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets [4, 5, 12]. For these and other reasons, the World Economic Forum suggests that all entities should document the provenance, creation, and use of machine learning datasets in order to avoid discriminatory outcomes [25].

Although data provenance has been studied extensively in the databases community [3, 8], it is rarely discussed in the machine learning community. Documenting the creation and use of datasets has received even less attention. Despite the importance of data to machine learning, there is currently no standardized process for documenting machine learning datasets.

To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, we propose that every

Authors' addresses: Timnit Gebru, Black in AI; Jamie Morgenstern, University of Washington; Briana Vecchione, Cornell University; Jennifer Wortman Vaughan, Microsoft Research; Hanna Wallach, Microsoft Research; Hal Daumé III, Microsoft Research; University of Maryland; Kate Crawford, Microsoft Research.

dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks.

After outlining our objectives below, we describe the process by which we developed datasheets for datasets. We then provide a set of questions designed to elicit the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions. We conclude with a summary of the impact to date of datasheets for datasets and a discussion of implementation challenges and avenues for future work.

1.1 Objectives

Datasheets for datasets are intended to address the needs of two key stakeholder groups: dataset creators and dataset consumers. For dataset creators, the primary objective is to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use. For dataset consumers, the primary objective is to ensure they have the information they need to make informed decisions about using a dataset. Transparency on the part of dataset creators is necessary for dataset consumers to be sufficiently well informed that they can select appropriate datasets for their chosen tasks and avoid unintentional misuse.[1]

Beyond these two key stakeholder groups, datasheets for datasets may be valuable to policy makers, consumer advocates, investigative journalists, individuals whose data is included in datasets, and individuals who may be impacted by models trained or evaluated using datasets. They also serve a secondary objective of facilitating greater reproducibility of machine learning results: researchers and practitioners without access to a dataset may be able to use the information in its datasheet to create alternative datasets with similar characteristics.

Although we provide a set of questions designed to elicit the information that a datasheet for a dataset might contain, these questions are not intended to be prescriptive. Indeed, we expect that datasheets will necessarily vary depending on factors such as the domain or existing organizational infrastructure and workflows. For example, some the questions are appropriate for academic researchers publicly releasing datasets for the purpose of enabling future

[1]We note that in some cases, the people creating a datasheet for a dataset may not be the dataset creators, as was the case with the example datasheets that we created as part of our development process.

research, but less relevant for product teams creating internal datasets for training proprietary models. As another example, Bender and Friedman [2] outline a proposal similar to datasheets for datasets specifically intended for language-based datasets. Their questions may be naturally integrated into a datasheet for a language-based dataset as appropriate.

We emphasize that the process of creating a datasheet is not intended to be automated. Although automated documentation processes are convenient, they run counter to our objective of encouraging dataset creators to carefully reflect on the process of creating, distributing, and maintaining a dataset.

2 Development Process

We refined the questions and workflow provided in the next section over a period of roughly two years, incorporating many rounds of feedback.

First, leveraging our own experiences as researchers with diverse backgrounds working in different domains and institutions, we drew on our knowledge of dataset characteristics, unintentional misuse, unwanted societal biases, and other issues to produce an initial set of questions designed to elicit information about these topics. We then "tested" these questions by creating example datasheets for two widely used datasets: Labeled Faces in the Wild [16] and Pang and Lee's polarity dataset [22]. We chose these datasets in large part because their creators provided exemplary documentation, allowing us to easily find the answers to many of the questions. While creating these example datasheets, we found gaps in the questions, as well as redundancies and lack of clarity. We therefore refined the questions and distributed them to product teams in two major US-based technology companies, in some cases helping teams to create datasheets for their datasets and observing where the questions did not achieve their intended objectives. Contemporaneously, we circulated an initial draft of this paper to colleagues through social media and on arXiv (draft posted 23 March 2018). Via these channels we received extensive comments from dozens of researchers, practitioners, and policy makers. We also worked with a team of lawyers to review the questions from a legal perspective.

We incorporated this feedback to yield the questions and workflow provided in the next section: We added and removed questions, refined the content of the questions, and reordered the questions to better match the key stages of the dataset lifecycle. Based on our experiences with product teams, we reworded the questions to discourage yes/no answers, added a section on "Uses," and deleted a section on "Legal and Ethical Considerations." We found that product teams were more likely to answer questions about legal and ethical considerations if they were integrated into sections about the relevant stages of the dataset lifecycle rather than grouped together. Finally, following feedback from the team of lawyers, we removed questions that explicitly asked about compliance with regulations, and introduced factual questions intended to

elicit relevant information about compliance without requiring dataset creators to make legal judgments.

3 Questions and Workflow

In this section, we provide a set of questions designed to elicit the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions. The questions are grouped into sections that roughly match the key stages of the dataset lifecycle: motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance. This grouping encourages dataset creators to reflect on the process of creating, distributing, and maintaining a dataset, and even alter this process in response to their reflection. We note that not all questions will be applicable to all datasets; those that do not apply should be skipped.

To illustrate how these questions might be answered in practice, we provide in the appendix an example datasheet for Pang and Lee's polarity dataset [22]. We answered some of the questions with "Unknown to the authors of the datasheet." This is because we did not create the dataset ourselves and could not find the answers to these questions in the available documentation. For an example of a datasheet that was created by the creators of the corresponding dataset, please see that of Cao and Daumé [6].[2] We note that even dataset creators may be unable to answer all of the questions provided in this section. We recommend answering as many questions as possible rather than skipping the datasheet creation process entirely.

3.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
- **Any other comments?**

[2] See https://github.com/TristaCao/into_inclusivecoref/blob/master/GICoref/datasheet-gicoref.md.

3.2 Composition

Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
- **How many instances are there in total (of each type, if appropriate)?**
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
- **Is there a label or target associated with each instance?** If so, please provide a description.
- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

If the dataset does not relate to people, you may skip the remaining questions in this section.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.
- **Any other comments?**

3.3 Collection Process

As with the questions in the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics. Again, questions that apply

only to datasets that relate to people are grouped together at the end of the section.

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

If the dataset does not relate to people, you may skip the remaining questions in this section.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
- **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
- **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or**

for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
- **Any other comments?**

3.4 Preprocessing/cleaning/labeling

Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
- **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
- **Any other comments?**

3.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

- **Has the dataset been used for any tasks already?** If so, please provide a description.
- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.
- **What (other) tasks could the dataset be used for?**

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
- **Any other comments?**

3.6 Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
- **When will the dataset be distributed?**
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
- **Any other comments?**

3.7 Maintenance

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions

in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

- **Who will be supporting/hosting/maintaining the dataset?**
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
- **Is there an erratum?** If so, please provide a link or other access point.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
- **Any other comments?**

4 Impact and Challenges

Since circulating an initial draft of this paper in March 2018, datasheets for datasets have already gained traction in a number of settings. Academic researchers have adopted our proposal and released datasets with accompanying datasheets [e.g., 7, 10, 23, 26]. Microsoft, Google, and IBM have begun to pilot datasheets for datasets internally within product teams. Researchers at Google published follow-up work on *model cards* that document machine learning models [20] and released a *data card* (a lightweight version of a datasheet) along with the Open Images dataset [17]. Researchers at IBM proposed *factsheets* [14] that document various characteristics of AI services, including whether the datasets used to develop the services are accompanied with datasheets. The Data Nutrition Project incorporated some of the questions provided in the previous section into the latest release of their Dataset Nutrition Label [9]. Finally, the Partnership on AI, a multi-stakeholder organization focused on

studying and formulating best practices for developing and deploying AI technologies, is working on industry-wide documentation guidance that builds on datasheets for datasets, model cards, and factsheets.[3]

These initial successes have also revealed implementation challenges that may need to be addressed to support wider adoption. Chief among them is the need for dataset creators to modify the questions and workflow provided in the previous section based on their existing organizational infrastructure and workflows. We also note that the questions and workflow may pose problems for dynamic datasets. If a dataset changes only infrequently, we recommend accompanying updated versions with updated datasheets.

Datasheets for datasets do not provide a complete solution to mitigating unwanted societal biases or potential risks or harms. Dataset creators cannot anticipate every possible use of a dataset, and identifying unwanted societal biases often requires additional labels indicating demographic information about individuals, which may not be available to dataset creators for reasons including those individuals' data protection and privacy [15].

When creating datasets that relate to people, and hence their accompanying datasheets, it may be necessary for dataset creators to work with experts in other domains such as anthropology, sociology, and science and technology studies. There are complex and contextual social, historical, and geographical factors that influence how best to collect data from individuals in a manner that is respectful.

Finally, creating datasheets for datasets will necessarily impose overhead on dataset creators. Although datasheets may reduce the amount of time that dataset creators spend answering one-off questions about datasets, the process of creating a datasheet will always take time, and organizational infrastructure and workflows—not to mention incentives—will need to be modified to accommodate this investment.

Despite these implementation challenges, there are many benefits to creating datasheets for datasets. In addition to facilitating better communication between dataset creators and dataset consumers, datasheets provide an opportunity for dataset creators to distinguish themselves as prioritizing transparency and accountability. Ultimately, we believe that the benefits to the machine learning community outweigh the costs.

Acknowledgments

We thank Peter Bailey, Emily Bender, Yoshua Bengio, Sarah Bird, Sarah Brown, Steven Bowles, Joy Buolamwini, Amanda Casari, Eric Charran, Alain Couillault, Lukas Dauterman, Leigh Dodds, Miroslav Dudík, Michael Ekstrand, Noémie Elhadad, Michael Golebiewski, Nick Gonsalves, Martin Hansen, Andy Hickl, Michael Hoffman, Scott Hoogerwerf, Eric Horvitz, Mingjing Huang, Surya

[3] https://www.partnershiponai.org/about-ml/

Kallumadi, Ece Kamar, Krishnaram Kenthapadi, Emre Kiciman, Jacquelyn Krones, Erik Learned-Miller, Lillian Lee, Jochen Leidner, Rob Mauceri, Brian Mcfee, Emily McReynolds, Bogdan Micu, Margaret Mitchell, Sangeeta Mudnal, Brendan O'Connor, Thomas Padilla, Bo Pang, Anjali Parikh, Lisa Peets, Alessandro Perina, Michael Philips, Barton Place, Sudha Rao, Jen Ren, David Van Riper, Anna Roth, Cynthia Rudin, Ben Shneiderman, Biplav Srivastava, Ankur Teredesai, Rachel Thomas, Martin Tomko, Panagiotis Tziachris, Meredith Whittaker, Hans Wolters, Ashly Yeo, Lu Zhang, and the attendees of the Partnership on AI's April 2019 ABOUT ML workshop for valuable feedback.

References

[1] Don A Andrews, James Bonta, and J Stephen Wormith. 2006. The recent past and near future of risk and/or need assessment. *Crime & Delinquency* 52, 1 (2006), 7–27.

[2] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[3] Anant P. Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshpande, Aaron J. Elmore, Samuel Madden, and Aditya G. Parameswaran. 2014. DataHub: Collaborative Data Science & Dataset Version Management at Scale. *CoRR* abs/1409.0798 (2014).

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. 77–91.

[6] Yang Trista Cao and Hal Daumé. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. abs/1910.13913.

[7] Yang Trista Cao and Hal Daumé, III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

[8] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2009. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases* 1, 4 (2009), 379–474.

[9] Kasia Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2020. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. In *NeurIPS Workshop on Dataset Curation and Security*.

[10] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[11] Glennda Chui. 2017. Project will use AI to prevent or minimize electric grid failures. [Online; accessed 14-March-2018].

[12] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[13] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. 2016. *The Perpetual Line-Up: Unregulated Police Face Recognition in America.* Georgetown Law, Center on Privacy &

Technology, New Jersey Ave NW, Washington, DC.

[14] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R. Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. *CoRR* abs/1808.07261 (2018).

[15] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *2019 ACM CHI Conference on Human Factors in Computing Systems*.

[16] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Technical Report 07-49. University of Massachusetts Amherst.

[17] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. OpenImages: A public dataset for large-scale multi-label and multi-class image classification.

[18] Tom CW Lin. 2012. The new investor. *UCLA Law Review* 60 (2012), 678.

[19] G Mann and C O'Neil. 2016. Hiring Algorithms Are Not Neutral. https://hbr.org/2016/12/hiring-algorithms-are-not-neutral.

[20] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. 220–229.

[21] Mary Catherine O'Connor. 2017. How AI Could Smarten Up Our Water System. [Online; accessed 14-March-2018].

[22] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. 271.

[23] Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. 2018. Baselines and a datasheet for the Cerema AWP dataset. *CoRR* abs/1806.04016 (2018). http://arxiv.org/abs/1806.04016

[24] Doha Suppy Systems. 2017. Facial Recognition. [Online; accessed 14-March-2018].

[25] World Economic Forum Global Future Council on Human Rights 2016–2018. 2018. How to Prevent Discriminatory Outcomes in Machine Learning. https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning.

[26] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

A Appendix

In this appendix, we provide an example datasheet for Pang and Lee's polarity dataset [22] (figure 1 to figure 4).

Movie Review Polarity **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

[1] All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

Fig. 1. Example datasheet for Pang and Lee's polarity dataset [22], page 1.

Movie Review Polarity **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is entirely self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
Unknown to the authors of the datasheet.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Some movie reviews might contain moderately inappropriate or offensive language, but we do not expect this to be the norm.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
Some personal information is retained from the newsgroup posting in the "raw form" of the dataset (as opposed to the "preprocessed" version, in which these are automatically removed), including the name and email address the author posted under (note that these are already public on the internet newsgroup archive).

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
Aside from the aforementioned name/email addresses, no.

Any other comments?
None.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data was mostly observable as raw text, except that the labels were extracted by the process described below. The data was collected by downloading reviews from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, at http://reviews.imdb.com/Reviews.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
Unknown to the authors of the datasheet.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The sample of instances collected is English movie reviews from the `rec.arts.movies.reviews` newsgroup, from which a "number of stars" rating could be extracted. The sample is limited to forty reviews per unique author in order to achieve broader coverage by authorship. Beyond that, the sample is arbitrary.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
Unknown to the authors of the datasheet.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
Unknown to the authors of the datasheet.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
Unknown to the authors of the datasheet.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
As described above, the data was collected from newsgroups.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
No. The data was crawled from public web sources, and the authors of the posts presumably knew that their posts would be public, but the authors were not explicitly informed that their posts were to be used in this way.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
No (see previous question).

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
N/A.

Any other comments?

Fig. 2. Example datasheet for Pang and Lee's polarity dataset [22], page 2.

Movie Review Polarity **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

None.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

Instances for which an explicit rating could not be found were discarded. Also only instances with strongly-positive or strongly-negative ratings were retained. Star ratings were extracted by automatically looking for text like "`**** out of *****`" in the review, using that as a label, and then removing the corresponding text. When the star rating was out of five stars, anything at least four was considered positive and anything at most two negative; when out of four, three and up is considered positive, and one or less is considered negative. Occasionally half stars are missed which affects the labeling of negative examples. Everything in the middle was discarded. In order to ensure that sufficiently many authors are represented, at most 20 reviews (per positive/negative label) per author are included.

In a later version of the dataset (v1.1), non-English reviews were also removed.

Some preprocessing errors were caught in later versions. The following fixes were made: (1) Some reviews had rating information in several places that was missed by the initial filters; these are removed. (2) Some reviews had unexpected/unparsed ranges and these were fixed. (3) Sometimes the boilerplate removal removed too much of the text.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes. The dataset itself contains all the raw data.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

No.

Any other comments?

None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

At the time of publication, only the original paper (http://xxx.lanl.gov/pdf/cs/0409058v1). Between then and 2012, a collection of papers that used this dataset was maintained at http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

There is a repository, maintained by Pang/Lee through April 2012, at http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html.

What (other) tasks could the dataset be used for?

The dataset could be used for anything related to modeling or understanding movie reviews. For instance, one may induce a lexicon of words/phrases that are highly indicative of sentiment polarity, or learn to automatically generate movie reviews.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

There is minimal risk for harm: the data was already public, and in the preprocessed version, names and email addresses were removed.

Are there tasks for which the dataset should not be used? If so, please provide a description.

This data is collected solely in the movie review domain, so systems trained on it may or may not generalize to other sentiment prediction tasks. Consequently, such systems should not—without additional verification—be used to make consequential decisions about people.

Any other comments?

None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset is distributed on Bo Pang's webpage at Cornell: http://www.cs.cornell.edu/people/pabo/movie-review-data. The dataset does not have a DOI and there is no redundant archive.

When will the dataset be distributed?

The dataset was first released in 2002.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The crawled data copyright belongs to the authors of the reviews unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: *Thumbs up? Sentiment classification using machine learning techniques.* Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Proceedings of EMNLP, 2002.

Fig. 3. Example datasheet for Pang and Lee's polarity dataset [22], page 3.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
Unknown to authors of the datasheet.

Any other comments?
None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
Bo Pang is supporting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The curators of the dataset, Bo Pang and Lillian Lee, can be contacted at https://sites.google.com/site/bopang42/ and http://www.cs.cornell.edu/home/llee, respectively.

Is there an erratum? If so, please provide a link or other access point.
Since its initial release (v0.9) there have been three later releases (v1.0, v1.1, and v2.0). There is not an explicit erratum, but updates and known errors are specified in higher version `README` and `diff` files. There are several versions of these: v1.0: http://www.cs.cornell.edu/people/pabo/movie-review-data/README; v1.1: http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/README.1.1 and http://www.cs.cornell.edu/people/pabo/movie-review-data/diff.txt; v2.0: http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/poldata.README.2.0.txt. Updates are listed on the dataset web page. (This datasheet largely summarizes these sources.)

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
This will be posted on the dataset webpage.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
The dataset has already been updated; older versions are kept around for consistency.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?
None.

Fig. 4. Example datasheet for Pang and Lee's polarity dataset [22], page 4.

Ministerie van Justitie en Veiligheid

> Retouradres Postbus 20301 2500 EH Den Haag

Aan de Voorzitter van de Tweede Kamer
der Staten-Generaal
Postbus 20018
2500 EA Den Haag

Directoraat-Generaal Politie en Veiligheidsregio's

Turfmarkt 147
2511 DP Den Haag
Postbus 20301
2500 EH Den Haag
www.rijksoverheid.nl/jenv

Ons kenmerk
2829671

Bij beantwoording de datum en ons kenmerk vermelden. Wilt u slechts één zaak in uw brief behandelen.

Datum 18 februari 2020
Onderwerp Beantwoording schriftelijke vragen AI bij de politie

Hierbij doe ik u, mede namens de Minister voor Rechtsbescherming, de antwoorden toekomen op de schriftelijke vragen over de kamerbrief 'Artificiële intelligentie bij de politie'[1] die de vaste commissie voor Justitie en Veiligheid op 10 februari 2020 aan mijn ministerie heeft gericht. Ik dank de commissie voor de interesse en belangstelling die uit de 82 gestelde vragen blijkt.

Ik maak graag van de gelegenheid gebruik om eerst in het algemeen de ontwikkeling en toepassing van artificiële intelligentie (AI) bij de politie toe te lichten.

Algoritmes worden al decennia gebruikt bij de politie voor vele politieprocessen. In de basis is een algoritme een serie van instructies in een workflow om een wiskundig probleem op te lossen. Niet elk algoritme (en ook niet elke data science toepassing) is AI. Pas als het gaat om algoritmes waarbij systemen intelligent gedrag kunnen vertonen en gemaakt zijn om in meer of mindere mate zelfstandig te kunnen leren en acties te kunnen ondernemen spreken we van AI.

In de praktijk betekent dit dat veel van de data science toepassingen die de politie gebruikt, waaronder het Criminaliteit anticipatie systeem (CAS), geen AI component hebben. Ook niet als deze meer complexe algoritmes bevatten dan voorheen of gebruik maken van technieken die ook voor AI gebruikt worden. Het gaat in die gevallen vaak om het gebruik van meer eenvoudige algoritmes of algoritmes die eenvoudig zijn te verklaren. Deze meer eenvoudige algoritmes zijn reeds geruime tijd in gebruik bij de politie. Het CAS is daar een goed voorbeeld van. De zogenoemde hotspot benadering, waar op basis van bestaande data wordt bepaald waar het waarschijnlijk is dat bepaalde vormen van openbare orde verstoring of criminaliteit voorkomt, wordt al vele jaren gebruikt. Het CAS is daar de meest recente en geavanceerde versie van. CAS geeft een verwachting voor een bepaald gebied en geen output die tot personen te herleiden zijn.

Hoewel het helder is dat veel vormen van gebruik van data geen AI zijn, is het minder eenvoudig om exact te stellen wanneer er wel sprake is van AI.

[1] Kamerstukken II, 2019/20, 26643, 652.

Directoraat-Generaal Politie en Veiligheidsregio's

Datum
18 februari 2020

Ons kenmerk
2829671

Slechts in een aantal gevallen ontwikkelt en gebruikt de politie op dit moment AI. Dit aantal zal in de toekomst gaan groeien. Zoals ik in genoemde brief uiteen heb gezet is het gebruik van AI bij de politie geen luxe maar noodzaak om effectief te kunnen blijven optreden. De verwachting is dat AI uiteindelijk tot een breed geaccepteerde en gebruikte techniek zal uitgroeien. Zoals bij meer nieuwere vormen van techniek is het van groot belang om de ontwikkeling zorgvuldig vorm te geven. In mijn brief en uit de antwoorden van onderstaande vragen blijkt dat dit bij de politie voorop staat.

Om de beantwoording van vragen gestructureerd vorm te geven heb ik er voor gekozen om de vragen per onderwerp te clusteren. Ik zal eerst starten met de beantwoording van vragen over toepassingen van AI bij de politie.

De Minister van Justitie en Veiligheid,

Ferd Grapperhaus

Lijst van vragen en antwoorden over de brief van 3 december 2019 inzake Artificiële intelligentie bij de politie (26643, nr. 652).

Directoraat-Generaal Politie en Veiligheidsregio's

Datum
18 februari 2020

Ons kenmerk
2829671

Toepassingen van AI

17. Kunt u, zo nodig vertrouwelijk, een overzicht geven welke AI toepassingen op dit moment in bedrijf of in ontwikkeling zijn bij de politie?
37. Voor welke werkprocessen bouwt de politie op dit moment aan AI-toepassingen binnen de politieorganisatie?
42. Bij welke van haar taken wordt er reeds AI toegepast door de politie?

Antwoord vragen 17, 37 en 42: Er zijn momenteel nog slechts een beperkt aantal toepassingen van AI in de zin dat systemen intelligent gedrag vertonen en in meer of mindere mate zelfstandig kunnen leren en acties kunnen ondernemen. Er zijn bijvoorbeeld toepassingen die het werk van politiemedewerkers vergemakkelijken maar die geen aanmerkelijke gevolgen voor burgers hebben. Daarbij kan gedacht worden aan het doorzoeken van in beslaggenomen gegevensdrager op afbeeldingen met beeldherkenning om een bepaald object te vinden. Dit versnelt het werk van rechercheurs en maakt het werk effectiever en efficiënter. De uitkomst van de analyse heeft dezelfde gevolgen voor burgers als wanneer deze analyse handmatig was uitgevoerd, en is bovendien slechts een van de vele aanwijzingen in het totale onderzoek. Een ander voorbeeld is de in mijn brief van 3 december 2019 reeds genoemde keuzehulp die wordt ingezet bij meldingen van internetoplichting.

Op de meeste terreinen wordt op dit moment dus nog geen gebruik gemaakt van AI. AI is een technologisch hulpmiddel waarmee diverse werkprocessen beter, sneller of effectiever kunnen worden uitgevoerd. AI biedt daarom voor de toekomst kansen indien het breder wordt toegepast om politiemedewerkers te ondersteunen bij hun werk. In potentie kan AI uiteindelijk bij alle politieprocessen worden ingezet.

Er wordt op dit moment vooral gewerkt aan naar de ontwikkeling van concepten en toepassingen die meerwaarde hebben in het ondersteunen van politiemensen in hun werk. Het gaat bijvoorbeeld om:

- spraak naar tekst (bv. mutaties via spraak opnemen in het systeem om bureauwerk te verminderen)
- beeldherkenning (bv. zoeken van afbeeldingen in grote bestanden)
- (natuurlijke) tekstanalyse (bv. keuzehulp internetoplichting)
- dialoogsystemen (bv. chatbot voor dienstverlening)
- explainable AI (bv. uitlegbare beeldherkenning)

43. Zijn er terreinen binnen het politiewezen waar nu nog geen gebruik wordt gemaakt van AI, maar die wel kansen bieden voor het gebruik van AI?
Antwoord op vraag 43: Echte AI toepassingen in de zin dat systemen intelligent gedrag vertonen en in meer of mindere mate zelfstandig kunnen leren en acties kunnen ondernemen komen binnen de politie nog beperkt voor. Op de meeste terreinen wordt op dit moment nog geen gebruik gemaakt van AI. AI is een technologisch hulpmiddel waarmee diverse werkprocessen beter, sneller of effectiever kunnen worden uitgevoerd. AI biedt daarom voor de toekomst kansen

indien het breder wordt toegepast om politiemedewerkers te ondersteunen bij hun werk. In potentie kan AI uiteindelijk bij alle politieprocessen worden ingezet.

44. Kan meer duidelijkheid worden gegeven over hoe wordt bepaald welke taken zullen worden uitgevoerd door de politie met hulp van AI?

Antwoord op vraag 44: Op dit moment worden AI toepassingen vooral op kleine schaal ontwikkeld en toegepast dicht bij de uitvoering, waar een operationele behoefte ligt en collega's actief meewerken in proeftuinen. Naarmate de conceptontwikkeling verder is gevorderd en toepassingen breder beschikbaar komen, worden hiervoor ook op tactisch en strategisch niveau richtinggevende kaders gegeven, in aansluiting met de bij de politie in ontwikkeling zijnde innovatiestrategie waarin AI als sleuteltechnologie een belangrijke rol heeft.

4. Hoe wordt AI toegepast op het gebied van de financiële misdaad?
Antwoord vraag 4: Zoals reeds gemeld komen echte AI toepassingen in de zin dat systemen intelligent gedrag vertonen en in meer of mindere mate zelfstandig kunnen leren en acties kunnen ondernemen binnen de politie nog zeer beperkt voor. Op dit moment wordt binnen de politie AI nog niet toegepast op het gebied van de financiële misdaad.

46. Bij welke politietaken wordt AI reeds gebruikt in het kader van de aanpak van ondermijning? Is er ruimte binnen dit domein om meer AI te gebruiken? Hoe wordt dit getoetst?

Antwoord vraag 46:
Er is ruimte om binnen alle politieprocessen te onderzoeken in hoeverre AI toepassingen meerwaarde hebben en hiermee te experimenteren. Ook binnen de aanpak van ondermijning. Hiervoor gelden dan dezelfde uitgangspunten zoals aangegeven in de brief. In het nieuw in te richten MIT team worden enkele Data Scientists opgenomen om de aanpak van ondermijning te versterken.

8. Wordt overwogen of al geëxperimenteerd met het gebruiken van het Criminaliteits Anticipatie Systeem (CAS) voor misdrijven die buiten de categorie van High Impact Crimes vallen?

Antwoord vraag 8: Sinds enkele jaren wordt in Nederland de term 'high impact crimes' (HIC) gebruikt om delicten aan te duiden die een grote impact op het slachtoffer, diens directe omgeving en het veiligheidsgevoel in de maatschappij hebben. Onder de klassieke HIC-delicten worden (gewelddadige) vermogensdelicten geschaard, om precies te zijn woninginbraak, straatroof en overvallen. Het Criminaliteits Anticipatie Systeem (CAS) wordt gebruikt ter voorkoming van diefstal/inbraak woning, straatroof, diefstal van personenauto's of uit personenauto's, zakkenrollerij, diefstal van snor/brom/fiets, diefstal uit bedrijf of kantoor, overlast jeugd, vernieling, openbare schennis der eerbaarheid en verdovende middelen / drugshandel. De aanwezigheid van de politie op plekken waar een verhoogde kans is op deze en andere soorten delicten kan bijdragen aan de voorkoming ervan. CAS wordt dus al gebruikt voor high impact crimes maar is niet beperkt tot deze categorie. Een regionale eenheid binnen het politiekorps kan op basis van lokaal gedefinieerde speerpunten en voldoende beschikbare data, een gewenste categorie in CAS opnemen. Wanneer er geen duidelijk patroon wordt gevonden, wordt dit niet op de kaart getoond. Daarnaast wordt de uitkomst van CAS altijd door een analist beoordeeld en al dan niet

meegenomen in het geheel van de informatiepositie. Overigens hecht ik er aan om op te merken dat het CAS geen AI bevat.

30. In hoeverre wordt AI nu al toegepast door de politie als het gaat om het beoordelen van meldingen van oplichting? Is hetgeen te zien is in het tv-programma TROS Radar maatgevend voor hoe AI functioneert? Klopt het dat meer dan de helft van de opgelichte mensen van de politie het advies krijgt om een advocaat in de arm te nemen, en dat de politie de aangifte niet wil opnemen, omdat het civiele zaken zou betreffen en geen strafrechtelijke zaken? In hoeverre kan AI werken, als de menselijke beoordeling van oplichtingszaken nu al niet werkt? Wordt de toepassing van AI dan niet een foutenfestival in het kwadraat?

Antwoord vraag 30: AI wordt ingezet voor het adviseren van burgers over meldingen van internetoplichting. Internetoplichting is een vorm van oplichting die de laatste jaren is opgekomen en waar in toenemende mate melding van wordt gedaan. De grens tussen wanneer er sprake is van oplichting (misdrijf in het Wetboek van Strafvordering) en wanneer er sprake is van bijvoorbeeld een wanprestatie (die moet worden opgelost via het civiel recht) is niet altijd even duidelijk voor burgers. Dat zorgde er voor dat er veel meldingen werden gedaan van feiten die niet strafbaar zijn, namelijk een wanprestatie. De keuzehulp geeft burgers die melding willen doen van internetoplichting een advies over welke route waarschijnlijk het meest kansrijk is. Als het voorval waarover de burger melding wil doen mogelijk een strafbaar feit betreft dan is melding (en aangifte) de geëigende weg. Als het bijvoorbeeld gaat om een wanprestatie dan is de route via de civiele rechter de juiste route. Dit is dan het advies dat de burger krijgt. Het staat de burger altijd vrij om bij de politie melding of aangifte te doen. Het gegeven advies is dus geen dwingend advies en kan door de melder worden genegeerd. Op elk moment gedurende de interactie met het systeem kan deze interactie afgebroken worden en de aangifte direct worden gedaan, zonder enig gevolg voor de behandeling van de aangifte.

Het blijft wel zo dat de politie alleen een melding in behandeling neemt als er ook daadwerkelijk sprake is van een mogelijk strafbaar feit. De politie heeft niet tot taak om civielrechtelijke geschillen op te lossen.

38. Hoe wordt bepaald of de toepassing van AI-technologie daadwerkelijk leidt tot een reductie van administratieve lasten?
Antwoord vraag 38: Het verlagen van administratieve lasten is niet altijd het doel van de toepassing van AI. Naast vermindering van administratieve lasten kan AI ook leiden tot andere effecten zoals sneller, effectiever of efficiënter optreden of een zorgvuldigere analyse. Om te bepalen of nieuwe toepassingen van AI effect hebben, vindt voordat van bredere uitrol sprake is, een beoordeling van het effect plaats.

60. In hoeverre bestaat er oefenruimte voor de politie om te kunnen experimenteren met AI?

Antwoord vraag 60: De politie kan, binnen de grenzen van de wet, experimenteren met AI. Daarvoor kan bijvoorbeeld een proefomgeving worden gebruikt. Een mogelijke drempel is dat in het kader van de gegevensbescherming in een ontwikkelomgeving gewerkt moet worden met dummy-data of synthetische

data, terwijl geavanceerde AI systemen beter werken als ze getraind worden met operationele data. Dit heeft mijn aandacht bij de herziening van de Wpg.

64. Op welke manier zijn de informatie-extractie, de juridische toetsing en de optimalisatie van de vraagvolgorde uit elkaar te houden? Deze hangen toch allemaal met elkaar samen?

Antwoord vraag 64: De diverse taken zijn opgenomen in verschillende modules, die los van elkaar informatie-extractie, juridische toetsing en de keuze van de volgende vraag voor hun rekening nemen. De verbanden tussen deze modules zijn hoofdzakelijk op ontwerpniveau van belang. De juridische toetsing wordt niet beïnvloed door de manier waarop observaties gemaakt worden, of hoe de vergaring van de missende informatie geoptimaliseerd wordt. En hoewel de juridische toetsing bepaalt welke observaties bekeken worden en kijkt welke informatie nog mist, heeft de toetsing zelf geen invloed op hoe deze observaties gemaakt worden of op bepaling van de vraagvolgorde. Door de modules op deze wijze van elkaar te scheiden blijft elke afzonderlijke stap beheers- en toetsbaar en wordt het geheel meer transparant.

Juridische en ethische kaders

5. Op welke wijze wordt de naleving van het legaliteitsbeginsel bij de inzet van AI door de politie gegarandeerd?

Antwoord vraag 5: AI wordt door de politie ontwikkeld en toegepast als technisch hulpmiddel voor taken waar zij reeds op grond van de wet toe bevoegd is. Zoals voor alle vormen van taakuitoefening door de politie geldt, dient er, conform het legaliteitsbeginsel, altijd een wettelijke grondslag te zijn voor het handelen van de politie. Dat geldt dus ook voor taakuitvoering die wordt ondersteund door artificiële intelligentie.

20. Kent u de bijlage 'Toelichting op typen data-analyses waarvoor wettelijke waarborgen zullen gaan gelden' (bijlage 2 bij Kamerstuk 26643, nr. 641, d.d. 8 oktober 2019)? Wat betekent het in deze bijlage geformuleerde voornemen om tot wettelijke waarborgen voor twee typen data-analyse te komen, profilering en gebiedsgebonden analyse, voor de ontwikkeling van AI binnen de politie? Behoren niet juist deze twee typen data-analyse tot de kern van het politiewerk en dreigt hiermee niet een rem te ontstaan op innovaties?

Antwoord vraag 20: Het voornemen om mogelijk tot wettelijke waarborgen te komen heeft tot doel om bepaalde risico's die zich voor kunnen doen bij data-analyse zoals profilering en gebiedsgebonden analyse tot een minimum te beperken. Er is geen sprake van een potentieel verbod, dus een voornemen tot wettelijke waarborgen betekent niet dat het onmogelijk wordt om van deze typen data analyse gebruik te maken.

Zowel ikzelf, alsook de korpschef, vinden het heel belangrijk dat innovatie altijd gebeurt met voldoende zorgvuldigheid en oog voor eventuele risico's. Wettelijke waarborgen en richtlijnen zoals de bij voornoemde brief opgenomen 'richtlijnen voor het toepassen van algoritmen door overheden' van de minister voor Rechtsbescherming, kunnen er daarbij voor zorgen dat er een duidelijk kader

geldt waarbinnen geëxperimenteerd, geïnnoveerd en gewerkt kan worden. Daarmee kunnen dergelijke waarborgen zelfs bijdragen aan innovatie.

21. Is er, gezien het feit dat uit de brief duidelijk wordt dat ethiek in de ontwikkeling van AI toepassingen binnen de politie een belangrijke rol speelt, sprake van een politie-specifiek ethisch document of een ethische guideline aan de hand waarvan AI toepassingen worden ontwikkeld en getoetst? Welk houvast hebben ontwikkelaars, leidinggevenden, interne en externe toezichthouders bij het beoordelen en toetsen van de voortgang van projecten? Wordt hier de hulp ingeroepen van externe deskundigen?

Antwoord vraag 21:
Op het gebied van ethische richtlijnen zoekt de politie samenwerking met diverse partners, zoals het ECP, platform voor Informatiesamenleving en is de politie aangesloten bij de Nederlandse AI coalitie. De politie neemt verder deel aan de pilot "*Ethics Guidelines of Trustworthy AI*" van de High Level Expert Group on AI van de Europese Commissie en aan het Transparantielab van het Ministerie van Binnenlandse zaken.

Vanwege het belang van privacy en ethiek voor het politiewerk is binnen de politie recent een specifieke portefeuillehouder privacy en ethiek aangesteld. Daarnaast is er in opdracht van de politie onderzoek gedaan naar ethische afwegingen voor verantwoordelijk gebruik van AI binnen de politie.[2]

53. Wordt er een juridisch en ethisch kader opgesteld voor de toepassing van AI?

Antwoord vraag 53: Voor de toepassing van AI gelden de gebruikelijke wettelijke kaders, waaronder de Wet politiegegevens (Wpg). De Minister voor Rechtsbescherming heeft daarnaast op 18 oktober 2019 'richtlijnen voor het toepassen van algoritmen door overheden' naar uw Kamer gestuurd. Zoals reeds aangekondigd in zijn brief van 18 oktober 2019 wil het kabinet toewerken naar waarborgen die in wetgeving kunnen worden opgenomen.[3] Deze richtlijnen komen bij de politie terug in het interne Kwaliteitskader Big Data waarin aandacht wordt besteed aan de vraag wie verantwoordelijk is voor de AI-toepassing en op basis van welke wettelijke en juridische grondslagen en ethische toetsing de ontwikkeling wordt gedaan. Daarnaast is er in opdracht van de politie onderzoek gedaan naar ethische afwegingen voor verantwoordelijk gebruik van AI binnen de politie.[4]

57. Wordt bij elke nieuwe toepassing van een AI techniek ook een nieuwe juridische en ethische toetsing gedaan?

---

[2] Dechesne, F., Dignum, V., Zardiashvili, L., & Bieger, L. J. (2019). AI & Ethics at the Police, via https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/artificiele-intelligentie-en-ethiek-bij-de-politie/ai-and-ethics-at-the-police-towards-responsible-use-of-artificial-intelligence-at-the-dutch-police-2019..pdf

[3] Kamerstukken II, 2019/20, 26643, 641.

[4] Dechesne, F., Dignum, V., Zardiashvili, L., & Bieger, L. J. (2019). AI & Ethics at the Police, via https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/artificiele-intelligentie-en-ethiek-bij-de-politie/ai-and-ethics-at-the-police-towards-responsible-use-of-artificial-intelligence-at-the-dutch-police-2019..pdf

Antwoord op vraag 57
De politie dient zich bij de inzet van technische toepassingen, waaronder AI, te houden aan de bestaande normering. Indien van toepassing, wordt op grond van de WPG het wettelijk kader via het instrument van de DPIA in kaart gebracht, waarbij rekening wordt gehouden met de stand van de techniek, uitvoeringskosten, alsook met de aard, omvang, de context, verwerkingsdoeleinden en risico's voor de rechten en vrijheden van personen worden meegewogen.

Bij elke nieuwe ontwikkeling van het Nationaal Politielab AI worden juridisch, ethische en sociale aspecten mee genomen en beoordeeld vooraf, gedurende de ontwikkeling en tijdens de experimenten. Hierbij wordt tevens gebruik gemaakt van het Kwaliteitskader Big Data. De werkwijze van de politie op het gebied van het ontwikkelen en toepassen van algoritmes en AI wordt de komende tijd conform dit kwaliteitskader ingericht.

59. Welke ethische standaarden zijn er uit het onderzoek van het Nationaal Politielab AI gekomen en welke daarvan worden ook echt toegepast?

Antwoord op vraag 59:
Uit onderzoek van het Nationaal Politielab AI, onder andere in een whitepaper[5], komt naar voren dat, om AI verantwoord in te zetten algemene en abstracte principes moeten worden omgezet naar concrete vereisten op technisch, individueel en maatschappelijk gebied. De verkregen inzichten worden op dit moment geanalyseerd en waar nodig opgenomen in de verdere planvorming met betrekking tot de AI ontwikkeling binnen de politie.

<u>Gegevensbescherming en rechten van burgers</u>

19. Hoe gaat de politie, specifiek als het gaat om AI toepassingen, om met de 'Richtlijnen inzake publieksvoorlichting over data-analyses' (bijlage 1.2 bij Kamerstuk 26643, nr. 641, d.d. 8 oktober 2019)? Klopt het dat deze richtlijn als uitgangspunt heeft een actieve informatieverstrekking vanuit de betrokken overheidsorganisatie aan het publiek, inclusief expliciete vermelding in een privacystatement op de eigen website van in dit geval de politie? Deelt u de observatie dat dergelijke informatie op dit moment niet te vinden is op de website van de politie? Gaat hier op korte termijn verandering in komen?

Antwoord vraag 19: De 'Richtlijnen inzake publieksvoorlichting over data-analyses' richten zich, zoals in deze richtlijnen zelf ook is aangegeven, voornamelijk op gegevensverwerkingen die vallen onder de reikwijdte van de Algemene Verordening Gegevensbescherming (AVG). De verwerking van persoonsgegevens door de politie valt echter grotendeels buiten de werking van deze verordening maar onder de werking van de Wet Politiegegevens, waarin de

[5] Dechesne, F., Dignum, V., Zardiashvili, L., & Bieger, L. J. (2019). AI & Ethics at the Police, via https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/artificiele-intelligentie-en-ethiek-bij-de-politie/ai-and-ethics-at-the-police-towards-responsible-use-of-artificial-intelligence-at-the-dutch-police-2019..pdf

Richtlijn (EU) 2016/680 inzake gegevensbescherming opsporing en vervolging is geïmplementeerd. Het feit dat voor de uitvoering van de politietaak een apart gegevensbeschermingsregime geldt, is gelegen in de bijzondere positie waarin de politie als handhavings- en opsporingsinstantie verkeert. In aanvulling op hetgeen ik eerder aan uw kamer heb gemeld (bijlage bij Kamerstuk 26643, nr. 426, d.d. 11 november 2016) geldt dat het voor de politie in voorkomende gevallen noodzakelijk is om (delen van) de gegevensverwerking niet inzichtelijk te maken. Dit kan nodig zijn om te voorkomen dat personen zich kunnen onttrekken aan een effectieve taakuitoefening door de politie. Inzicht in de gebruikte analysemethode kan immers aanleiding zijn om het gedrag bewust zodanig aan te passen dat men in de gegevensanalyse buiten zicht blijft. Daarnaast kan geheimhouding nodig zijn omdat inzicht in de gegevensverwerking raakt aan de nationale veiligheid.

Dat neemt niet weg dat de politie zoveel als mogelijk zal aansluiten bij de onderhavige richtlijnen en dus ook in dit kader uitvoering zal geven aan de actieve informatieplicht die op de politie rust. Dit zal op termijn, en met de voornoemde kaders en beperkingen, leiden tot aanvulling van de informatie op de website van de politie.

Bijvoorbeeld bij de keuzehulp internetoplichting worden de burgers actief geïnformeerd over het feit dat de keuzehulp geautomatiseerd is, en er dus geen sprake is van contact met een politiemedewerker tijdens het invullen van de keuzehulp.

77. Kunnen burgers inzien welke data van hen gebruikt worden door AI bij de politie?

Antwoord vraag 77: De politie verwerkt bij de toepassing van AI politiegegevens zoals bedoeld in de Wpg. Een ieder heeft, op grond van artikel 25 Wpg, recht op inzage in de gegevens die over hem worden verwerkt door de politie, dus ook die door AI toepassingen worden gebruikt. Vanwege de in het antwoord op vraag 19 genoemde bijzondere positie van de politie geeft de Wpg (art. 27) hierop wel een aantal uitzonderingen Het recht op inzage wordt bijvoorbeeld beperkt ter bescherming van de openbare en nationale veiligheid of indien inzage nadelige gevolgen heeft voor het opsporingsonderzoek. Deze uitzonderingsgronden wijken enigszins af van die van de AVG.

55. Hoe zijn op dit moment de rechten van inwoners beschermd die deel uitmaken van een experiment, proeftuin of pilot?
56. Worden inwoners adequaat geïnformeerd over de gevolgen voor hun privacy-rechten? Zo ja, hoe worden zij geïnformeerd? Zo nee, waarom niet?

Antwoord vragen 55 en 56: De politie dient in het geval van (operationele) experimenten, proeftuinen of pilots binnen hetzelfde wettelijke kader te opereren als wanneer er geen sprake is van een experiment. De rechten van inwoners die deel uitmaken van een experiment, proeftuin of pilot zijn dus op eenzelfde wijze beschermd als normaal gesproken. Het wettelijk kader dat van toepassing is op de vorm van de toepassing (is er bijvoorbeeld sprake van opsporing of van preventie van criminaliteit) bepaalt mede in welke mate de politie gegevens kan delen.

Zoals reeds opgemerkt in mijn antwoord vraag 19 is het niet altijd mogelijk om op eenzelfde wijze transparant te zijn over het gebruik van data door de politie als bij AVG data. Zoals gemeld in mijn antwoord op vraag 78 hebben burgers, op grond van artikel 25 Wpg, recht op inzage in de gegevens die over hen worden verwerkt door de politie, dus ook die door AI toepassingen worden gebruikt. Vanwege de in het antwoord op vraag 19 genoemde bijzondere positie van de politie geeft de Wpg (art. 27) hierop wel een aantal uitzonderingen die enigszins afwijken van die van de AVG.

Bijvoorbeeld bij de keuzehulp internetoplichting worden de burgers actief geïnformeerd over het feit dat de keuzehulp geautomatiseerd is, en er dus geen sprake is van contact met een politiemedewerker tijdens het invullen van de keuzehulp.

Data

18. Als in de brief gesproken wordt over (big) data, wordt dan enkel gedoeld op databestanden die de politie zelf aanhoudt en grotendeels vallen onder de werking van de Wet politiegegevens, of wordt dan ook gedoeld op publieke databestanden, hetzij van andere overheden hetzij publieke data die via internet benaderbaar zijn? Klopt het beeld dat de politie nu al gebruik maakt van data-mining? Kunt u, zo nodig vertrouwelijk, een overzicht geven van de toepassing van data-mining binnen en door de politie?

Antwoord vraag 18: Alle persoonsgegevens die de politie verkrijgt in het kader van de uitoefening van haar taak (minus de bestuursrechtelijke taken) zijn politiegegevens. Dus ook gegevens verstrekt uit samenwerkingsverbanden of door samenwerkingspartners en de gegevens die de politie verzamelt bijvoorbeeld tijdens een opsporingsonderzoek uit open bronnen zoals Kadastergegevens en voor internetgegevens. Ten aanzien van dat laatste spreken we conform het advies van de Commissie Koops liever van publiektoegankelijke en gesloten bronnen dan van open bronnen.

De politie maakt al langere tijd gebruik van datamining, in de afgelopen jaren is dit getransformeerd naar het interdisciplinaire vakgebied data science, Waar datamining zich doorgaans meer richt op beschrijvende analyses, biedt data science ook mogelijkheden om predictief en prescriptief acties te kunnen ondernemen. Data science kan zowel voor de kerntaken als voor specifieke taken van de politie worden ingezet. Daarbij kan o.a. worden gedacht aan de ontwikkeling van (risicotaxatie)modellen, het herkennen van patronen of anomaliteiten, tekst- en beeldherkenning, het inzichtelijk maken van criminele netwerken en markten, het doorgronden van grote datasets maar ook om de bedrijfsvoering efficiënter te maken.

78. Worden er enkel data gebruikt van burgers die in contact zijn geweest met justitie, of gaan de data ook over 'reguliere' burgers?

Antwoord vraag 78: De politie verwerkt voor haar taak persoonsgegevens (zogenaamde politiegegevens). Het gaat hier niet enkel om gegevens over de verdachte in een opsporingsonderzoek maar ook over gegevens van personen die een aangifte doen, slachtoffer zijn, getuige zijn van een delict, personen die

staande worden gehouden of die bijvoorbeeld een melding doen in het kader van de openbare orde handhaving. Ook deze politiegegevens zijn nodig om de taak goed uit te oefenen. De omvang, aard en indringendheid van deze politiegegevens is vanzelfsprekend niet voor alle categorieën betrokkenen gelijk. De registratie van politiegegevens bij een overlastmelding zijn veel beperkter dan de politiegegevens die worden verzameld over een verdachte van een ernstig strafbaar feit.

Bias

24. In hoeverre zijn beslisbomen daadwerkelijk altijd neutraal? Kunnen beslisbomen impliciete aannames bevatten, die helemaal niet neutraal zijn?
25. Klopt het dat beslisbomen niet altijd neutraal zijn, maar vaak impliciete aannames bevatten en in een bepaalde richting kunnen tenderen? Herinnert u zich de beslisboom van het ministerie van Economische Zaken (EZ) van maart 2000, over marktwerking en privatisering? Klopt het dat het invullen van deze beslisboom in bijna alle gevallen leidde tot het advies om over te gaan tot privatisering en/of marktwerking? Klopt het dat dit een weerspiegeling was van het toenmalige marktdenken op het ministerie van EZ? Klopt het dat beslisbomen helemaal niet neutraal zijn? Waarom is het dan "helder dat hier geen bijzonder beleid voor nodig is"?

Antwoord vragen 24 en 25: Het klopt dat algoritmes (waaronder beslisbomen) en dus ook AI bias of fouten kunnen bevatten of gebaseerd kunnen zijn op (onterechte) aannames. Dat geldt evenzo voor beslissingen die alleen door mensen worden gemaakt. Als het gaat om bias en aannames in eenvoudige algoritmes (zoals vergelijkingen in een Excel sheet) en in beslisbomen geldt dat deze eenvoudiger met het blote oog of via standaard controle stappen te herkennen zijn, omdat de redeneerlijn in een beslisboom stap voor stap te volgen is. Voor meer complexe algoritmes is het ontdekken van een aanname of bias minder eenvoudig met het blote oog te doen. Dit onderscheid tussen de complexiteit van algoritmes maakt dat er een grotere noodzaak gevoeld wordt om met specifiek beleid te komen voor bias in complexe algoritmes, dan voor een eenvoudiger algoritme.

9. Hoe wordt de noodzaak om bias te voorkomen geadresseerd bij het gebruik van kunstmatige intelligentie binnen de politie?
26. Als de privatiseringsbeslisboom van het ministerie van EZ van 2000 niet neutraal was, kunt u dan uitsluiten dat de algoritmes, beslisbomen en risicotaxaties van de politie evenmin neutraal zijn? Op welke manier kan een dergelijke vooringenomenheid of bias uitgesloten worden?
47. Op welke wijze zet de overheid zich in om actief discriminatie te voorkomen in de AI systemen van politie?

Antwoord vragen 9, 26 en 47: De politie werkt bij het ontwikkelen van AI volgens het principe '*ethics by design*'. Dit betekent dat bepaalde waarden bij de ontwikkeling van een AI systeem al worden meegenomen in het ontwerp. Er wordt dus reeds bij de ontwikkeling van algoritmes en AI aandacht besteed aan de mogelijkheid dat er bias in de trainingsdata zit die effect kan hebben op de werking van het algoritme. Indien bekend wordt dat er bias in de data aanwezig is

zijn er mogelijkheden om te compenseren voor deze eventuele bias in de data die worden toegepast. Dit compenseren kan zowel in de data zelf als door hier in het algoritme rekening mee te houden.

28. Waar ligt de grens tussen intelligent gedrag van systemen en systemen die eerder invoerde aannames uitvoeren? Klopt het dat deze grens niet scherp te trekken is, en er een groot grijs gebied is tussen intelligent gedrag en ingevoerd of geïnduceerd gedrag? Hoe beoordeelt u in dit kader de uitspraak "Het is gevaarlijk als alleen witte mannen algoritmes maken" van dhr. Frans Muller (Forum, "Frans Muller: bedrijven moeten ethisch omgaan met data")?
29. Zijn algoritmes en AI niet te vergelijken met een zichzelf bevestigende grabbelton, onder het motto "wat je er in stopt, haal je er ook weer uit"? Lopen AI toepassingen daarmee niet het risico een echoput te worden van vooroordelen en/of vooringenomenheid, waaronder ook racisme en discriminatie? Hoe kunt u dit uitsluiten? Wie of welke instantie toetst hierop? Is er een toetsing denkbaar die analoog is aan de toetsing uit de Sleepwet van 2018, met een commissie die noodzaak en proportionaliteit beoordeelt? Zo nee, op welke manier is dan een toetsing mogelijk?

Antwoord vragen 28 en 29: Ik beaam dat het belangrijk is om rekening te houden met diversiteit en eventuele bias in de ontwikkeling en toepassing van AI en algoritmes. Een belangrijk aandachtspunt in data science is 'garbage in, garbage out', wat grofweg betekent dat het gebruik van slechte, vervuilde of onvolledige data ervoor zorgt dat ook de output slecht, vervuild of onvolledig kan zijn. Tegelijk zijn er in de data science diverse manieren om hiervoor te compenseren.

Zoals ik in mijn brief aan uw Kamer heb aangegeven moet gemotiveerd worden waarom en welke data er wordt gebruikt en er moet gekeken worden naar de kwaliteit van de gebruikte data. Dit is mede de professionele taak van de data scientists binnen de politie om dit voorkomen. Het is daarbij uitermate belangrijk dat er bij de ontwikkeling van AI toepassingen voldoende duidelijk is wat de eventuele gebreken in een dataset zijn, zodat hiervoor kan worden gecompenseerd. Daarbij komt bijvoorbeeld ook de vraag naar voren of kan worden volstaan met geanonimiseerde data en op welke wijze privacy by design wordt toegepast. Overigens is het zo dat de politie zelf ook geen baat heeft bij een slecht werkende AI toepassing. Er wordt daarom doorlopend gecontroleerd of de modellen van voldoende kwaliteit blijven.

Op 20 december 2019 hebben de Ministers van Binnenlandse Zaken en Koninkrijksrelaties en de Minister voor Rechtsbescherming het onderzoek "Toezicht op gebruik van algoritmen door de overheid" aan uw Kamer aangeboden.[6] De onderzoekers constateerden op basis van een analyse van wet- en regelgeving geen juridische lacune in de toezichtstaken. In de begeleidende brief is uw Kamer een reactie op de aanbevelingen in dit rapport toegezegd. In de beantwoording van de voorliggende vragen wil ik niet vooruitlopen op die reactie.

48. Op welke wijze toetst de politie de afwezigheid van discriminatie in AI systemen?

[6] Kamerstukken II, 2019/20, 26643, 657.

71. Op welke wijze wordt discriminatie als onbedoeld gevolg van de toepassing van AI voorkomen?
81. Welke mechanismen worden in werking gesteld op het moment dat discriminatie wordt opgemerkt in het ontwerp, gebruik of de uitkomst van AI systemen?

Antwoord vragen 48, 71 en 81: Ongewenste effecten van algoritmen, onder andere mogelijk door bias, vereisen een brede aanpak. De basis is professionaliteit bij de ontwikkelaars als het gaat om (datagedreven) AI toepassingen. Het kennisniveau van hen moet daarom bij blijven bij de laatste ontwikkelingen. De politie organiseert daarom bijv. colloquia en meetups, en geeft experts de ruimte om te blijven leren. Sommige effecten openbaren zich pas in het werkproces en zijn niet inherent aan de technologie. Daarom is het belangrijk dat eindgebruikers in contact staan met de ontwikkelaars. Hierdoor kan de technologie waar mogelijk aangepast worden of juist de eindgebruiker waar nodig beterebegrip krijgen van het middel.

Bij de inzet van algoritmes voor het politiewerk wordt doorlopend geëvalueerd en gecontroleerd of de modellen van voldoende kwaliteit blijven. Daarbij is nadrukkelijk aandacht voor bias, zowel in de data als in de modellen, en ook in de uitkomst.

De monitoring op de toepassing van AI (ook wel intern toezicht) wordt ingevuld door middel van het 'three lines of defence'-model waarbij de eerste lijn wordt vervuld vanuit het lijnmanagement, de tweede lijn door de korpscontroller en de derde lijn door de afdeling Concernaudit.[7]
Indien bekend wordt dat er bias in de data aanwezig is zijn er mogelijkheden om te compenseren voor deze eventuele bias in de data die worden toegepast. Dit compenseren kan zowel in de data zelf als door hier in het algoritme rekening mee te houden.

82. Welke stappen zijn er ondernomen om te voorkomen dat de AI systemen van de politie voornamelijk worden gericht op gemarginaliseerde groepen in de samenleving? Hoe vindt de evaluatie van dit soort waarborgen plaats?

Antwoord vraag 82: Voor al het politiewerk geldt dat het niet de bedoeling is dat dit specifiek wordt gericht op bepaalde gemarginaliseerde groepen in de samenleving. Dat geldt dus ook voor het toepassen van AI. Bij de politie wordt er veel aandacht besteed aan het voorkomen van discriminatie en etnische profilering in den brede. Ik heb u hierover onder andere geïnformeerd in mijn brief van 12 december 2019.[8] Deze aanpak geldt ook voor het gebruiken van het ondersteunende middel AI.
Specifiek voor technische hulpmiddelen zoals AI geldt dat er daarnaast ook al tijdens de ontwikkeling van de toepassing aandacht worden besteed aan het voorkomen van bias, door te ontwikkelen volgens het principe van 'ethics by design'.

Bij de inzet van algoritmes voor het politiewerk wordt doorlopend geëvalueerd en gecontroleerd of de modellen van voldoende kwaliteit blijven. Daarbij is

[7] Kamerstukken II 2018/19, 29628, 835 (blg 1).
[8] Kamerstukken II, 2019/20, 30950, 183.

nadrukkelijk aandacht voor bias, zowel in de data als in de modellen, en ook in de uitkomst. Zie hiervoor ook het antwoord op vragen 48, 71 en 81.

SyRI

16. Wat voor invloed heeft de SyRi-uitspraak op de manier waarop de politie AI toepast?

Antwoord vraag 16: De uitspraak omtrent SyRI is pas zeer recentelijk gepubliceerd. Zowel mijn departement als de politie zullen de uitspraak zeer zorgvuldig bestuderen, en bezien op eventuele consequenties voor de toepassing van AI door de politie.

Verantwoording en toezicht

1. Wordt aan de Tweede Kamer inzicht verschaft over de toepassing van artificiële intelligentie (AI) door de politie en zo ja, hoe en wanneer?
69. In hoeverre wordt de Tweede Kamer geïnformeerd over de uitkomsten van onderzoek op het gebied van AI bij de politie?

Antwoord vraag 1, 69: Mijn kamerbrief van 3 december 2019 over AI bij de politie had tot doel het verschaffen van transparantie over AI bij de politie. De korpschef en ik blijven transparant over de toepassing van AI bij de politie. Indien er nieuwe ontwikkelingen zijn op het gebied van AI bij de politie zal ik uw Kamer daarover informeren.

Het wetenschappelijk onderzoek dat wordt uitgevoerd in het Nationaal Politielab AI, door de wetenschappers die daaraan zijn verbonden, wordt gepubliceerd in wetenschappelijke tijdschriften. Daarmee is het openbaar en ook voor Kamerleden te bestuderen indien gewenst. In de bijlage van mijn brief van 3 december jl. heb ik een recent overzicht van het uitgevoerde onderzoek bijgevoegd.

23. Welke rol speelt de functionaris gegevensbescherming op korpsniveau in het toezicht op de ontwikkeling van AI toepassingen?

Antwoord vraag 23 De Functionaris voor Gegevensbescherming (FG) is onafhankelijk in zijn advies en toezicht op de naleving van de Wpg en AVG binnen de Politie. De FG kan dus eigenstandig proactief onderzoek doen en toezicht houden op de toepassing van AI binnen de politie, daar waar het de verwerking van persoonsgegevens betreft. Daarbij wordt de FG geïnformeerd wanneer een Gegevensbeschermingseffectbeoordeling (GEB) is opgesteld en kan hij langs die lijn adviseren en toezicht houden.

72. Waarom stelt u in de brief dat "de politie zijn eigen strengste criticaster is"? Hoe relateert u deze uitspraak aan het functioneren van de politie in de Arnhemse moordzaak, niet alleen in 1998-1999, maar ook nu? Is er in de afgelopen 22 jaar enige vorm van zelfreflectie geweest? Klopt het dat de politie helemaal niet "zijn eigen strengste criticaster" is maar juist het tegenovergestelde? Namelijk dat door een interne angst- en suppressiecultuur geen enkele zelfreflectie mogelijk is, afgezien van een enkeling die inmiddels met pensioen is?

Antwoord vraag 72:
Het stelsel van toezicht en waarborgen op en binnen de politie heb ik in mijn brief van 5 december 2018[9] toegelicht. Dit start bij intern toezicht binnen de politie,, dat het zogenoemde 'three lines of defence' model volgt, waarbij de eerste lijn wordt vervuld vanuit het lijnmanagement, de tweede lijn door de korpscontroller en de derde lijn door de afdeling Concernaudit. Basis van het toezicht op de politie is dat de politie het interne toezicht streng en goed op orde moet hebben. Dat wordt er bedoeld met de frase dat de politie zijn eigen strengste criticaster is. Dat betekent niet dat er geen fouten gemaakt worden, maar het uitgangspunt moet zijn dat het interne toezicht van de politie een strenge controleur is. Dit vind ik ook bij AI van groot belang. Hoe beter de politie zelf de waarborgen en het interne toezicht op orde heeft, hoe meer zekerheid de burger heeft dat het optreden van de politie legitiem is.

80. Bij wie ligt de verantwoordelijkheid voor monitoring van de toepassing van AI bij politietaken? Zit daar een bepaalde evaluatietermijn aan vast?

Antwoord vraag 80: De monitoring op de toepassing van AI (ook wel intern toezicht) wordt ingevuld door middel van het 'three lines of defence'-model waarbij de eerste lijn wordt vervuld vanuit het lijnmanagement, de tweede lijn door de korpscontroller en de derde lijn door de afdeling Concernaudit.[10] Dit interne toezicht arrangement geldt ook voor toepassingen die gebruik maken van AI.

27. Wie of welke instantie beoordeelt de neutraliteit van algoritmes? Welke rol kunnen de Autoriteit Persoonsgegevens (AP), de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) en de Algemene Rekenkamer daarbij spelen?

Antwoord op vraag 27: Op 20 december 2019 hebben de Ministers van Binnenlandse Zaken en Koninkrijksrelaties en de Minister voor Rechtsbescherming het onderzoek "Toezicht op gebruik van algoritmen door de overheid" aan uw Kamer aangeboden.[11] De onderzoekers constateerden op basis van een analyse van wet- en regelgeving geen juridische lacune in de toezichtstaken. In de begeleidende brief is uw Kamer een reactie op de aanbevelingen in dit rapport toegezegd. In de beantwoording van de voorliggende vragen wil ik niet vooruitlopen op die reactie.

73. Op welke manier houdt de Inspectie voor Justitie en Veiligheid toezicht op het gebruik van AI door de politie? Hoeveel fte heeft de Inspectie daarvoor beschikbaar? Heeft de Inspectie daarvoor de benodigde expertise en welke vorm neemt deze expertise aan?

Antwoord op vraag 73: De Inspectie Justitie en Veiligheid houdt toezicht op de kwaliteit van de taakuitvoering van onder andere de politie. Als bij die taakuitvoering AI wordt ingezet, dan wordt dit betrokken in het toezicht van de

[9] Kamerstukken II 2018/19, 29628, 835 (blg 1).
[10] Kamerstukken II 2018/19, 29628, 835 (blg 1).
[11] Kamerstukken II, 2019/20, 26643, 657.

Inspectie. Het toezicht is dan gericht op de bijdrage van de techniek aan de uitvoering. Daarnaast wordt de legitimiteit van de inzet van AI bezien en risico's op ethisch ongewenste effecten, zoals uitsluiting of discriminatie van personen. Tot op heden heeft de Inspectie nog geen onderzoek verricht naar AI. Echter door de toename van het gebruik van AI en gerelateerde technieken zal dit een meer prominente rol in het toezicht gaan innemen. De Inspectie heeft de ambitie om haar toezicht de komende jaren nadrukkelijk te richten op deze technieken.

De totale capaciteit van de Inspectie is ca. 100 FTE. De capaciteit voor toezicht wordt zoveel mogelijk risico gestuurd ingezet op het gehele brede toezichtterrein. Het afgelopen jaar is geïnvesteerd in de werving van medewerkers met kennis van het digitale domein en technologieën, zoals AI. Deze capaciteit zal de komende jaren verder toenemen. Daarnaast investeert de Inspectie in het verder versterken van de kennis door expliciet de verbinding te zoeken met de wetenschap.

75. Wie controleert de effecten van een Data Protection impact assessment? Welke rol spelen de AP en de Auditdienst Rijk hierin?

Antwoord op vraag 75: Een Gegevensbeschermingseffectbeoordeling (GEB) wordt voorafgaand aan de start van een hoog risico verwerking opgesteld. Degene die formeel bevoegd is te beslissen of persoonsgegevens worden verwerkt is verantwoordelijk voor het opstellen van de GEB.
De verantwoordelijke wordt hierbij ondersteund door een Privacyfunctionaris bij de Politie. Eventueel kan de Privacyfunctionaris ondersteuning vragen aan de Gegevensautoriteit binnen de Politie. De GEB wordt toegestuurd aan de onafhankelijke Functionaris voor Gegevensbescherming (FG) binnen de Politie en indien nodig wordt de Autoriteit Persoonsgegevens voorafgaand aan de verwerking geraadpleegd. De Auditdienst Rijk heeft geen rol bij een Gegevensbeschermingseffectbeoordeling.

54. Op welke wijze legt de politie verantwoording af over de wijze hoe AI wordt toegepast?

Antwoord vraag 54: De politie moet verantwoording afleggen over de taakuitvoering. Afhankelijk van het type inzet van AI en de benodigde data voor de ontwikkeling van de AI wordt op verschillende wijzen verantwoording afgelegd. De politie voert haar taken uit onder de verantwoordelijkheid van het gezag. Indien het gaat om de toepassing van AI in het kader van de opsporing is het bevoegd gezag het Openbaar Ministerie. Als het gaat om vervolging zal deze vervolgens in het strafproces moeten kunnen uitleggen hoe het opsporingsproces is verlopen. Als daar AI bij is gebruikt moet deze voor de rechter toetsbaar zijn.

Daarnaast legt de korpschef voor het gevoerde beheer en de kwaliteit van de taakuitvoering verantwoording af aan mij. Dat omvat ook de inzet en kwaliteit van AI in het algemeen.

70. Kunt u het interne kwaliteitskader big data dat de politie samen met het OM in het kader van het programma Toekomstbestendig Opsporen en Vervolgen heeft opgesteld zo nodig vertrouwelijk delen met de Tweede Kamer? Hoe verhoudt dit interne kader zich tot de ministerieel vastgelegde `richtlijnen voor het toepassen van algoritmes door

overheden' (bijlage 1 bij Kamerstuk, 26643, nr. 641, dd. 8 oktober 2019)?
76. Welke externe toezichthouders hebben of krijgen inzicht in het interne kwaliteitskader dat door politie en OM wordt gehanteerd?

Antwoord vraag 70 en 76: Het interne Kwaliteitskader Big Data dat de politie samen met het OM in het kader van het programma Toekomstbestendig Opsporen en Vervolgen heeft opgesteld bevindt zich momenteel in de afrondende fase van het doorlopen van het interne proces van goedkeuring binnen de politie. De verwachting is dat dit proces begin april 2020 zal zijn afgerond. Na vaststelling door de korpsleiding kan het Kwaliteitskader met Uw Kamer worden gedeeld. Het Kwaliteitskader is een praktische uitwerking van de Richtlijnen voor het toepassen van algoritmes door overheden. De documenten worden bij de implementatie van het Kwaliteitskader Big Data in de politieorganisatie in samenhang met elkaar bekeken. Zodra het Kwaliteitskader Big Data naar Uw Kamer is verzonden is deze openbaar geworden. Uiteraard hebben ook alle toezichthouders die dat wensen inzicht in het Kwaliteitskader.

Verantwoordelijkheid en foutmarges

14. Wie is er verantwoordelijk voor (eventuele) fouten van AI?

Antwoord vraag 14: De korpschef is ingevolge artikel 27 Politiewet 2012 belast met de leiding en het beheer van de politie. Voor AI is de verantwoordelijkheid niet anders dan voor alle andere technieken en hulpmiddelen die in de taakuitvoering worden gebruikt. Zie verder het antwoord op vraag 54 over de wijze waarop de politie verantwoording aflegt.

50. Hoe ziet de foutmarge van AI eruit? Welke gevolgen kunnen foute beslissingen hebben, die genomen worden op basis van AI?

Antwoord vraag 50: Foutmarges bij AI toepassingen verschillen per toepassing. Het is dus niet mogelijk om in algemeenheid uitspraak te doen over een foutmarge bij AI.

Zoals gemeld in mijn brief van 3 december 2019[12] is aandacht voor foutmarges bij toepassing van AI door de politie zeer belangrijk. Hierbij geldt dat het type proces, de actie maar met name het potentiële gevolg bepalend is voor de acceptatie van de marge van fouten van een AI toepassing. Voor ingrijpende processen zal AI in principe politiemedewerkers slechts ondersteunen in hun taak.

Het uitgangspunt is dat er bij het gebruik van AI altijd sprake is van een menselijke tussenkomst bij het nemen van besluiten ("*human in the loop*"). Desalniettemin is ook bij besluitondersteuning door AI zorgvuldigheid geboden.

51. Kan AI ertoe leiden dat onschuldige verdachten worden aangewezen, zoals bij de Arnhemse moordzaak in 1998? Klopt het dat bij de Arnhemse moorzaak in 1998 sprake was van een discriminerende vooringenomenheid, die heeft geleid tot negen personen die onschuldig zijn veroordeeld tot jarenlange gevangenisstraffen, wat bij één persoon

[12] Kamerstukken II, 2019/20, 26643, 652

heeft geleid tot zelfmoord? Hoe kan het dat de Arnhemse politie op basis van een snipper foutief bewijs in een onvoorstelbare tunnelvisie terecht is gekomen? Kunt u uitsluiten dat juist AI de politie aan foutieve snippers bewijs gaat helpen, dat leidt tot nieuwe tunnelvisies?

Antwoord vraag 51: AI wordt door de politie gebruikt als ondersteuning van het politiewerk. Daarnaast vereisen de Wpg en de richtlijn (EU) 2016/680 inzake gegevensbescherming opsporing en vervolging, dat er voor geautomatiseerde besluitvorming een wettelijke grondslag is. Er is geen sprake van geautomatiseerde besluitvorming bij de politie, noch is dat bij wet voorzien. De vrees dat AI op zichzelf kan leiden tot uitkomsten in strafzaken, of die nu goed of foutief zijn, is dus ongegrond. Tegelijkertijd is bij elk stuk bewijs, bij elke stap binnen de opsporing zorgvuldigheid geboden. Zoals ik in mijn brief reeds aangaf, moet AI bij de politie toetsbaar zijn voor de rechter. Ook als dit slechts ondersteunend is geweest in het proces. Deze waarborg zorgt ervoor dat in een proces de rechter zelf kan beoordelen wat de invloed van AI in de procesgang is geweest.

Opleiding en competenties

2. Hoe staat het met de digitale geletterdheid van de politieagenten in Nederland?
Antwoord vraag 2: De politie investeert op verschillende manieren in de digitale geletterdheid van de organisatie. Sinds de reorganisatie is het aantal digitaal experts in de inrichting toegenomen en het aantal digitaal experts stijgt nog steeds. Zo vindt op dit moment een versterking plaats voor de aanpak van cybercrime met 145 fte. Hiervan is in 2019 ruim 80 fte ingestroomd. De rest volgt in 2020. De cybercrimeteams in de eenheden vormen een vliegwiel voor de aanpak van gedigitaliseerde criminaliteit in de eenheden. Zij leren kennis uit aan de districten, basisteams en aan de intake-en servicemedewerkers. De politie beschikt verder over specialistische teams digitale Opsporing die de opsporing ondersteunen met hun digitale expertise. Verder biedt de politieacademie opleidingen en trainingen aan op het gebied van gedigitaliseerde criminaliteit en opsporing.

45. In hoeverre is de theorie en toepassing van AI verankerd in de politieopleiding?

Antwoord vraag 45: De kennis van Data Science en AI ontwikkeling maakt geen onderdeel uit van de politieopleiding. Data science en AI is een specialisme waarvoor de politie in principe hoger opgeleiden van buiten de organisatie aantrekt die de korte opleiding Politiemedewerker Specifieke Inzet volgen, die speciaal gericht is op zij-instromende specialisten in de Politie-organisatie. Dit specialisme is bovendien kennisintensief. Om continu aan te sluiten bij die laatste ontwikkelingen is het Nationaal politielab AI gestart, waarin in ICAI verband wordt samengewerkt met Universiteiten.
Indien een AI toepassing op dit moment in de praktijk wordt toegepast in pilots worden de betrokken collega's hiervoor opgeleid. Hierbij kan gedachte worden aan het leren wat de uitkomst van de AI betekent en hoe die beoordeeld moet worden. Dit betreft maatwerk, afhankelijk van de toepassing. Indien AI in de toekomst breed ingezet gaat worden ter ondersteuning van werkprocessen is te verwachten dat dit onderdeel wordt van de opleiding.

22. Welke kennis rond AI en big data is aanwezig bij het lijnmanagement, bij de korpscontroller en bij de afdeling Concernaudit? Kunt u deze vraag specifiek beantwoorden, inclusief verwijzing naar relevante werkervaring en opleidingen?

Antwoord vraag 22: Bij concernaudit zitten zowel operational auditors als IT auditors. Zowel voor de technische als de operationele aspecten van de toepassing van AI is daarmee in basis kennis aanwezig. Per situatie wordt ingeschat of er aanvullende kennis ingehuurd moet worden.

Bij de korpscontroller en de afdeling Concernaudit is de kennis over AI in opbouw. Deskundigheid wordt deels van buiten aangetrokken en is deels aanwezig bij de zittende mensen of wordt door hen verder ontwikkeld. De inhoudsdeskundigen van de genoemde afdelingen zijn aangehaakt bij de korpsbrede ontwikkeling en toepassing van algoritmes en data analysemethoden.

De implementatie van het Kwaliteitskader Big Data kent een opleidingstraject. Hier kan ook door het lijnmanagement, de korpscontroller en de afdeling Concernaudit gebruik van worden gemaakt.

58. Hoe worden beslissers bij de politie juridisch en ethisch getraind en opgeleid om verantwoorde besluiten te kunnen nemen op basis van door AI ondersteunde processen? Welke rol moet de Politieacademie vervullen om aan deze vaardigheidseisen te voldoen?

Antwoord vraag 58: Voor de toepassing van AI gelden de gebruikelijke wettelijke kaders, waaronder de privacywetgeving en WPG Wpg. Deze wettelijke kaders zitten verankerd in het basispolitieonderwijs. Daarnaast informeert de Gegevensautoriteit van politie de medewerkers middels een nieuwsbrief over privacy- en WPG-aangelegenheden.

Politie heeft bijgedragen aan de ontwikkeling van de begeleidingsethiek door bureau ECP, platform voor de informatiesamenleving. Deze methode wordt nu ingevoerd en bestaat er uit dat o.a. de inzet van AI in het politiewerk vanuit diverse perspectieven wordt beoordeeld. De beslissers leren hierdoor de specifieke kenmerken en ethische risico's van zo'n toepassing in de praktijk te herkennen. Ze leren ook, hoe en waar in de organisatie ze hun vragen of eventuele zorgen neer kunnen leggen.
Ook is de politie aangesloten bij de Nederlandse AI coalitie. De politie neemt verder deel aan de pilot " Ethics guidelines of trustworthy AI". Van de High level expert group on AI van de Europese Commissie en aan het Transparantielab van het Ministerie van Binnenlandse zaken.

Vanwege het belang van privacy en ethiek voor het politiewerk is recent een Politiechef als portefeuillehouder privacy en ethiek aangesteld. Voor de toepassing van AI gelden de gebruikelijke wettelijke kaders, waaronder de Wpg. Deze wettelijke kaders zitten verankerd in het basispolitieonderwijs. Daarnaast informeert de Gegevensautoriteit van politie de medewerkers middels een nieuwsbrief over privacy- en WPG-aangelegenheden.

Tenslotte wordt ook in het Kwaliteitskader big data aandacht besteed aan juridische en ethische aspecten. Bij de implementatie van het kwaliteitskader big

data wordt in samenwerking met de politieacademie onderzocht op welke wijze het kwaliteitskader ingebed kan worden in leer- en kwaliteitsprocessen en welke doelgroepen hierbij in aanmerking komen.

63. In hoeverre zet u het verlagen van de werkdruk in als strategie om de negatieve effecten van AI binnen de opsporing te voorkomen (zie het rapport Beginselen Digitaal (Rapport van een verkennend Onderzoek in opdracht van het WODC, Kamerstuk 29279, nr. 388, waaruit blijkt dat bij een hoge werkdruk uitkomsten van de computer vaker klakkeloos worden overgenomen)?

Antwoord vraag 63: Bij het gebruik van AI is het belangrijk dat de menselijke autonomie en de menselijke controle op de AI toepassing die wordt gebruikt betekenisvol is. Dat betekent dat de politiemedewerker voldoende in staat moet zijn om de uitkomst van een AI toepassing kritisch te bezien, en in staat moet zijn om de uitkomst indien nodig naast zich neer te leggen. Daar hoort ook bij dat de medewerker zich niet genoodzaakt voelt de uitkomst van de AI toepassing te accepteren door een verhoogde werkdruk.
Omdat AI een verandering in het werk van de politie teweeg brengt wordt ook in het kader van het Nationaal Politielab AI onderzoek gedaan naar personale en organisatorische consequenties van de toepassing van AI binnen politieprocessen en de randvoorwaarden die daarbij horen.

Predictive policing

10. Kan predictive policing een opsporingsmethode genoemd worden?

12. Hoe verhoudt predictive policing zich tot het beginsel dat er sprake moet zijn van een redelijk vermoeden van een strafbaar feit, een verdenking, voordat overgegaan kan worden tot opsporing?

Antwoord vragen 10 en 12: Predictive policing wordt binnen de politie gebruikt als hulpmiddel om een inschatting te maken of er een (verhoogd) risico is op bepaalde vormen van criminaliteit in een bepaald gebied zodat hierop vooraf kan worden geanticipeerd bij de inzet van de gebiedsgebonden politie. Het is geen opsporingsmethode en wordt ook niet gebruikt voor opsporen. Bij predictive policing is er dus geen sprake van een redelijk vermoeden van een strafbaar feit of een verdenking. Het risico wordt bepaald op basis van historische data, waargenomen trends en patronen waarbij alleen wordt gekeken naar het optreden van een (verhoogd) risico in een bepaald gebied. Het is niet gericht op individuen of groepen.

13. Wordt overwogen iets te doen aan predictive identification, een vorm van predictive policing waarin personen of groepen aangewezen kunnen worden als potentiële misdadigers of slachtoffers van bepaalde strafbare feiten?

Antwoord vraag 13: De politie maakt gebruik van (klassieke) risicotaxatiemodellen om inschattingen te kunnen maken ten aanzien van het potentieel gebruik maken van geweld of het optreden van recidive bij personen met antecedenten zoals vastgelegd in de politiesystemen. Het taxatiemodel geeft

alleen een inschatting of er een verhoogd risico is op het gebruik van geweld of het optreden van recidive bij een geselecteerde groep personen.
Met betrekking tot het kunnen identificeren van potentiele slachtoffers van bepaalde strafbare feiten zijn modellen ontwikkeld om een inschatting te kunnen maken of er sprake is van discriminatie of kindermishandeling bij het screenen van cases. Daarnaast wordt onderzocht of het mogelijk is om potentieel kwetsbare asielzoekers, mensenhandel of onvrijwillige prostitutie te kunnen herkennen.

Gezichtsherkenning

7. Wordt overwogen te gaan werken of experimenteren met realtime gezichtsherkenning?

Antwoord vraag 7:
De maatschappij ontwikkelt zich en de politie ontwikkelt daarin mee. Het past in die lijn dat er ook binnen de politie wordt nagedacht over de mogelijkheden van gezichtsherkenning. Ik heb dit ook aan uw Kamer bericht in mijn brief van november 2019.[13] Daarin heb ik benadrukt dat de politie uiteraard moet kunnen experimenteren met vormen van gezichtsherkenningstechnologie, maar dat deze niet operationeel mag worden ingezet dan nadat met betrokken partijen inzichtelijk is gemaakt wat het wettelijk kader is, welke waarborgen er getroffen zijn en wat de uitkomst is van de juridisch ethische toets. Op dit moment zijn er binnen de politie geen experimenten met realtime gezichtsherkenning.

11. Past het gebruik van realtime gezichtsherkenning door voormalig EU-collega Engeland binnen het Coordinated Plan on Artificial Intelligence van de EU?

Antwoord vraag 11:
Het is niet aan mij om te beoordelen of het gebruik van technologie van een voormalige EU lidstaat past binnen het *Coordinated Plan on Artificial Intelligence* van de EU.

Politielab AI

52. Op welke manier gaat het wetenschappelijke karakter van het Nationale Politielab AI bewaakt worden? Welke waarborgen heeft u daarvoor ingebouwd? Welke rol gaat het Transparantielab hierbij spelen?

Antwoord vraag 52: Het Nationaal Politielab AI is een samenwerkingsverband in ICAI-verband tussen de politie en de Universiteit van Amsterdam en de Universiteit Utrecht. Er zijn twee wetenschappelijk directeuren betrokken, die mede de wetenschappelijke standaarden bewaken.

In het Nationaal Politielab AI wordt onderzoek uitgevoerd door promovendi en post docs die onderzoek doen aan Universiteit Utrecht of de Universiteit van Amsterdam, maar die ook (parttime) in dienst zijn van de politie. Op deze wijze

[13] Kamerstukken II 2019/20, 32761, 152.

wordt een unieke mix van wetenschappelijke kennis en onderzoek en operationele toepasbaarheid gecreëerd.

De politie neemt ook deel aan het Transparantielab van het Ministerie van Binnenlandse Zaken, waar de 'richtlijnen voor het toepassen van algoritmen door overheden'[14] worden getest.

61. Hoeveel is reeds begroot voor het opzetten van het Nationaal Politielab AI?

Antwoord vraag 61: De financiering wordt gerealiseerd vanuit de incidentele gelden voor digitalisering en cybercrime. Indien blijkt dat het Nationaal Politielab AI bijdraagt aan de doelstellingen van de politie op dit gebied zal gekeken worden naar een meer structurele financiering.

62. In hoeverre werkt het Nationaal Politielab AI samen met het Openbaar Ministerie (OM)?

Antwoord vraag 62: Voor zover het gaat om het ontwikkelen en toepassen van AI binnen de taakuitvoering dat onder het gezag van het OM plaatsvindt is het OM betrokken bij de wijze waarop dit wordt ingericht.

Ook de ontwikkeling van het Kwaliteitskader Big Data heeft plaatsgevonden in samenwerking tussen het OM en de politie.

65. Wat voor soort publicaties over AI wil de politie in wetenschappelijke tijdschriften gaan plaatsen? Hoe voorkomt de politie daarbij dat de criminelen meelezen? Of wordt er gemakshalve vanuit gegaan dat criminelen dit niet gaan doen?

Antwoord vraag 65: Er wordt gepubliceerd over wetenschappelijke inzichten die worden opgedaan in het onderzoek, en niet over operationele informatie. Met de universiteiten is vastgelegd dat de politie het recht heeft niet over te gaan tot publicatie als dit ongewenste inzichten biedt op politietactieken of de staatsveiligheid in gevaar kan brengen.

<u>Betrokkenheid bedrijven en wetenschap</u>

6. Worden private bedrijven betrokken bij het ontwerp van AI systemen bij de politie en zo ja, welke? Hoever strekt die samenwerking?
49. Welke externe partijen zoals (universiteiten, bedrijven en individuen werken op dit moment samen met de politie bij het ontwikkelen van AI toepassingen? Kunt u beschrijven op welke wijze deze externe partijen met de politie samenwerken, indachtig de opmerking in de brief dat de politie "niet te afhankelijk" moet worden van andere partijen?
66. Met welke private partijen werkt de politie samen in het kader van AI? Indien geen private partijen worden betrokken bij de ontwikkeling en toepassing van kunstmatige intelligentie in de politiepraktijk, waarom niet?

[14] Kamerstukken II, 2019/20, 26643, 641.

67. Met welke nationale of internationale kennispartners werkt de politie samen? Hoe wordt geborgd dat zij geen veiligheidsrisico kunnen vormen?

68. Hoe wordt de betrouwbaarheid van kennispartners van de politie getoetst en wordt deze toetsing regulier uitgevoerd?

Antwoord vragen 6, 49, 66, 67, 68: De politie werkt in het kader van het Nationaal Politielab AI samen met de Universiteit Utrecht en de Universiteit van Amsterdam. Incidenteel wordt er ook samengewerkt met andere universiteiten. Daarnaast werkt de politie op het gebied van AI samen met diverse Europese lidstaten, zowel in Europol verband als ook bilateraal.

In bepaalde gevallen werkt de politie samen met bedrijven bij de ontwikkeling van AI. Daarbij is het wel vereist dat deze bedrijven voldoen aan de ethische vereisten (transparantie, uitlegbaarheid, etc.).

Voor de politie is het van groot belang dat de partners en kennisinstellingen met wie zij samenwerkt betrouwbare partners zijn. Dat geldt niet alleen in het kader van AI, maar algemeen voor samenwerking en kennisuitwisseling van de politie. Voor de beoordeling van de betrouwbaarheid van partners worden de gebruikelijke procedures gevolgd.

Met het wetsvoorstel screening ambtenaren van politie en politie-externen[15] dat recentelijk door uw Kamer is aangenomen wordt een wettelijke grondslag gecreëerd voor het screenen van externen en worden de screeningsbevoegdheden van de politie uitgebreid.

36. Hoe gaat de overheid en dus de politie waarborgen dat men niet te afhankelijk wordt van private partijen die dit soort technologieën leveren?

Antwoord vraag 36: De politie haalt onder andere in het Nationaal Politielab AI kennis en kunde naar binnen. Promovendi die in dienst zijn van de politie werken gedeeltelijk aan de Universiteit en gedeeltelijk dicht bij de uitvoering om AI toepassingen te ontwikkelen die bijdragen aan de taakuitvoering van de politie. Door de kennis zelf in huis te halen is de politie niet afhankelijk van externe leveranciers. Indien samengewerkt wordt met externe leveranciers moeten zij voldoen aan de ethische vereisten (transparantie, uitlegbaarheid, etc.).

<u>Europa</u>

15. Kan worden aangegeven waar Nederland staat in vergelijking tot andere Europese landen als het gaat om de ontwikkeling en toepassing van AI?

Antwoord vraag 15: De politie in Nederland heeft in vergelijking met politiekorpsen in andere landen in Europa relatief veel kennis in huis over AI en is een van de voorlopers in de ontwikkeling van AI. Dat komt onder andere door het samenwerkingsverband in het Nationaal Politielab AI waarbij ontwikkeling van AI binnen de politie plaatsvindt in samenwerking met de wetenschap.

[15] *Kamerstukken II* 2018/19, 35 170, 2.

39. Kan worden aangegeven in hoeverre nu wordt ingezet op kennisuitwisseling met andere Europese landen als het aankomt op AI?

Antwoord vraag 39: Het is belangrijk om bij te blijven met de nieuwste ontwikkelingen op het gebied van AI, onder andere om ook de ontwikkelingen die in het criminele veld plaatsvinden bij te houden. Daarom wordt door de politie nauw samengewerkt en kennis uitgewisseld in Europol verband, maar ook bilateraal met andere Europese landen. Deze positie is alleen mogelijk omdat de politie over eigen AI experts beschikt. Daarnaast wordt er vanuit de politie en mijn departement deelgenomen aan diverse werkgroepen over AI en nieuwe technologieën in de EU.

Cybercrime

34. In hoeverre maken criminelen ook gebruik van AI? Gebruiken cybercriminelen ook algoritmes en AI om slachtoffers te kiezen voor hackaanvallen? Op welke manier kan dit tegengegaan worden?

Antwoord vraag 34: Over het algemeen blijkt dat (cyber)criminelen gebruik maken van nieuwe technologische mogelijkheden voor het plegen van criminaliteit. De politie heeft nog geen concrete indicaties van het gebruik van zelflerende algoritmes om keuzes te maken in het cybercriminele proces. Wel is het zo dat cybercriminelen processen automatiseren, optimaliseren en schaalbaar maken, zowel voor het plegen als het beschermen van cybercrime.

Binnen de industrie wordt gewaarschuwd voor het ontwikkelen van AI-toepassingen voor (cyber)crimineel gebruik, waarbij meestal wordt verwezen naar het ontwikkelen en gebruiken van deepfakes ten behoeve van phishing en fraude. In de praktijk heeft de politie hier in Nederland nog geen voorbeelden van gezien.

35. Hoe kan de overheid en dus ook de politie de criminelen voor blijven in de technologische wedloop? Kunt u daarbij betrekken wat daarover is gezegd in Nieuwsuur van 5 februari 2020? Deelt u de daarin geuite mening dat het helemaal niet vaststaat dat de overheid deze wedloop gaat winnen van de criminelen

Antwoord vraag 35: Om criminelen voor te kunnen blijven in de technologische wedloop is het van belang dat de politie ook gebruik kan maken van de nieuwe technologische ontwikkelingen ten behoeve haar taakuitvoering. Ik vind het daarom ook van belang dat de politie de ruimte krijgt om te experimenteren met de inzet van nieuwe technologie, uiteraard binnen de geldende kaders. Het Nationaal Politielab AI is hier een voorbeeld van.

79. Welke strafbare feiten met behulp van algoritmes en AI worden er gepleegd, en hoe vaak kwamen deze strafbare feiten in 2019 voor?

Antwoord vraag 79: De politie heeft op dit moment geen aanwijzingen dat in Nederland al strafbare feiten gepleegd worden met AI.

AI Coalitie

40. Welke partijen nemen naast de overheid nog meer deel aan de Nederlandse AI Coalitie?

Antwoord 40: Naast overheden gaat om het grootbedrijf, MKB , kennis- en onderwijsinstellingen, brancheorganisaties en ziekenhuizen. U vindt een overzicht van de deelnemende partijen via https://nlaic.com/coalitiepartners

41. Welke criteria worden gebruikt om partijen te selecteren voor de AI Coalitie?

Antwoord vraag 41: Van deelnemende partijen worden concrete bijdragen verwacht aan de werkzaamheden van de Coalitie. Het moet gaan om in Nederland gevestigde rechtspersonen (geen privé personen), die aantoonbaar activiteiten hebben of plannen hebben, die bijdragen aan de ontwikkeling van Nederland op AI-gebied.

Overig

3. Wanneer vindt de evaluatie van de wet Computercriminaliteit III plaats?

Antwoord vraag 3: Twee jaar na de inwerkingtreding op 1 maart 2019 wordt de wet CCIII geëvalueerd door het WODC.

31. Wordt ook overwogen om het Juridisch Loket te vervangen door AI? Zal zo'n Juridische Robot een positief of een negatief effect hebben op de kwaliteit van de advisering?

Antwoord vraag 31: In de tweede voortgangsrapportage van het programma Rechtsbijstand (brief d.d. 19 december 2019, Kamerstukken II, 31753, nr.190) is uiteengezet op welke manier het Juridisch Loket bezig is de eigen dienstverlening dichter bij de burger te brengen. Onderdeel van het programma is ook de verbetering van online informatie en advies. Het einddoel is dat meer mensen terecht kunnen in een betrouwbare, begrijpelijke online omgeving, die met meer functionaliteiten op interactieve wijze helpt om problemen op te lossen. De komende tijd wordt een plan van aanpak gemaakt voor deze online omgeving. De minister voor Rechtsbescherming sluit niet bij voorbaat uit dat in deze online omgeving ook artificiële intelligentie (AI) wordt ingezet. De inzet van AI en algoritmen in de rechtspleging dient steeds op verantwoorde wijze te geschieden. IJkpunten daartoe heeft de minister voor Rechtsbescherming geschetst in zijn brief aan de Eerste Kamer.[16]

Robocop en andere filmverwijzingen

32. Zijn politieagenten op straat ook te vervangen door AI? Aan welke concrete verschijningsvorm kan daarbij gedacht worden? Moet daarbij gedacht worden aan een variant op Robocop (een figuur uit een Amerikaanse sciencefiction-actiefilm uit 1987 en 2014)?

[16] Kamerstukken I, 2018/19, 34775-VI, AH.

33. Wat waren de positieve en negatieve kanten aan het functioneren van Robocop, in uw ogen? Kunt u hierbij ook ingaan op de kwaliteiten van Judge Dredd (een Amerikaanse actiefilm uit 1995)? Valt de Terminator (1984 e.v.) volgens u ook onder de gewenste AI?

Antwoord vragen 32 en 33: Er is geen sprake van het vervangen van politieagenten op straat door AI. AI wordt gebruikt ter ondersteuning van het politiewerk. Zoals vele organisaties (bijvoorbeeld in de zorg) experimenteert de politie wel op kleine schaal met robotisering. Het is niet aan mij als Minister van Justitie en Veiligheid om in te gaan op de kwaliteiten, positieve en negatieve kanten van fictieve karakters uit Hollywoodfilms.

74. Klopt het dat het boek "1984" (met daarin Big Brother) uit 1948, en de daarin beschreven surveillancestaat, in China inmiddels een realiteit geworden is? Kunt u uitsluiten dat elementen van zo'n politiestaat in Nederland ingang zullen vinden?

Antwoord vraag 74: Helder moge zijn dat ik geen surveillance staat voor sta. De ontwikkelingen over AI bij de politie zoals die uit brief blijken geven ook geen aanleiding voor vrees daartoe. De Nederlandse wetgeving, de Europese richtlijnen en het EVRM zijn en blijven leidend bij de verdere ontwikkeling van AI. Ten aanzien van de vraag over China verwijs ik naar de brief van de minister van Buitenlandse zaken dd. 15 mei jl.[17]

[17] Kamerstukken II, 2018/19, 35207, 1.

ECLI:NL:PHR:2017:1051

Instantie	Parket bij de Hoge Raad
Datum conclusie	25-09-2017
Datum publicatie	20-10-2017
Zaaknummer	17/01448
Formele relaties	Arrest Hoge Raad: ECLI:NL:HR:2018:1316, Gevolgd
Rechtsgebieden	Belastingrecht
Bijzondere kenmerken	-
Inhoudsindicatie	

Voor bijlage zie ECLI:NL:PHR:2017:1081.

A-G IJzerman heeft conclusie genomen naar aanleiding van het beroep in cassatie van het college van burgemeester en wethouders van de gemeente Waalwijk tegen de uitspraak van Gerechtshof 's-Hertogenbosch van 10 februari 2017, nr. 15/01429.

Deze procedure maakt deel uit van de zeven verwante zaken waarin vandaag conclusie wordt genomen. Bij die conclusies behoort een gemeenschappelijke bijlage. Alle zaken hebben gemeen dat zij, op enigerlei wijze, zien op de vraag of bepaalde fysieke stukken of computerbestanden, zijn aan te merken als ‘op de zaak betrekking hebbende stukken’.

In deze zaak gaat het met name om de vraag of een grondstaffel (de gehanteerde kavelwaarde en de opbouw hiervan binnen een perceel) is aan te merken als een op de zaak betrekking hebbend stuk, als bedoeld in artikel 7:4, tweede lid, van de Awb.

Het Hof heeft overwogen dat de Heffingsambtenaar de grondstaffel eerst heeft verstrekt als onderdeel van de nieuwe matrix bij zijn verweerschrift in hoger beroep. Het Hof heeft geconstateerd dat indien de nieuwe matrix en de oorspronkelijke matrix met elkaar worden vergeleken, blijkt dat er voor wat betreft de grondwaardes grote verschillen zijn, welke zijn ontstaan doordat verschillende grondstaffels zijn gebruikt. Naar het oordeel van het Hof volgt uit artikel 40 van de Wet WOZ dat een belastingplichtige recht heeft op alle gegevens die van belang kunnen zijn voor het controleren van de voor zijn onroerende zaak vastgestelde waarde, een en ander met inachtneming van het belang van de privacy van anderen. Het Hof heeft vervolgens overwogen dat ook de grondstaffel in dit verband van belang kan zijn. In casu is daarvan volgens het Hof inderdaad sprake.

Het Hof heeft voorts geoordeeld dat ook afgezien van artikel 40 van de Wet WOZ de Heffingsambtenaar de grondstaffel in de bezwaarfase ter beschikking had moeten stellen van belanghebbende, nu belanghebbendes gemachtigde hierom reeds in de bezwaarfase had verzocht. Immers ingevolge artikel 7:4, tweede lid, van de Awb dient de Heffingsambtenaar alle op de zaak betrekking hebbende stukken voorafgaand aan het horen ter inzage te leggen voor belanghebbende gedurende ten minste een week. Naar het oordeel van het Hof had de

Heffingsambtenaar de grondstaffel, als zijnde een op de zaak betrekking hebbend stuk als bedoeld in de zin van artikel 7:4, tweede lid, van de Awb ter inzage moeten leggen, en indien de

belanghebbende bij de inzage om een afschrift daarvan had verzocht, hem dat moeten verstrekken. Gelet hierop had de Heffingsambtenaar het verzoek van belanghebbende (een afschrift van) de grondstaffel te verstrekken niet mogen weigeren.

Desalniettemin vormt het ten onrechte weigeren om de grondstaffel te verstrekken voor het Hof geen aanleiding om de uitspraken op bezwaar te vernietigen, omdat de beschikte waarde, naar blijkt uit vergelijkbare objecten, niet te hoog is vastgesteld. Wel heeft het Hof de Heffingsambtenaar veroordeeld in de proceskosten van de beroepsfase en de hoger beroepsfase.

Het College heeft bij tegen de Hofuitspraak beroep in cassatie ingesteld, onder aanvoering van vier middelen.

Het eerste middel van het College houdt in dat het Hof ten onrechte heeft aangenomen dat de grondstaffel is neergelegd in een stuk dat voor verstrekking in aanmerking komt.

De A-G ziet als ‘de op de zaak betrekking hebbende stukken’ ook gegevens die zijn opgenomen in applicaties die door de inspecteur of heffingsambtenaar bij de belastingheffing zijn gebruikt. Verder merkt de A-G op dat de Hoge Raad bij arrest van 20 december 2013 heeft overwogen dat onder ‘op de zaak betrekking hebbende stukken’ mede zijn verstaan afdrukken, als in afschrift verstrekt, van in elektronische vorm vastgelegde gegevens. Nu de grondstaffel in een dergelijke elektronische vorm was opgenomen en (uiteindelijk) in afschrift is verstrekt, meent de A-G dat reeds daarom het eerste middel van het College faalt.

In het tweede middel stelt het College dat de uitspraak van het Hof rechtsongelijkheid creëert, nu deze uitspraak afwijkt van jurisprudentie van de Afdeling Bestuursrechtspraak van de Raad van State. Daarover merkt de A-G op dat in fiscale zaken moet worden gevaren op de rechtsoordelen van (de belastingkamer van) de Hoge Raad. Overigens meent de A-G dat de door belanghebbende gepercipieerde rechtsongelijkheid berust op een onjuiste rechtsopvatting.

Met het derde middel wordt geklaagd over de overweging van het Hof dat de grondstaffel voorafgaand aan de hoorzitting ter inzage had moeten worden gelegd. De A-G meent daarentegen dat nu de grondstaffel behoort tot ‘de op de zaak betrekking hebbende stukken’, het College de grondstaffel reeds op eigen initiatief had moeten overleggen.

Overigens heeft het College gesteld het verstrekken van de verzochte gegevens afbreuk zou doen aan de strekking van de in artikel 40 van de Wet WOZ neergelegde regeling die een vorm van beperkte openbaarheid beoogt en niet valt in te zien waarom ten aanzien van artikel 7:4 van de Awb een andere redenering zou moeten worden gevolgd.

De A-G zou die benadering niet willen volgen. Een belangrijk verschil tussen beide regelingen is dat hier in de Woz wordt uitgegaan van een verzoek, terwijl het bij toepassing van artikel 7:4, tweede lid, van de Awb, gaat om een eigen verplichting van de inspecteur.

In het vierde middel bestrijdt het College 's Hofs overweging dat eerst in hoger beroep voldoende inzicht is geboden in de totstandkoming van de waarde. De A-G leidt echter af uit het procesverloop dat die overweging van het Hof juist te achten is, zodat ook het vierde middel faalt.

De conclusie strekt ertoe dat het beroep in cassatie van het college van burgemeester en wethouders van de gemeente Waalwijk ongegrond dient te worden verklaard.

Vindplaatsen

Rechtspraak.nl
Belastingblad 2017/467
V-N Vandaag 2017/2490
V-N 2017/54.3 met annotatie van Redactie
Viditax (FutD), 20-10-2017
FutD 2017-2627
NLF 2017/2814 met annotatie van Jits Berns

Conclusie

PROCUREUR-**G**ENERAAL BIJ DE **H**OGE **R**AAD DER **N**EDERLANDEN

MR. R.L.H. IJZERMAN

ADVOCAAT-GENERAAL

Conclusie van 25 september 2017 inzake:

Nr. Hoge Raad: 17/01448	**B & W gemeente Waalwijk**
Nr. Gerechtshof: 15/01429 Nr. Rechtbank: 14/7668	
Derde Kamer B	**tegen**
Wet waardering onroerende zaken 2014	**[X]**

1 Inleiding

1.1 Heden neem ik conclusie in de zaak met nummer 16/04497 naar aanleiding van het beroep in cassatie van het college van burgemeester en wethouders van de gemeente Waalwijk (hierna: het College) tegen de uitspraak van Gerechtshof 's-Hertogenbosch (hierna: het Hof) van 10 februari 2017.[1]

1.2 Belanghebbende, is eigenaar van een onroerende zaak, plaatselijk bekend als [a-straat 1] te [Z] (hierna: de onroerende zaak). Bij beschikking, als bedoeld in artikel 22 van de Wet waardering onroerende zaken (hierna: Wet WOZ) van 28 februari 2014, is de waarde per peildatum 1 januari 2013, voor het belastingjaar 2014 vastgesteld op € 455.000. Bij gelijktijdig gegeven en in hetzelfde geschrift vervatte beschikking is aan belanghebbende terzake van de onroerende zaak een aanslag in de onroerendezaakbelasting (hierna: OZB) over het jaar 2014 opgelegd.

1.3 Op het aanslagbiljet staat dat de vastgestelde waarde van de onroerende zaak is bepaald op basis van een vergelijkbaarheidsanalyse. Voor meer informatie over de waardebepaling wordt verwezen naar 'www.waalwijk.nl -> digitaal loket'. Voorts wordt vermeld dat aldaar ook het taxatieverslag van de onroerende zaak kan worden gevonden.

1.4 Belanghebbende is op 31 maart 2014 tegen de WOZ-beschikking en de aanslag OZB in bezwaar gekomen. Op 28 augustus 2014 heeft in dat kader een hoorzitting plaatsgevonden. De gemachtigde van belanghebbende heeft daar, op de voet van artikel 40 van de Wet WOZ, aan de Heffingsambtenaar verzocht bij de uitspraak op bezwaar (…) de kavelwaarde en de opbouw hiervan bekend te maken' (hierna: de grondstaffel).

1.5 Het hiervoor genoemde artikel 40, eerste lid, van de Wet WOZ (oud) biedt grondslag voor de Heffingsambtenaar om 'op verzoek (…) het waardegegeven van een bepaalde onroerende zaak (…) [te verstrekken, A-G] aan een ieder die kan aantonen uit hoofde van de belastingheffing te zijnen aanzien een gerechtvaardigd belang te hebben bij de verkrijging daarvan'. In het tweede lid staat dat 'een afschrift van de gegevens die ten grondslag liggen aan de vastgestelde waarde' op verzoek uitsluitend wordt verstrekt 'aan degene te wiens aanzien een beschikking is genomen'.[2]

1.6 Belanghebbendes bezwaren zijn op 11 november 2014, bij in één geschrift vervatte uitspraken op bezwaar, door de Heffingsambtenaar ongegrond verklaard. Bij de uitspraken op bezwaar heeft de Heffingsambtenaar niet de door belanghebbende opgevraagde gegevens verstrekt.

1.7 Belanghebbende is tegen de uitspraken op bezwaar bij beroepschrift van 23 december 2014 in beroep gekomen bij Rechtbank Zeeland-West-Brabant (hierna: de Rechtbank). Belanghebbende heeft in beroep wederom verzocht om overlegging van de grondstaffel.

1.8 Hangende de procedure bij de Rechtbank is in opdracht van de gemeente door [A] (hierna: de taxateur), na een uitpandige opname van de onroerende zaak, een taxatierapport d.d. 28 januari 2015 opgemaakt. Aanvankelijk was met belanghebbende een afspraak gemaakt voor een inpandige opname. Deze afspraak is door de gemachtigde van belanghebbende geannuleerd, waarbij de gemachtigde heeft aangegeven geen toestemming te verlenen voor een inpandige opname. Blijkens het taxatierapport, waarin vermeld worden de referentieobjecten [b-straat 1] te [Z], [c-straat 1] te [Q] en [d-straat 1] te [Q], is de WOZ-waarde van de onroerende zaak € 470.000. Bij het taxatierapport is een matrix (de oorspronkelijke matrix) gevoegd waarin de voor de waardebepaling van de onroerende zaak en de referentieobjecten relevante gegevens zijn opgenomen.

1.9 De door belanghebbende in bezwaar en beroep opgevraagde grondstaffel maakt geen onderdeel uit van voornoemde matrix. De grondstaffel is ook later in de procedure niet overgelegd door de Heffingsambtenaar verstrekt.

1.10 De Rechtbank heeft belanghebbendes beroep ongegrond verklaard bij uitspraak van 27 oktober 2015. De Rechtbank heeft daartoe eerst overwogen 'dat de Heffingsambtenaar met het taxatierapport aannemelijk heeft gemaakt dat de vastgestelde waarde niet te hoog is'.

1.11 Vervolgens heeft de Rechtbank overwogen dat de Heffingsambtenaar is gehouden 'op grond van artikel 7:4, tweede lid, van de Awb (…) alle op de zaak betrekking hebbende stukken voorafgaand aan het horen ter inzage [te] leggen voor belanghebbende gedurende ten minste een week. Het

derde lid schrijft voor dat bij oproeping van het horen wordt vermeld waar en wanneer de stukken ter inzage zullen liggen. Het vierde lid bepaalt dat belanghebbende van de stukken tegen vergoeding van ten hoogste de kosten afschriften verkrijgen'.

1.12 De Rechtbank heeft daarna overwogen dat 'de grondstaffel (…) een stuk [is] dat van belang is geweest voor de primaire besluitvorming, en (…) daardoor een op de zaak betrekking hebbend stuk [is] zoals bedoeld in artikel 7:4, tweede lid, van de Awb (vgl. HR 20 maart 2009, ECLI:NL:HR:2009:BH6420). Dit betekent (…) dat (het stuk met) de grondstaffel ter inzage had moeten worden gelegd'.

1.13 Een en ander heeft de Rechtbank tot de conclusie geleid dat 'de heffingsambtenaar ten onrechte [heeft] geweigerd om een afschrift van (het stuk met) de grondstaffel aan belanghebbende te verstrekken. De omstandigheid dat belanghebbende bij zijn verzoek heeft gerefereerd aan artikel 40 van de Wet WOZ maakt dat niet anders.'

1.14 De Rechtbank heeft daar aan toegevoegd dat 'uit artikel 40, tweede lid, van de Wet WOZ niet [kan] worden afgeleid dat deze bepaling een beperking vormt van de verplichtingen op grond van artikel 7:4 van de Awb. In artikel 40, tweede lid, van de Wet WOZ is bepaald dat de heffingsambtenaar uitsluitend aan degene te wiens aanzien een beschikking is genomen, op verzoek een afschrift verstrekt van de gegevens die ten grondslag hebben gelegen aan de vastgestelde waarde (beperkte openbaarheid). Het woord 'uitsluitend' heeft betrekking op 'aan degene te wiens aanzien een beschikking is genomen'. De bepaling houdt dus niet in dat aan diegene 'uitsluitend' bedoeld afschrift mag worden verstrekt (dus met uitsluiting van andere afschriften).'

1.15 Desalniettemin heeft de Rechtbank in de schending van artikel 7:4 van de Awb geen aanleiding gezien 'om de uitspraken op bezwaar te vernietigen (artikel 6:22 van de Awb), maar wel om te gelasten dat de heffingsambtenaar wordt veroordeeld in de proceskosten van de beroepsfase en het door belanghebbende betaalde griffierecht. Bij dat oordeel heeft de rechtbank in acht genomen dat de gemachtigde van belanghebbende ter zitting geloofwaardig heeft verklaard dat hij om de grondstaffels heeft verzocht om een inschatting te kunnen maken of hij al dan niet beroep wilde aantekenen, alsmede dat de heffingsambtenaar ook in beroep de grondstaffel niet heeft ingebracht'.

1.16 De Heffingsambtenaar is, bij geschrift van 16 december 2015, tegen de uitspraak van de Rechtbank in hoger beroep gekomen bij het Hof. Bij brief van 19 februari 2016 heeft belanghebbende verweer gevoerd en gelijktijdig incidenteel hoger beroep ingesteld.

1.17 Het Hof heeft zowel het hoger beroep als het incidenteel hoger beroep ongegrond verklaard.

1.18 Het Hof heeft overwogen dat 'de Heffingsambtenaar (…) in het onderhavige geval eerst in hoger beroep de grondstaffel [heeft] overgelegd'. Daarna is bij de zienswijze van de Heffingsambtenaar op het incidentele hoger beroep van belanghebbende een nieuwe matrix gevoegd waarin de grondstaffel is vermeld. 'Deze nieuwe matrix wijkt voor wat betreft de grondwaarde af van de oorspronkelijke matrix (…). Het argument van de Heffingsambtenaar dat het geven van inzage in de grondstaffel in het onderhavige jaar op praktische problemen stuitte, wordt door het Hof verworpen. De Heffingsambtenaar dient zijn administratie zodanig in te richten dat de noodzakelijke bouwstenen voor de waardebepaling hieruit op doelmatige wijze zijn te destilleren. Dit klemt te meer nu de (…) verschillen pas in hoger beroep aan het licht zijn gekomen. Het in artikel 40 van de Wet WOZ besloten beginsel dat belanghebbende in staat moet worden gesteld de waardebepaling te controleren, zou anders inhoudsloos worden. De omstandigheid dat artikel 40 van de Wet WOZ een lex specialis is ten opzichte van de Wet openbaarheid bestuur (hierna: WOB) maakt dit, gelijk de Rechtbank terecht heeft beslist, niet anders. Gezien het voren overwogene is het Hof van oordeel dat de Heffingsambtenaar aan het verzoek van de gemachtigde om de grondstaffel te verstrekken, tegemoet had moeten komen.'

1.19 Overigens is het Hof tot het oordeel gekomen ‘dat de Heffingsambtenaar aannemelijk heeft gemaakt dat de waarde van de onroerende zaak, alsmede de daarop gebaseerde aanslag, door de Heffingsambtenaar niet te hoog zijn vastgesteld’.

1.20 Het Hof heeft tot slot in ‘het ten onrechte weigeren om de grondstaffel te verstrekken’ aanleiding gezien ‘de Heffingsambtenaar te veroordelen in de proceskosten van de hoger beroepsfase. Belanghebbende is immers door de weigerachtige houding van de Heffingsambtenaar moeten blijven procederen tot in hoger beroep om een gefundeerd oordeel te kunnen vormen over de door de Heffingsambtenaar vastgestelde waarde om aan de hand van de ontvangen gegevens in alle redelijkheid te kunnen inschatten of hij al dan niet (hoger) beroep zou instellen. Pas na ontvangst van de zienswijze op het incidentele hoger beroep, met daarbij gevoegd de nieuwe matrix met de grondstaffel en de definitieve, gewijzigde, cijfers betreffende de grondwaardes van de onroerende zaak en de referentieobjecten, kon belanghebbende de hiervoor vermelde inschatting maken’.

1.21 Het College heeft bij geschrift van 22 maart 2017 beroep in cassatie ingesteld.

1.22 Het College is tegen de uitspraak van het Hof in cassatie gekomen onder aanvoering van vier middelen.

1.23 Het eerste middel luidt dat ‘het Hof niet ingaat op de grief dat de grondstaffel niet is neergelegd in een stuk dat voor verstrekking in aanmerking komt’.

1.24 Met het tweede middel stelt het College dat de uitspraak van het Hof rechtsongelijkheid creëert, nu, zakelijk weergegeven, deze uitspraak afwijkt van jurisprudentie van de Afdeling Bestuursrechtspraak van de Raad van State (hierna: de ABRvS).

1.25 Met het derde middel wordt geklaagd over de oordelen van het Hof ‘dat de grondstaffel A) voorafgaand aan de hoorzitting ter inzage had moeten worden gelegd en B) valt onder de gegevens die op grond van artikel 40 van de Wet WOZ moeten worden verstrekt’.

1.26 In het vierde middel bestrijdt het College ’s Hofs overweging ‘dat eerst in hoger beroep voldoende inzicht is geboden in de totstandkoming van de waarde’.

1.27 Deze conclusie is verder als volgt opgebouwd. In onderdeel 2 worden de feiten en het procesverloop weergegeven, gevolgd door een beschrijving van het geding dat nu in cassatie voorligt in onderdeel 3. Onderdeel 4 omvat een overzicht van relevante wet- en regelgeving, wetsgeschiedenis, jurisprudentie en literatuur.[3] Onderdeel 5 bevat de beoordeling van de middelen, gevolgd door de conclusie in onderdeel 6.

1.28 Deze procedure maakt deel uit van een cluster van zeven verwante zaken waarin ik vandaag conclusie neem. Bij die conclusies behoort een gemeenschappelijke bijlage.[4] Alle zaken hebben gemeen dat zij, op enigerlei wijze, zien op de vraag of bepaalde fysieke stukken of computerbestanden, zijn aan te merken als ‘op de zaak betrekking hebbende stukken’. Of hiervan in een bepaald geval sprake is, hangt vaak mede af van de feiten van het geval, dus niet alleen van de uitleg van het recht; het gaat veelal om gemengde oordelen. Indien en voor zover geoordeeld moet worden dat er sprake is van ‘op de zaak betrekking hebbende stukken’, dient de Inspecteur die, in de bezwaarfase, ter inzage te geven aan de belanghebbende en dient die, in de beroepsfase, in te zenden aan de belastingrechter.

1.29 In deze zaak gaat het met name om de vraag of een grondstaffel is aan te merken als een op de zaak betrekking hebbend stuk, als bedoeld in artikel 7:4, tweede lid, van de Awb. Ook is in geschil of een grondstaffel door de heffingsambtenaar desverzocht moet worden overgelegd op grond van artikel 40 van de Wet WOZ. Zie ook de vergelijkbare conclusie in de met deze zaak

samenhangende zaak nr. 16/04497.

2 De feiten en het geding in feitelijke instanties

Feiten

2.1 Het Hof heeft de feiten vastgesteld:[5]

2.1. Belanghebbende is eigenaar van de onroerende zaak. De onroerende zaak heeft als bouwjaar 1939, heeft een inhoud van circa 816 m3, een perceeloppervlakte van 1.390 m2, een garage van 71 m3 en twee dakkapellen.

2.2. De beschikking WOZ en aanslagbiljet gemeentelijke belastingen voor het belastingjaar 2014 is gedagtekend 28 februari 2014. Op het aanslagbiljet is vermeld dat de gemeente de verkoopwaarde heeft bepaald op basis van nauwkeurig onderzoek onder vergelijkbare panden als het pand waarvan belanghebbende gebruik maakt. Voor meer informatie over de waardebepaling wordt verwezen naar 'www.waalwijk.nl -> digitaal loket'. Voorts wordt vermeld dat aldaar ook het taxatieverslag van de onroerende zaak kan worden gevonden, opdat de juistheid van de door de gemeente bepaalde prijs kan worden gecontroleerd.

2.3. Belanghebbende is op 31 maart 2014 in bezwaar gekomen tegen de WOZ-beschikking en de aanslag OZB.

2.4. Op 28 augustus 2014 heeft een hoorzitting plaatsgevonden waarbij de gemachtigde van belanghebbende een pleitnota heeft voorgelezen. In die pleitnota is onder meer opgenomen:

"Ik verzoek u bij de uitspraak op bezwaar mij de kavelwaarde en de opbouw hiervan bekend te maken, (de door u gehanteerde grondstaffel).

Dit op basis van artikel 40 Wet WOZ.

Daarnaast verzoek ik u deze gegevens ook te verstrekken van de door u gehanteerde vergelijkingsobjecten in deze procedure."

2.5. Bij de uitspraken op bezwaar is de gevraagde grondstaffel door de Heffingsambtenaar niet verstrekt.

2.6. Op 23 december 2014 heeft belanghebbende beroep ingesteld tegen de uitspraken op bezwaar. Belanghebbende verzoekt wederom om de grondstaffel. Voor zover van belang schrijft belanghebbende:

"Artikel 40 Wet WOZ

Tijdens de hoorzitting heb ik om de kavelwaarde van de nu in het geding zijnde object gevraagd. En tevens heb ik gevraagd om de kavelwaarde en de opbouw daarvan, van de vergelijkingsobjecten. Beiden worden door de gemeente Waalwijk niet verstrekt. Naar mijn mening een onjuiste uitleg van artikel 40 Wet WOZ. De gemeente stelt namelijk dat op basis van artikel 40 Wet WOZ alleen een taxatieverslag verstrekt hoeft te worden. In artikel 40 Wet WOZ wordt in lid 1 en lid 2 het volgende gesteld: "Op verzoek kan het waardegegeven van een bepaalde onroerende zaak door de in artikel 1, tweede lid, bedoelde gemeenteambtenaar worden verstrekt aan een ieder die kan aantonen uit hoofde van de belastingheffing te zijnen aanzien een gerechtvaardigd belang te hebben bij de verkrijging daarvan. De in artikel 1, tweede lid, bedoelde gemeenteambtenaar verstrekt uitsluitend aan degene te wiens aanzien een beschikking is genomen, op verzoek een afschrift van de gegevens die ten grondslag liggen aan de vastgestelde waarde."

Hier staat niet dat er een enkel een taxatieverslag geleverd hoeft te worden, er staat dat de

heffingsambtenaar op verzoek een afschrift van de gegevens die ten grondslag liggen aan de vastgestelde waarde dient te verstrekken, en naar mijn mening is o.a. de kavelwaarde en de opbouw daarvan een relevant waardegegeven en dat weigert de gemeente Waalwijk te verstrekken."

2.7. Opgemaakt en getekend 28 januari 2015 heeft [A] (hierna: [A]), voornoemd, in opdracht van de gemeente een taxatierapport (hierna: het taxatierapport) opgemaakt. Het betreft een uitpandige opname. In het taxatierapport wordt opgemerkt dat de woning al eerder inpandig is opgenomen in het kader van het bezwaar voor het belastingjaar 2013.

2.7.1. Blijkens het taxatierapport is aan de onroerende zaak een waarde toegekend van € 470.000. Het taxatierapport bevat gegevens van de onroerende zaak, alsmede foto's van de onroerende zaak, de omgeving daarvan en telkens één foto en gegevens van de referentieobjecten [b-straat 1] te [Z], [c-straat 1] te [Q] en [d-straat 1] te [Q]. Bij het taxatierapport is een matrix (de oorspronkelijke matrix) gevoegd waarin voor de waardebepaling van de onroerende zaak en de referentieobjecten relevante gegevens zijn opgenomen. De grondstaffel maakt geen onderdeel uit van de matrix, en deze is ook niet later in de beroepsprocedure aan belanghebbende verstrekt.

2.8. Op 19 februari 2016 heeft belanghebbende incidenteel hoger beroep ingesteld tegen de uitspraak van de Rechtbank. Belanghebbende verzoekt opnieuw, onder andere verwijzend naar het beroepschrift, om de grondstaffel.

2.9. Na ontvangst van het incidentele hoger beroepschrift heeft de Heffingsambtenaar [A] verzocht de vastgestelde waarde opnieuw te beoordelen. [A] heeft haar bevindingen neergelegd in een nieuwe matrix (hierna: de nieuwe matrix), alwaar zij komt tot een waarde van € 508.776. In de nieuwe matrix, gevoegd als bijlage bij de schriftelijke zienswijze op het incidentele hoger beroep, zijn grondstaffels opgenomen die zijn gehanteerd in de waardegebieden waarin de onroerende zaak en de referentieobjecten zijn gelegen.

Rechtbank Zeeland-West-Brabant

2.2 Na ongegrondverklaring van belanghebbendes bezwaren door de Heffingsambtenaar, bij in één geschrift vervatte uitspraken van 11 november 2014, heeft de Rechtbank belanghebbendes daartegen gerichte beroep bij mondelinge uitspraak van 17 december 2015 ongegrond verklaard.

2.3 De Rechtbank heeft, voor zover thans van belang in cassatie, overwogen:[6]

2.4. De heffingsambtenaar verdedigt de vastgestelde waarde. Belanghebbende bepleit een waarde van € 369.000.

(…)

2.7. De rechtbank is van oordeel dat de heffingsambtenaar met het taxatierapport aannemelijk heeft gemaakt dat de vastgestelde waarde niet te hoog is. Zij overweegt daartoe als volgt.

2.7.1. De rechtbank is van oordeel dat het vergelijkingsobject [b-straat 1] (verkoopprijs € 495.000) het meest vergelijkbaar is met de pastoriewoning. Dit vergelijkingsobject is net als de pastoriewoning gelegen in [Z] en is verder gelet op de uitstraling, bouwkundige kwaliteit, onderhoud, voorzieningen en perceelsomvang het best vergelijkbaar met de pastoriewoning (de factoren voor bouwkundige kwaliteit en onderhoud zijn voor beide objecten vastgesteld op 2 respectievelijk 3).

2.7.2. [b-straat 1] heeft een verkoopprijs van € 495.000, gecorrigeerd door de taxateur naar € 481.518 in verband met de waardepeildatum. De vastgestelde waarde voor de

pastoriewoning is € 455.000. De vraag is of het verschil tussen deze waarde en de (gecorrigeerde) verkoopprijs groot genoeg is gelet op de verschillen tussen de pastoriewoning en het vergelijkingsobject. De verschillen zijn dat het vergelijkingsobject een groter kavel, een grotere garage en twee bergingen, alsmede een hoger voorzieningenniveau heeft, maar dat daartegenover staat dat de pastoriewoning een substantieel grotere inhoud en een dakkapel/koekoek heeft, en van een veel jonger bouwjaar is. Het taxatierapport maakt voldoende inzichtelijk hoe de taxateur met deze verschillen is omgegaan bij zijn taxatie. De rechtbank acht met het taxatierapport aannemelijk gemaakt dat voornoemde vraag bevestigend moet worden beantwoord. Daarbij verdient opmerking dat het taxatierapport de pastoriewoning zelfs op een hogere waarde (€ 470.000) taxeert dan de vastgestelde waarde, zodat het door belanghebbende aangevoerde verschil in waarde tussen een dakkapel en een koekoek niet tot een ander oordeel kan leiden. Met het taxatierapport heeft de heffingsambtenaar naar het oordeel van de rechtbank dan ook aannemelijk gemaakt dat vastgestelde waarde van € 455.000 niet te hoog is.

2.7.3. Hetgeen belanghebbende heeft aangevoerd, heeft onvoldoende bewijskracht tegenover het taxatierapport om tot een andere oordeel te komen. Belanghebbende heeft zijn standpunt dat de waarde € 369.000 bedraagt niet onderbouwd met een taxatierapport maar de motivering daarvan beperkt tot de verwijzing naar verkoopresultaten van de objecten [a-straat 2], [e-straat 1] en [f-straat 1], alle gelegen te [Z]. Het object [a-straat 2] komt echter niet voor vergelijking in aanmerking aangezien dit meer dan 1 jaar na de waardepeildatum is verkocht en een ander type woning betreft (woonboerderij). Wat betreft Van de [e-straat 1] en [f-straat 1] heeft belanghebbende onvoldoende inzichtelijk gemaakt wat de verschillen zijn ten opzichte van de pastoriewoning. en hoe met die verschillen rekening is gehouden om tot een lagere waarde dan € 455.000 te komen.

2.8 Gelet op het vorenstaande is de rechtbank van oordeel dat de waarde van de pastoriewoning en de daarop gebaseerde aanslag niet te hoog zijn vastgesteld.

(…)

4. Heeft de heffingsambtenaar terecht geweigerd de grondstaffels te verstrekken?

2.11. Belanghebbende heeft tijdens het horen verzocht om bij de uitspraak op bezwaar de gehanteerde grondstaffel bekend te maken, zulks met verwijzing naar artikel 40 van de Wet WOZ. De heffingsambtenaar heeft bij de uitspraak op bezwaar dit verzoek gemotiveerd afgewezen. Ook in beroep is de grondstaffel niet bekendgemaakt.

Belanghebbende heeft in zijn beroepschrift erover geklaagd dat de grondstaffel niet is overgelegd bij de uitspraak op bezwaar. Volgens belanghebbende geeft de heffingsambtenaar een onjuiste uitleg aan artikel 40 van de Wet WOZ.

Tijdens de zitting heeft de rechtbank aan de heffingsambtenaar voorgehouden dat de rechtbank, gelet op de strekking van het verzoek in de bezwaarfase en van de klacht in beroep, de klacht ook zal onderzoeken – dus met aanvulling van de rechtsgronden – in het licht van artikel 7:4 van de Awb. De heffingsambtenaar heeft daarop aangevoerd dat artikel 40 van de Wet WOZ een lex specialis is, en dat hij gelet op dat artikel niet verplicht is om de grondstaffel over te leggen. De rechtbank overweegt als volgt.

2.11.1. Op grond van artikel 7:4, tweede lid, van de Awb moet de heffingsambtenaar alle op de zaak betrekking hebbende stukken voorafgaand aan het horen ter inzage leggen voor belanghebbende gedurende ten minste een week. Het derde lid schrijft voor dat bij oproeping van het horen wordt vermeld waar en wanneer de stukken ter inzage zullen liggen. Het vierde lid bepaalt dat belanghebbende van de stukken tegen vergoeding van ten hoogste de kosten afschriften verkrijgen.

2.11.2. De heffingsambtenaar heeft ter zitting desgevraagd verklaard dat de grondstaffel is gebruikt voor de vaststelling van de waarde bij de (primaire) beschikking. De grondstaffel is derhalve een stuk dat van belang is geweest voor de primaire besluitvorming, en is daardoor een op de zaak betrekking hebbend stuk zoals bedoeld in artikel 7:4, tweede lid, van de Awb (vgl. HR 20 maart 2009, ECLI:NL:HR:2009:BH6420).

Dit betekent, gelet op hetgeen in 2.11.1 is overwogen, dat (het stuk met) de grondstaffel ter inzage had moeten worden gelegd, en dat indien belanghebbende bij de inzage om een afschrift ervan had verzocht, hij dat had moeten krijgen.

2.11.3. Uit de stukken van het geding blijkt niet dat een inzage heeft plaatsgevonden, en blijkt evenmin dat er een oproeping voor de hoorzitting is geweest waarbij vermeld is waar en wanneer de stukken ter inzage zullen liggen. Gelet op dit laatste heeft de heffingsambtenaar in strijd met artikel 7:4, lid 3, van de Awb gehandeld. Onder deze omstandigheden kan aan belanghebbende niet worden tegengeworpen dat hij eerst tijdens het hoorgesprek om verstrekking van de grondstaffels heeft verzocht, en brengt artikel 7:4, vierde lid van de Awb mee dat ook in deze latere fase (dan vóór het horen) belanghebbende een afschrift van (het stuk met) de grondstaffel had moeten kunnen verkrijgen. Gelet hierop heeft de heffingsambtenaar ten onrechte geweigerd om een afschrift van (het stuk met) de grondstaffel aan belanghebbende te verstrekken. De omstandigheid dat belanghebbende bij zijn verzoek heeft gerefereerd aan artikel 40 van de Wet WOZ maakt dat niet anders.

2.11.4. In tegenstelling tot hetgeen de heffingsambtenaar heeft gesteld, kan uit artikel 40, tweede lid, van de Wet WOZ niet worden afgeleid dat deze bepaling een beperking vormt van de verplichtingen op grond van artikel 7:4 van de Awb. In artikel 40, tweede lid, van de Wet WOZ is bepaald dat de heffingsambtenaar uitsluitend aan degene te wiens aanzien een beschikking is genomen, op verzoek een afschrift verstrekt van de gegevens die ten grondslag hebben gelegen aan de vastgestelde waarde (beperkte openbaarheid). Het woord 'uitsluitend' heeft betrekking op 'aan degene te wiens aanzien een beschikking is genomen'. De bepaling houdt dus niet in dat aan diegene 'uitsluitend' bedoeld afschrift mag worden verstrekt (dus met uitsluiting van andere afschriften). Hoe dan ook, de tekst van artikel 40, tweede lid, van de Awb geeft geen aanknopingspunten voor de conclusie dat het artikel een beperking inhoudt van de verplichtingen in de bezwaarfase met betrekking tot op de zaak betrekking hebbende stukken. Ook de wetsgeschiedenis biedt daarvoor geen aanknopingspunten (vgl. Kamerstukken II, 1993/94, 22 885, nr. 10, p. 3). De omstandigheid dat artikel 40 van de Wet WOZ (wel) een lex specialis is ten opzichte van de Wet openbaarheid van bestuur (bijv. ABRvS 2 april 2014, ECLI:NL:RVS:2014:1161), kan aan het voorgaande niet afdoen. Ook de verwijzing van de heffingsambtenaar naar de uitspraak van de Rechtbank Zeeland-West-Brabant van 4 december 2014 (ECLI:NL:RBZWB:2014:8315), baat hem niet, reeds omdat in die procedure vaststond dat er geen grondstaffel was.

2.11.5. Gelet op het vorenstaande is de rechtbank van oordeel dat de weigering van de heffingsambtenaar om een afschrift van de grondstaffel(s) te verstrekken, leidt tot schending van artikel 7:4 van de Awb. De rechtbank ziet daarin geen aanleiding om de uitspraken op bezwaar te vernietigen (artikel 6:22 van de Awb), maar wel om te gelasten dat de heffingsambtenaar wordt veroordeeld in de proceskosten van de beroepsfase en het door belanghebbende betaalde griffierecht. Bij dat oordeel heeft de rechtbank in acht genomen dat de gemachtigde van belanghebbende ter zitting geloofwaardig heeft verklaard dat hij om de grondstaffels heeft verzocht om een inschatting te kunnen maken of hij al dan niet beroep wilde aantekenen, alsmede dat de heffingsambtenaar ook in beroep de grondstaffel niet heeft ingebracht.

2.11.6. Voor de te vergoeden proceskosten sluit de rechtbank aan bij het Besluit proceskosten bestuursrecht voor de door een derde beroepsmatig verleende rechtsbijstand, zijnde € 980 (1 punt voor het indienen van het beroepschrift en 1 punt voor de zitting, met een waarde

per punt van € 490 en een wegingsfactor 1).

2.11.7. Vanwege samenhang met de procedure AWB 14/7667 voor het griffierecht is alleen in AWB 14/7667 griffierecht geheven. Nu heffingsambtenaar in die procedure reeds is veroordeeld tot vergoeding van het griffierecht is veroordeling daarvan in de onderhavige procedure achterwege gebleven.

Gerechtshof 's-Hertogenbosch

2.4 De Heffingsambtenaar is, bij geschrift van 16 december 2015, tegen de uitspraak van de Rechtbank in hoger beroep gekomen bij het Hof. Bij (principaal) verweerschrift van 19 februari 2016 heeft belanghebbende tevens incidenteel hoger beroep ingesteld.

2.5 Het Hof heeft geoordeeld:[7]

Vraag I. Had de Heffingsambtenaar op verzoek de grondstaffel moeten verstrekken.

4.1. Sedert de bezwaarfase heeft belanghebbende, met een beroep op artikel 40 van de Wet WOZ, verzocht om een afschrift van de door de Heffingsambtenaar toegepaste grondstaffel, om zich zodoende een oordeel te vormen van de grondwaarde van de onroerende zaak en de referentieobjecten, en van de opbouw daarvan. De Heffingsambtenaar heeft de grondstaffel eerst verstrekt als onderdeel van de nieuwe matrix bij zijn zienswijze op het incidentele hoger beroep.

4.2. Indien de nieuwe matrix en de oorspronkelijke matrix met elkaar worden vergeleken, blijkt dat er voor wat betreft de grondwaardes verschillen zijn, welke zijn ontstaan doordat verschillende grondstaffels zijn gebruikt. Schematisch kunnen die verschillen als volgt in beeld worden gebracht:

Object	**Grondwaarde oorspronkelijk**	**Grondwaarde nieuw**	**Prijs per m2 oorspronkelijk**	**Prijs per m2 nieuw**
Onroerende zaak	€ 193.410	€ 230.410	€ 139	€ 166
[b-straat 1]	€ 237.000	€ 237.050	€ 126	€ 126
[c-straat 1]	€ 261.120	€ 261.120	€ 318	€ 318
[d-straat 1]	€ 214.520	€ 214.520	€ 281	€ 281

De grondwaarde en de prijs per m2 van de onroerende zaak zijn gewijzigd.

4.3. Hof 's-Hertogenbosch heeft op 28 juli 2016, nr. 15/00962, ECLI:NL:GHSHE:2016:3344 overwogen, welke overwegingen in de onderdelen 4.12.2 tot en met 4.12.5 het Hof voor zover hieronder opgenomen tot de zijne maakt:[8]

(...)

4.4. De Heffingsambtenaar heeft in het onderhavige geval eerst in hoger beroep de grondstaffel overgelegd. Bij de zienswijze van de Heffingsambtenaar op het incidentele hoger beroep van belanghebbende, gedagtekend 25 april 2016 en door het Hof ontvangen op 26 april 2016, is een nieuwe matrix gevoegd waarin de grondstaffel is vermeld. Deze nieuwe matrix wijkt voor wat betreft de grondwaarde af van de oorspronkelijke matrix (zie onderdeel 4.2). Het argument van de Heffingsambtenaar dat het geven van inzage in de grondstaffel in het onderhavige jaar op praktische problemen stuitte, wordt door het Hof verworpen. De Heffingsambtenaar dient zijn administratie zodanig in te richten dat de noodzakelijke bouwstenen voor de waardebepaling hieruit op doelmatige wijze zijn te destilleren. Dit klemt te meer nu de in onderdeel 4.2 weergegeven verschillen pas in hoger beroep aan het licht zijn gekomen. Het in artikel 40 van de Wet WOZ besloten beginsel dat belanghebbende in staat moet worden gesteld de waardebepaling te controleren, zou anders inhoudsloos worden. De omstandigheid dat artikel 40 van de Wet WOZ een lex specialis is ten opzichte van de Wet openbaarheid bestuur maakt dit, gelijk de Rechtbank terecht heeft beslist, niet anders. Gezien het voren overwogene is het Hof van oordeel dat de Heffingsambtenaar aan het verzoek van de gemachtigde om de grondstaffel te

verstrekken, tegemoet had moeten komen.

Vraag I dient derhalve bevestigend te worden beantwoord.

Vraag II. Is de waarde van de onroerende zaak te hoog vastgesteld?

4.5. Krachtens artikel 17, lid 1, van de Wet WOZ, wordt aan een onroerende zaak een waarde toegekend. Ingevolge het tweede lid van dit artikel wordt deze waarde bepaald op de waarde die aan de onroerende zaak dient te worden toegekend indien de volle en onbezwaarde eigendom daarvan zou kunnen worden overgedragen en de verkrijger de zaak in de staat waarin die zich bevindt, onmiddellijk en in volle omvang in gebruik zou kunnen nemen. Daarbij heeft als waarde te gelden de waarde in het economische verkeer. Dit is de prijs, die bij aanbieding ten verkoop op de voor die onroerende zaak meest geschikte wijze, na de beste voorbereiding, door de meest biedende gegadigde voor de onroerende zaak zou zijn betaald.

4.6. Op grond van artikel 18, lid 1, van de Wet WOZ wordt de waarde van een onroerende zaak bepaald naar de waarde die de zaak op de peildatum heeft naar de staat waarin de zaak op die datum verkeert. Daarbij geldt in het onderhavige geval als peildatum 1 januari 2013.

4.7. Ingevolge artikel 4, lid 1, van de Uitvoeringsregeling instructie waardebepaling Wet WOZ, wordt de in artikel 17, lid 2, van de Wet WOZ bedoelde waarde voor woningen bepaald door middel van een methode van systematische vergelijking met woningen waarvan marktgegevens beschikbaar zijn (zogeheten referentieobjecten).

4.8. De Heffingsambtenaar, op wie de bewijslast rust ter zake van de juistheid van de in geschil zijnde waarde van de onroerende zaak, beroept zich op het taxatierapport, waaruit een waarde volgt van € 470.000, alsmede, naar het Hof begrijpt, op de nieuwe matrix.

4.9. Naar het oordeel van het Hof heeft de Heffingsambtenaar met het taxatierapport en de nieuwe matrix de door hem verdedigde waarde van € 455.000 voldoende inzichtelijk gemaakt. In dit verband overweegt het Hof als volgt.

4.10. De Heffingsambtenaar heeft de waardebepaling van de onroerende zaak onderbouwd door middel van de waarde gegevens van drie referentiepanden, te weten [b-straat 1] gelegen te [Z], [c-straat 1] en [d-straat 1], beide gelegen te [Q]. De vergelijkingsobjecten zijn kort vóór of kort na de peildatum verkocht.

Anders dan de Rechtbank is het Hof van oordeel dat het vergelijkingsobject [b-straat 1] geen goed vergelijkingsobject vormt. Hoewel eveneens als de onroerende zaak behorend tot het waardegebied met nummer 9 is het pand [b-straat 1], naar belanghebbende onweersproken heeft gesteld ter zitting, ongeveer 10 kilometer gelegen van de onroerende zaak. Gelet op de uitstraling en het bouwjaar 1875 is het evenmin goed vergelijkbaar. De referentiepanden [c-straat 1] en [d-straat 1] zijn daarentegen, evenals de onroerende zaak, gebouwd in de jaren dertig van de vorige eeuw, hebben ook een piramidedak en zijn verder gelet op uitstraling, bouwkundige kwaliteit, onderhoud en voorzieningen, naar ’s-Hofs oordeel, goed vergelijkbaar met de onroerende zaak. Daarnaast is rekening gehouden met de onderlinge verschillen zoals verschillen in oppervlakte van de verschillende kavels (afnemend grensnut) en inhoud van de woningen. Daar doet niet aan af dat deze referentiepanden, naar partijen ter zitting eenparig stelden, in een andere kern liggen dan de onroerende zaak alsmede behoren tot een waardegebied met nummer 102. Beide kernen zijn, slechts gescheiden door de A59, immers tegen elkaar aangebouwd. Met het verschil in ligging is voorts in voldoende mate rekening gehouden met behulp van de grondstaffel. De bijzonder gunstige ligging van het pand [c-straat 1] komt tot uitdrukking in de verhoging van de waarde van de grond met 20%.

4.11. De gemachtigde van belanghebbende heeft ter zitting in zijn reactie op het stuk van de Heffingsambtenaar van 25 april 2016 gesteld dat het feit dat de gemeente de grond niet

consequent taxeert voor hem mede een reden is geweest om de toegepaste grondstaffel op te vragen. De gemeente blijft de grondwaarde maar wijzigen, aldus de gemachtigde van belanghebbende. De gemachtigde van belanghebbende geeft aan om deze reden er niet op te kunnen vertrouwen dat de getaxeerde grondwaarde juist is. De toelichting ter zitting door de Heffingsambtenaar dat het verschil in de kavelwaarde van de onroerende zaak in het taxatierapport van € 193.410 en in de nieuwe matrix van € 230.410 berust op een aanvankelijk foutieve grondstaffel die is toegepast ter zake van de onroerende zaak, acht het Hof aannemelijk. Het staat de Heffingsambtenaar voorts in beginsel vrij om in een latere fase van het geding de door hem verdedigde waarde met behulp van een herziene matrix te ondersteunen. Daarbij wijst het Hof erop dat de door de taxateur in de nieuwe matrix vastgestelde waarde van € 508.776 ruimschoots is gelegen boven de beschikte waarde van €455.000.

4.12. Belanghebbende heeft voorts gesteld dat – kort gezegd – de Heffingsambtenaar onvoldoende rekening heeft gehouden met de onderhoudstoestand en gebreken van de onroerende zaak. Belanghebbende verwijst daartoe naar een rapportage van de Monumentenwacht alsmede naar een meerjarig onderhoudsplan dat door een door het College van Kerkrentmeesters opgerichte onderhoudscommissie is opgesteld. Uit laatst vermeld plan blijkt, aldus belanghebbende, dat de komende jaren € 90.120 moet worden geïnvesteerd om de onroerende zaak weer op orde te krijgen. De Heffingsambtenaar heeft deze stelling gemotiveerd weersproken onder meer door te wijzen op diens bevindingen naar aanleiding van een inpandige opname die heeft plaatsgevonden op 27 augustus 2013 en door fotomateriaal te overleggen van de onroerende zaak.

4.13. Het Hof verwerpt voornoemde grief. Met hetgeen belanghebbende heeft gesteld heeft zij weliswaar aannemelijk gemaakt dat ten aanzien van de keuken en badkamer sprake is van een zekere gedateerdheid en dat enig schilder- en voegwerk noodzakelijk is, doch daarmee heeft de Heffingsambtenaar bij de waardebepaling met het hanteren van de factor 3 voor onderhoud en de factor 2 voor voorzieningen in voldoende mate rekening gehouden. Een en ander komt in de nieuwe matrix in voldoende mate tot uitdrukking in de gehanteerde lagere m3-prijs van de onroerende zaak ten opzichte van die van de [c-straat 1] en [d-straat 1].

4.14. Gelet op het voren overwogene is het Hof van oordeel dat de Heffingsambtenaar aannemelijk heeft gemaakt dat de waarde van de onroerende zaak, alsmede de daarop gebaseerde aanslag, door de Heffingsambtenaar niet te hoog zijn vastgesteld.

Vraag II dient derhalve ontkennend te worden beantwoord.

4.15. Het ten onrechte weigeren om de grondstaffel te verstrekken vormt voor het Hof aanleiding de Heffingsambtenaar te veroordelen in de proceskosten van de hoger beroepsfase. Belanghebbende is immers door de weigerachtige houding van de Heffingsambtenaar moeten blijven procederen tot in hoger beroep om een gefundeerd oordeel te kunnen vormen over de door de Heffingsambtenaar vastgestelde waarde om aan de hand van de ontvangen gegevens in alle redelijkheid te kunnen inschatten of hij al dan niet (hoger) beroep zou instellen. Pas na ontvangst van de zienswijze op het incidentele hoger beroep, met daarbij gevoegd de nieuwe matrix met de grondstaffel en de definitieve, gewijzigde, cijfers betreffende de grondwaardes van de onroerende zaak en de referentieobjecten, kon belanghebbende de hiervoor vermelde inschatting maken.

Slotsom

4.16. De slotsom is dat zowel het hoger beroep als het incidenteel hoger beroep ongegrond zijn.

2.6 Kats heeft hierbij geannoteerd in NTFR:[9]

Een uitspraak om des keizers baard, zo lijkt het op het eerste gezicht. Toch is dat niet helemaal waar. Voor de tweede keer maakt Hof Den Bosch namelijk duidelijk dat met ‘gegevens die ten grondslag liggen aan de vastgestelde waarde’ niet slechts het bekende taxatieverslag wordt

bedoeld, maar *alle* gegevens die van belang kunnen zijn voor het controleren van de voor de onroerende zaak vastgestelde waarde. Dit in tegenstelling tot wat veel gemeenten veronderstellen.

Het taxatieverslag bevat, als het goed is, in ieder geval gegevens die aan de vastgestelde waarde ten grondslag liggen, maar vaak niet alle. Zijn er meer gegevens gebruikt en wordt daarom verzocht, dan dienen deze dus ook verstrekt te worden. Dit sluit mijns inziens goed aan bij de bedoeling van de wetgever. Ter zake meldt hij namelijk dat aan degene te wiens aanzien de beschikking is genomen op verzoek de aan die taxatie onderliggende gegevens ter inzage worden gegeven (Kamerstukken II, 1993-1994, 22 885, nr. 10, p. 3). Een geheel juiste conclusie dus wat mij betreft.

2.7 Boone heeft aangetekend in Belastingblad:[10]

Bij de openbaarheid van WOZ-informatie is er de afgelopen jaren een ontwikkeling waar te nemen die op het eerste gezicht wat tegenstrijdig lijkt. Aan de ene kant zijn vanaf 1 oktober 2016 de WOZ-waarden van alle woningen in Nederland volledig openbaar geworden (Besluit van 9 december 2015, Stb. 2015, 506, Belastingblad 2016/42). Voor informatie over de onderbouwing van de WOZ-waarde, geldt echter het tegenovergestelde. Die informatie valt niet onder de Wet openbaarheid van bestuur (hierna: Wob) en is daarmee niet openbaar, zo blijkt uit een reeks van uitspraken van de Afdeling bestuursrechtspraak van de Raad van State (zie voor een overzicht mijn beschouwing Openbaarheid van WOZ-waarden en WOZ-gegevens van woningen: de stand van zaken in Belastingblad 2014/319). Specifiek voor grondstaffels oordeelde de afdeling nog vrij recent dat deze niet onder de Wob vallen (RvS 16 november 2016, ECLI:NL:RVS:2016:3022, Belastingblad 2017/8, m.nt. J.P. Kruimel).

Informatie die gemeenten gebruiken om de WOZ-waarde te onderbouwen, is dus niet door iedereen op te vragen. Dat betekent gelukkig niet dat voor deze informatie een absolute geheimhouding geldt. Een belanghebbende die de WOZ-waarde van zijn eigen onroerende zaak aanvecht, kan namelijk wel inzage krijgen in deze informatie en wel op grond van art. 40 lid 2 Wet WOZ en eventueel (als het komt tot een bezwaarprocedure) art. 7:4 Awb. De bepaling van art. 40 lid 2 Wet WOZ geeft iedereen die van de gemeente een WOZ-beschikking heeft gekregen recht op een afschrift van de gegevens die ten grondslag liggen aan de WOZ-waarde. Gemeenten hebben lang gedacht dat het in dit wetsartikel (uitsluitend) zou gaan om een afschrift van het taxatieverslag. Zo ook de gemeente Waalwijk die dit standpunt in 2014 nog met succes verdedigde bij Rechtbank Zeeland-West-Brabant. De rechtbank volgde toen nog het standpunt van de heffingsambtenaar dat grondstaffels niet onder het bereik vallen van art. 40 Wet WOZ (Rb. Zeeland-West-Brabant 4 december 2014, ECLI:NL:RBZWB:2014:8315, Belastingblad 2015/108, m.nt. E.G. Borghols). Omdat het in de WOZ-waardering gaat om de eindwaarde, zo overwoog de rechtbank destijds, is een grondstaffel niet een gegeven dat ten grondslag ligt aan de vastgestelde waarde als bedoeld in art. 40 lid 2 Wet WOZ. Dat was toen. In de onderhavige procedure oordelen zowel rechtbank als hof dat een grondstaffel wel degelijk valt onder art. 40 Wet WOZ en dat de gemeente Waalwijk een afschrift van de staffel moet verstrekken dan wel inzage hierin moet geven (let wel: dit is dus iets anders dan openbaar maken). Het hof oordeelt dat art. 40 lid 2 Wet WOZ ziet op alle gegevens die van belang kunnen zijn voor het controleren van de WOZ-waarde (vgl. RvS 11 december 2013, ECLI:NL:RVS:2013:2326, Belastingblad 2014/75, m.nt. Noordermeer van Loo). Dit is een ruime omschrijving, hetgeen betekent dat gemeenten niet te kinderachtig moeten zijn bij de verstrekking van WOZ-gegevens. De informatieplicht is uiteraard ook niet onbegrensd. Gemeenten hoeven niet alles te verstrekken. Steeds moeten zij het belang van degene die verzoekt om gegevens te verstrekken, afwegen tegen de privacy van anderen, zo volgt ook uit deze hofuitspraak.

3 Het geding in cassatie

3.1 Het College heeft bij geschrift van 22 maart 2017 beroep in cassatie ingesteld. Belanghebbende

heeft met dagtekening 24 mei 2017 een verweerschrift ingediend. Het College heeft op dit verweerschrift gereageerd door middel van een conclusie van repliek. Belanghebbende heeft daarop een conclusie van dupliek ingediend.

Beroep in cassatie

3.2 Het College heeft in cassatie vier middelen aangevoerd.

3.3 Het eerste middel luidt:

De uitspraak is onvoldoende gemotiveerd doordat het Hof niet ingaat op de grief dat de grondstaffel niet is neergelegd in een stuk dat voor verstrekking in aanmerking komt

3.4 Het College heeft toegelicht:

Het Gerechtshof gaat uit van de veronderstelling dat de grondstaffel is neergelegd in een stuk dat voor verstrekking in aanmerking komt. Deze veronderstelling is evenwel niet juist en daarnaast ook onbegrijpelijk. In de motivering van het hoger beroep, aan het Hof toegezonden op 13 januari 2016, is uitvoerig uiteengezet dat de grondstaffel niet kan worden aangemerkt als een stuk in de zin van artikel 7:4 van de Awb. In dit kader werd het volgende opgemerkt:

"De waardebepaling zoals die in het kader van de Wet Waardering onroerende zaken (hierna: WOZ) geschiedt, vindt bij de taxatie in eerste aanleg geautomatiseerd plaats. Hiertoe worden door (nagenoeg uitsluitend externe) softwareontwikkelaars taxatiemodellen ontworpen die per gemeente worden gevuld met de relevante objectgegevens en de beschikbare verkoopinformatie. Op grond van modelmatige analyse worden in het taxatiemodel onder meer staffels berekend die kunnen worden gebruikt voor het bepalen van de grondwaarde van het te taxeren object, met inachtneming van het type object en de ligging van de onroerende zaak.

De grondstaffels zijn derhalve verwerkt in het taxatiemodel, welke programmatuur is neergelegd in een softwareprogramma van de ingeschakelde externe softwareleverancier. Deze staffels zijn dan ook niet neergelegd in een stuk, noch zijn zij uit de taxatiesoftware (in het onderhavige geval betreft dit het programma GeoTax 2.5) te destilleren.

Anders dan de rechtbank overweegt is de grondstaffel dus niet neergelegd in een stuk; van een op de zaak betrekking hebbend stuk zoals bedoeld in artikel 7:4 van de Awb is dan ook geen sprake. Evenmin is het de gemeente mogelijk deze grondstaffels uit de betreffende programmatuur te destilleren. Reeds om deze reden heeft de rechtbank niet kunnen oordelen dat sprake is van schending van dit wetsartikel. Dit geeft aanleiding om de uitspraak van de rechtbank op dit punt te vernietigen hetgeen eveneens moet gelden voor de door de rechtbank uitgesproken proceskostenveroordeling."

De vraag of sprake is van een stuk in de zin van artikel 7:4 van de Awb is van belang om te kunnen bepalen of terinzagelegging noodzakelijk is. Voornoemde grief, inhoudende dat van een voor terinzagelegging vatbaar stuk in de zin van artikel 7:4 van de Awb geen sprake is, is dus van belang voor de beantwoording van de vraag of sprake is van een verplichting tot terinzagelegging als bedoeld in dit artikel en of, wanneer van een dergelijke verplichting sprake is, men dan gehouden is om voorafgaand aan de hoorzitting alsnog niet bestaande stukken voor de terinzagelegging te vervaardigen. (Op deze laatste vraag wordt later in dit beroepschrift meer uitgebreid teruggekomen).

Het Hof is in de bestreden uitspraak echter op geen enkele manier ingegaan op de opgeworpen grief, zoals neergelegd in de motivering van het hoger beroepschrift van 13 januari 2016 (pagina's 2 en 3) en gaat zonder meer uit van de veronderstelling dat sprake is van een stuk als bedoeld in artikel 7:4 van de Awb. Gelet op de naar voren gebrachte grieven terzake had het Hof dit standpunt niet kunnen innemen zonder op deze grieven in te gaan en deze gemotiveerd te verwerpen. De omstandigheid dat het Hof in de bestreden uitspraak geen enkele overweging aan deze grieven wijdt, maakt de uitspraak naar de mening van het College onbegrijpelijk en

onvoldoende gemotiveerd. Een en ander geeft grond voor vernietiging van de uitspraak.

3.5 Het tweede middel luidt:

Bevestiging van de uitspraak van het hof creëert rechtsongelijkheid.

3.6 Het College heeft toegelicht:

De heersende jurisprudentie inzake de gegevensverstrekking die onder de werking van de Wet openbaarheid van bestuur (hierna: Wob) dient plaats te vinden (zie bijvoorbeeld de uitspraak van de Afdeling Bestuursrechtspraak van de Raad van State (ABRvS) van 5 juni 2013, nr. 201204362/1/A3, ECLI:NL:RVS:2013:CA2102) houdt in dat de Wob geen verplichting bevat om gegevens te vervaardigen die niet in bestaande documenten zijn neergelegd. Niet valt in te zien waarom een dergelijke redenering niet ook zou moeten worden gevolgd in het geval van een verzoek om verstrekking van gegevens waarop de Wob niet van toepassing is.

Voor een analoge toepassing van deze jurisprudentie bestaat naar de mening van het College alle aanleiding. Wanneer een verzoek om gegevens in het kader van artikel 40 van de Wet WOZ wordt gedaan zonder dat tevens bezwaar wordt gemaakt of beroep wordt ingesteld tegen een belastingaanslag, is in hoger beroep de ABRvS de bevoegde rechter (Gerechtshof Arnhem-Leeuwarden, 2 februari 2016, 15/00046, ECLI:NL:GHARL:2016:639). Deze legt als maatstaf voor de vraag of de gevraagde gegevens moeten worden verstrekt (mede) aan of de gevraagde gegevens al dan niet in bestaande documenten zijn neergelegd (ABRvS, 5 juni 2013, nr. 201204362/1/A3, ECLI:NL:RVS:2013:CA2102).

Wanneer het verzoek om gegevensverstrekking op basis van artikel 40 van de Wet WOZ wordt gedaan binnen het kader van een bezwaar- of beroepsprocedure tegen de waarde van een onroerende zaak, is de Wob niet op dit verzoek van toepassing maar moet het verzoek worden beoordeeld binnen het kader van de bezwaar- of beroepsprocedure tegen de belastingaanslag of de WOZ-beschikking (vgl. ABRvS 11 december 2013, nr. 201208181/1/A3, ECLI:NL:RVS:2013:2326 en ABRvS 2 april 2014, nr. 201308103/1/A3, ECLI:NL:RVS:2014:1161). In hoger beroep is dan het Gerechtshof bevoegd.

Wanneer de uitspraak van het Gerechtshof door Uw Raad wordt bevestigd, een uitspraak waarin het Hof kennelijk meent dat gegevens moeten worden verstrekt die niet in bestaande documenten zijn neergelegd, en hiermee tot leidende jurisprudentie wordt gemaakt, wordt rechtsongelijkheid gecreëerd. In dat geval heeft immers een verzoeker die om gegevens verzoekt buiten het kader van een belastingprocedure geen recht op verstrekking van gegevens die niet in documenten zijn vastgelegd (immers; bevoegde rechter is de ABRvS), terwijl een verzoeker die om gegevens verzoekt binnen het kader van een belastingprocedure wel recht heeft op verstrekking van deze gegevens (immers: bevoegde rechter is het Gerechtshof).

Een en ander is naar de mening van het College aanleiding de bestreden uitspraak te vernietigen.

3.7 Het derde middel luidt:

Ten onrechte overweegt het Hof dat de grondstaffel A) voorafgaand aan de hoorzitting ter inzage had moeten worden gelegd en B) valt onder de gegevens die op grond van artikel 40 van de Wet WOZ moeten worden verstrekt

3.8 Het College heeft toegelicht:

Voor zover Uw Raad oordeelt dat het voorgaande niet reeds noopt tot vernietiging van de bestreden uitspraak, en de grondstaffel wel moet worden aangemerkt als een op de zaak betrekking hebbend stuk in de zin van artikel 7:4 van de Awb, merkt het College nog het volgende op.

Het Hof overweegt in onderdeel 4.3 van de bestreden uitspraak (onder verwijzing naar een

eerdere uitspraak van hetzelfde Hof) dat de grondstaffel, als een op de zaak betrekking hebbend stuk, voorafgaand aan de hoorzitting voor belanghebbende gedurende tenminste een week voor belanghebbende ter inzage gelegd had moeten worden. Zoals in het voorgaande reeds is overwogen, kan de grondstaffel niet worden aangemerkt als een op de zaak betrekking hebbend stuk in de zin van artikel 7:4 van de Awb, aangezien de grondstaffel niet in een stuk was vastgelegd. Ook anderszins kan niet worden gesteld dat de grondstaffel voorafgaand aan de hoorzitting aan belanghebbende ter beschikking had moeten worden gesteld, aangezien namens belanghebbende eerst ter hoorzitting om verstrekking van de grondstaffel is verzocht. Het oordeel dat een niet bestaand stuk, waarom eerst tijdens de hoorzitting is verzocht, voorafgaand aan de hoorzitting ter inzage had moeten worden gelegd, is dan ook onbegrijpelijk.

Voorts miskent het Hof dat artikel 40 van de Wet WOZ een uitputtende regeling geeft voor openbaarmaking van de gegevens die aan de waardebepaling ten grondslag liggen (zie bijvoorbeeld ABRvS, 7 augustus 2013, 201208505/1/A3, ECLI:NL:RVS:2013:629), althans dat deze uitputtende regeling zich verzet tegen openbaarmaking langs de weg van artikel 7:4 van de Awb. Eveneens miskent het Hof dat de grondstaffel niet valt onder de gegevens die op grond van artikel 40 van de Wet WOZ aan belanghebbenden moeten worden verstrekt. Het artikel beoogt een beperkte openbaarmaking van gegevens (waaronder het taxatieverslag), aan de hand waarvan een belanghebbende de waarde van zijn woning kan controleren.

De grondstaffel valt niet onder deze gegevens en behoeft naar de mening van het College dus niet te worden verstrekt op grond van artikel 40 van de Wet WOZ. Een nadere beschouwing van de inhoud van dit artikel maakt zulks duidelijk. Het eerste lid van artikel 40 betreft de verstrekking van de waarde van andere onroerende zaken dan de getaxeerde onroerende zaak, welke waardegegevens op verzoek aan een belanghebbende kunnen worden verstrekt. De grondstaffel valt niet onder dit artikellid. Het tweede lid van artikel 40 heeft uitsluitend betrekking op verstrekking van het taxatieverslag van de te taxeren onroerende zaak en geeft geen verplichting om andere gegevens te verstrekken.

De ABRvS heeft in haar eerder genoemde uitspraak van 2 april 2014 (ECLI:NL:RVS:2014:1161) beslist dat uit de geschiedenis van de totstandkoming van artikel 40 van de Wet WOZ volgt dat de wetgever met deze bepaling een toegesneden regeling inzake openbaarmaking en geheimhouding heeft willen treffen voor de gegevens die worden gebruikt bij de waardevaststelling van woningen. De Afdeling heeft bij uitspraak van 11 augustus 2004 (ECLI:NL:RVS:2004:AQ6645) overwogen dat de wetgever als gegevens zoals bedoeld in artikel 40, tweede lid, van de Wet WOZ in de Memorie van Toelichting bij dit artikel nadrukkelijk het aan die WOZ-waarde ten grondslag liggende taxatieverslag heeft genoemd.

Artikel 40, tweede lid, van de Wet WOZ (in de destijds geldende nummering betrof het artikel 41, tweede lid) is aldus komen te luiden door een amendement van het Tweede Kamerlid Kamp (Kamerstukken II 1993/94, 22 885, nr. 22). In het wetsvoorstel (Kamerstukken II 1993/94, 22 885, nr. 2) was slechts sprake van een verplichting tot het geven van inzage. Het amendement is als volgt toegelicht:

"Met dit amendement wordt beoogd te bereiken dat degene te wiens aanzien een beschikking is genomen desgevraagd het taxatierapport in fotokopie toegestuurd krijgt".

Met betrekking tot het taxatieverslag geldt dat belanghebbenden, na ontvangst van de beschikking, dit via Internet kunnen raadplegen. Met het verslag wordt inzicht gegeven in de manier waarop de waarde tot stand is gekomen door vermelding van de vergelijkingsobjecten en de verkoopgegevens waarmee ter bepaling van de waarde is vergeleken. Dit op Internet te raadplegen taxatieverslag is een uitwerking van artikel 6 van het Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ en is vastgelegd in artikel 6 van de Uitvoeringsregeling instructie waardebepaling Wet WOZ. Het taxatieverslag voldoet daarmee dan ook aan de volgens de Wet WOZ te verstrekken gegevens.

Uit het bovenstaande volgt dat de bijzondere openbaarmakingsregeling van artikel 40, tweede lid van de Wet WOZ slechts verplicht tot verstrekking van het taxatieverslag. De bijzondere openbaarmakingsregeling van artikel 40 van de Wet WOZ verhindert naar de mening van het college dat met betrekking tot gegevens die op grond van deze uitputtende regeling niet behoeven te worden verstrekt, langs de weg van artikel 7:4 van de Awb alsnog kan worden bereikt dat deze gegevens aan belanghebbenden bekend moeten worden gemaakt.

De ABRvS heeft in haar reeds eerder aangehaalde uitspraak van 2 april 2014 (ECLI:NL:RVS:2014:1161), met betrekking tot de vraag of de algemene regeling zoals neergelegd in de Wob de openbaarmakingsregeling van artikel 40 van de Wet WOZ opzijzet, overwogen dat dit niet het geval is omdat "het verstrekken van de verzochte gegevens met toepassing van de Wob (...) afbreuk (zou) doen aan de strekking van de in artikel 40 van de Wet Woz neergelegde regeling die een vorm van 'beperkte openbaarheid' beoogt." Niet valt in te zien waarom ten aanzien van artikel 7:4 van de Awb een andere redenering zou moeten worden gevolgd.

Voorts overweegt het Hof naar de mening van het College ten onrechte dat uit artikel 40 van de Wet WOZ volgt dat een belastingplichtige recht heeft op "alle gegevens die van belang kunnen zijn voor het controleren van de voor zijn onroerende zaak vastgestelde waarde, een en ander met inachtneming van het belang van de privacy van anderen" (overweging 4.3, onder 4.12.3). Deze overweging is naar de mening van het College onjuist. Belastingplichtigen hebben op grond van dit artikel immers geen recht op verstrekking van alle gegevens die voor het controleren in voorbedoelde zin van belang kunnen zijn, maar slechts op verstrekking van die gegevens die de heffingsambtenaar op grond van artikel 40 van de Wet WOZ verplicht is te verstrekken. Deze gegevens, die verplicht moeten worden verstrekt, omvatten een kleinere kring van gegevens dan "alle gegevens" zoals omschreven door het Hof. Het uitgangspunt van waaruit het Hof beoordeelt of aan de verplichting tot informatievoorziening is voldaan, is derhalve in strijd met artikel 40 van de Wet WOZ en daarmee onjuist.

Gelet op het voorgaande is het College van oordeel dat, zo al moet worden aangenomen dat de grondstaffeling als stuk heeft te gelden, deze staffeling niet voorafgaand aan de hoorzitting ter inzage hoeft te worden gelegd. Voorts geldt dat de grondstaffel, zo deze al aan te merken valt als een stuk, niet valt onder de gegevens die op grond van artikel 40 van de Wet WOZ moeten worden verstrekt. De andersluidende uitspraak van het Hof komt naar de mening van het College dan ook voor vernietiging in aanmerking. Dit impliceert tevens dat geen recht bestaat op een proceskostenvergoeding zoals door het Hof is uitgesproken, en evenmin op vergoeding van het griffierecht.

3.9 Het vierde middel luidt:

Ten onrechte overweegt het Hof dat eerst in hoger beroep voldoende inzicht is geboden in de totstandkoming van de waarde.

3.10 Het College heeft toegelicht:

In onderdeel 4.15 van de bestreden uitspraak overweegt het Hof dat belanghebbende door de weigerachtige houding van de heffingsambtenaar tot in hoger beroep heeft moeten doorprocederen om een gefundeerd oordeel te kunnen vormen van de door de heffingsambtenaar vastgestelde waarde.

Naar de mening van het College is deze overweging niet juist. In de eerste plaats heeft de heffingsambtenaar reeds in de bezwaarfase alsook in de beroepsfase alle gegevens verstrekt die volgens het voornoemde modeltaxatieverslag als uitwerking van artikel 6 van het Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ geleverd dienen te worden, en hiermee voldaan aan de op hem rustende plicht tot informatievoorziening zoals neergelegd in artikel 40 van de Wet WOZ. Van een weigerachtige houding is derhalve geen sprake

geweest; slechts van een juiste wetstoepassing waarbij de verstrekking van niet bestaande gegevens of van gegevens tot verstrekking waarvan belanghebbende vanuit de bijzondere openbaarmakingsregeling van artikel 40 van de Wet WOZ geen recht heeft, achterwege is gelaten. Daarnaast is, zo blijkt uit hetgeen hierboven reeds is aangevoerd, de wetgever uitdrukkelijk van oordeel dat als gegevens, bedoeld in artikel 40, tweede lid, van de Wet WOZ, hebben te gelden de gegevens die zijn neergelegd in het taxatieverslag. Dat het Hof impliciet oordeelt dat de heffingsambtenaar niet aan zijn verplichting tot het leveren van informatie heeft voldaan, hoewel de gegevens zijn verstrekt die volgens de bijzondere openbaarmakingsregeling van artikel 40 van de Wet WOZ beschikbaar moeten worden gesteld, is naar de mening van het College dan ook onbegrijpelijk. Dit geeft aanleiding de uitspraak van het Hof te vernietigen.

Ten slotte vestigt het College Uw aandacht op het feit dat het ontbreken van een grondstaffel de Rechtbank niet heeft verhinderd om een oordeel over de vastgestelde waarde te vellen; een oordeel dat door het Hof wordt bevestigd. Naar de mening van het College is het onbegrijpelijk en ook tegenstrijdig dat het Hof enerzijds oordeelt dat het niet verstrekken van de grondstaffel, wat hier verder ook van zij, belanghebbende verhindert een juiste beoordeling van de juistheid van de waarde te maken, maar anderzijds kennelijk van oordeel is dat de Rechtbank tot een dergelijke beoordeling op basis van de beschikbare gegevens wel in staat is. Kennelijk is inzicht in de grondstaffel dus in het geheel niet benodigd om zich een oordeel omtrent de waarde van de onroerende zaak te kunnen vormen. Ook deze omstandigheid geeft naar het oordeel van het College aanleiding de bestreden uitspraak te vernietigen.

4 **Relevante wet- en regelgeving, wetsgeschiedenis, parlementaire behandeling, jurisprudentie en literatuur**

Wetgeving Algemene wet bestuursrecht [11]

4.1 Artikel 7:4 van de Awb luidt:

1 Tot tien dagen voor het horen kunnen belanghebbenden nadere stukken indienen.

2 Het bestuursorgaan legt het bezwaarschrift en alle verder op de zaak betrekking hebbende stukken voorafgaand aan het horen gedurende ten minste een week voor belanghebbenden ter inzage.

3 Bij de oproeping voor het horen worden belanghebbenden gewezen op het eerste lid en wordt vermeld waar en wanneer de stukken ter inzage zullen liggen.

4 Belanghebbenden kunnen van deze stukken tegen vergoeding van ten hoogste de kosten afschriften verkrijgen.

5 Voor zover de belanghebbenden daarmee instemmen, kan toepassing van het tweede lid achterwege worden gelaten.

6 Het bestuursorgaan kan, al dan niet op verzoek van een belanghebbende, toepassing van het tweede lid voorts achterwege laten, voor zover geheimhouding om gewichtige redenen is geboden. Van de toepassing van deze bepaling wordt mededeling gedaan.

7 Gewichtige redenen zijn in ieder geval niet aanwezig, voor zover ingevolge de Wet openbaarheid van bestuur de verplichting bestaat een verzoek om informatie, vervat in deze stukken, in te willigen.

8 Indien een gewichtige reden is gelegen in de vrees voor schade aan de lichamelijke of geestelijke gezondheid van een belanghebbende, kan inzage van de desbetreffende stukken worden voorbehouden aan een gemachtigde die hetzij advocaat hetzij arts is.

4.2 Artikel 8:1 van de Awb luidt:

Een belanghebbende kan tegen een besluit beroep instellen bij de bestuursrechter.

Wetgeving Wet waardering onroerende zaken (hierna: Wet WOZ)

4.3 Artikel 30, eerste lid, van de Wet WOZ luidt:

Met betrekking tot de waardebepaling en de waardevaststelling ingevolge de hoofdstukken III en IV zijn de artikelen 1, derde lid, 5, eerste lid, tweede volzin, 22j tot en met 30, 47 , 49 tot en met 51, 52a, 53a, 54 en 56 tot en met 60 van de Algemene wet inzake rijksbelastingen van overeenkomstige toepassing. Met betrekking tot natuurlijke personen die een bedrijf of zelfstandig een beroep uitoefenen, alsmede lichamen, is voorts artikel 52, vierde en vijfde lid, en - voor zoveel het betreft het bewaren van gegevensdragers - zesde lid, van de Algemene wet inzake rijksbelastingen van overeenkomstige toepassing.

4.4 Artikel 40 van de Wet WOZ luidde in 2014:[12]

1 Op verzoek kan het waardegegeven van een bepaalde onroerende zaak door de in artikel 1, tweede lid, bedoelde gemeenteambtenaar worden verstrekt aan een ieder die kan aantonen uit hoofde van de belastingheffing te zijnen aanzien een gerechtvaardigd belang te hebben bij de verkrijging daarvan.

2 De in artikel 1, tweede lid, bedoelde gemeenteambtenaar verstrekt uitsluitend aan degene te wiens aanzien een beschikking is genomen, op verzoek een afschrift van de gegevens die ten grondslag liggen aan de vastgestelde waarde.

3 Bij of krachtens algemene maatregel van bestuur kunnen regels worden gesteld met betrekking tot de vergoeding die in rekening kan worden gebracht ter zake van de verstrekking van een waardegegeven aan derden.

Wetgeving AWR

4.5 Artikel 1 van de AWR luidt, voor zover in cassatie van belang:

1. De bepalingen van deze wet gelden in Nederland bij de heffing van rijksbelastingen, de heffing van belastingrente, revisierente en bestuurlijke boeten welke ingevolge de belastingwet kunnen worden vastgesteld of opgelegd, alsmede bij de uitvoering van de basisregistratie inkomen, een en ander met uitzondering van de belastingen voor zover voor een belanghebbende na een door de inspecteur gedane uitspraak op bezwaar met betrekking tot deze belastingen beroep openstaat bij het Gerecht in eerste aanleg van Bonaire, Sint-Eustatius en Saba, bedoeld in hoofdstuk VIII van de Belastingwet BES.

2 Onder rijksbelastingen worden verstaan belastingen welke van rijkswege door de rijksbelastingdienst worden geheven.

3 Met betrekking tot de heffing van rijksbelastingen blijven titel 5.2 en afdeling 10.2.1 van de Algemene wet bestuursrecht buiten toepassing.

4.6 Artikel 2, eerste lid, onder a, van de AWR luidt:

1 Deze wet verstaat onder:

a. belastingwet: zowel deze wet als andere wettelijke bepalingen betreffende de heffing van de onder artikel 1 vallende belastingen;

4.7 Artikel 26 van de AWR luidt:

1 In afwijking van artikel 8:1 van de Algemene wet bestuursrecht kan tegen een ingevolge de belastingwet genomen besluit slechts beroep bij de bestuursrechter worden ingesteld, indien het betreft:

a. een belastingaanslag, daaronder begrepen de in artikel 15 voorgeschreven verrekening, of

b. een voor bezwaar vatbare beschikking.

2 De voldoening of afdracht op aangifte, dan wel de inhouding door een inhoudingsplichtige, van een bedrag als belasting wordt voor de mogelijkheid van beroep gelijkgesteld met een voor bezwaar vatbare beschikking van de inspecteur. De wettelijke voorschriften inzake bezwaar en beroep tegen zodanige beschikking zijn van overeenkomstige toepassing, voorzover de aard van de voldoening, de afdracht of de inhouding zich daartegen niet verzet.

Wetsgeschiedenis I Wet WOZ

4.8 Uit de Memorie van Toelichting bij het wetsvoorstel Wet Woz is het volgende opgenomen over artikel 40 van de Wet WOZ:[13], [14]

In het RAVI-rapport wordt openbaarheid van het waardegegeven aanbevolen om de objectiviteit van de waardebepaling te waarborgen. De belastingbetaler moet de waarde van zijn onroerende zaak kunnen vergelijken met die van andere (vergelijkbare) onroerende zaken. In het kabinetsstandpunt naar aanleiding van het rapport wordt dit standpunt onderschreven, zij het dat voor bedrijfsmatig gebruikte onroerende zaken een terughoudend standpunt ter zake wordt ingenomen. In het door de werkgroep WOG uitgebrachte rapport over de organisatie van de waardebepaling is met betrekking tot het vraagstuk van de openbaarheid opgemerkt, dat daaraan in het wetgevingsproces nader aandacht zal worden geschonken. Binnen genoemde werkgroep is onderzoek gedaan naar de aspecten die aan openbaarheid van het waardegegeven verbonden zijn. Daarbij is bezien of bepaalde doelstellingen van de wet - een vergroting van de rechtsgelijkheid en rechtszekerheid voor de burger - met zich meebrengen dat het waardegegeven van een onroerende zaak in beginsel openbaar moet zijn. Het waardegegeven is echter naar onze mening zozeer verweven met de strikt individuele financiële en fiscale positie van de belanghebbende dat geheimhouding voorop dient te staan. Dit geldt niet alleen de particuliere burger maar evenzeer, en uit allerlei economische overwegingen wellicht nog sterker, de gegevens betreffende zaken die onderdeel vormen van een ondernemingsvermogen. Wel is de werkgroep van mening dat een uitzondering dient te worden gemaakt - in zoverre zou men kunnen spreken van beperkte openbaarheid - indien het redelijk is dat aan derden het waardegegeven ter beschikking wordt gesteld. Dit met het oog op de mogelijkheid voor de burger om in de praktijk te kunnen toetsen of het college van burgemeester en wethouders het gelijke ook gelijk behandeld - in casu gewaardeerd - heeft. Het waardegegeven van een aan een ander toe te rekenen object zal uitsluitend gegeven kunnen worden onder de voorwaarde dat die derde kan aantonen er met het oog op de heffing van belasting te zijnen aanzien een gerechtvaardigd belang bij te hebben om kennis te dragen van dit gegeven. Het aan het waardegegeven onderliggende taxatierapport is alleen ter inzage van de eigenaar of beperkt gerechtigde of van de gebruiker en - uiteraard - de rechterlijke macht. De werkgroep heeft bij de keuze voor dit stelsel het belang dat wordt gehecht aan de mogelijkheid van controle door de burger in verband met de rechtszekerheid en rechtsgelijkheid, zeer zwaar laten wegen. Doorzichtigheid naar de burger toe is immers ook een van de doelstellingen die met het onderhavige wetsvoorstel beoogd worden. Ook pleidooien voor openbaarheid uit de belastingadviespraktijk heeft zij in haar beschouwingen betrokken (…)

Zoals opgemerkt, hebben onder andere de algemene beginselen van rechtsgelijkheid en rechtszekerheid ten grondslag gelegen aan de totstandkoming van deze wet. Mede daarom is in het voorgestelde artikel 41 bepaald dat het waardegegeven door het college van burgemeester en wethouders aan derden verstrekt mag worden. Echter, gezien het feit dat het waardegegeven een element is dat een rol speelt bij de vaststelling van iemands belastingschuld, en daarmee als privacygevoelig geldt, hebben wij geconcludeerd dat openbaarheid geen regel zal kunnen zijn. Daarmee is tevens de gedachte verworpen bij wet openbare registers in te stellen, die voor iedereen toegankelijk zijn. Gekozen is voor de variant, waarbij een recht tot kennisneming wordt geregeld voor degene die kan aantonen een gerechtvaardigd belang te hebben bij de kennisname van het waardegegeven. Dit recht strookt met de doelstelling van de wet om de waardebepaling van onroerende zaken ten behoeve van de heffing van belastingen voor de burger doorzichtiger en voor controle vatbaar te maken. Opgemerkt zij dat dit recht moet worden onderscheiden van het inzagerecht dat op grond van de Wet persoonsregistraties aan belanghebbende zelf toekomt. Om te voorkomen dat er «schaduw» registraties ontstaan of op niet met de Wet WOZ strokende wijze gebruik wordt gemaakt van het waardegegeven, wordt voor de kennisneming door derden de eis gesteld van een gerechtvaardigd belang in verband met belasting die van hem wordt geheven. Daarbij moet gedacht worden aan iemand die de waarde die aan zijn pand is toegekend wil vergelijken met de waarde die is vastgesteld ter zake van een vergelijkbaar pand. Aangezien de vrager van het waardegegeven alleen een (fiscaal) belang kan hebben indien hij zelf een soortgelijk pand gebruikt of in eigendom heeft - en dit is een van de elementen waaraan de gemeente zal moeten toetsen alvorens een waardegegeven te verstrekken aan een derde - is het

niet mogelijk de waarde van willekeurig welke onroerende zaak op te vragen. Voorts wordt met de beperking recht gedaan enerzijds aan het beginsel van de bescherming van de persoonlijke levenssfeer en anderzijds aan het recht van de belastingplichtige tot controle op een juiste waardevaststelling van zijn onroerende zaak. Met betrekking tot bedrijfsmatig gebruikte onroerende zaken, waarvan het waardegegeven een concurrentiegevoelig gegeven zou kunnen zijn, menen wij eveneens met de eis van een gerechtvaardigd belang de vrees voor oneigenlijk gebruik van het waardegegeven te hebben weggenomen. Er zij nogmaals op gewezen dat de onderliggende gegevens nooit aan derden worden verstrekt. De zakelijk gerechtigde en de gebruiker van een onroerende zaak hebben wel het recht tot inzage van deze gegevens. De gemeente zal zorgvuldig moeten toetsen of een gerechtvaardigd belang bestaat. Tegen een weigering om het waardegegeven te verstrekken kan de normale procedure worden ingesteld voor geschillen tussen burgers en openbaar bestuur. De hiervoor uiteengezette keuze om - wat betreft in individuele gevallen genomen beslissingen inzake voor de heffing van onderscheidene belastingen geldende waarderingen - het primaat te verlenen aan de vertrouwelijkheid die alom als wenselijk wordt ervaren en daarop slechts die uitzonderingen toe te laten die door een effectuering in de praktijk van het gelijkheidsbeginsel wordt gevergd, staat geheel los van de werkingssfeer van de Wet openbaarheid van bestuur. Ingevolge die wet zullen immers eventuele algemene beleidsregels voor zover die niet zijn neergelegd in gepubliceerde voorschriften en dergelijke kunnen worden gevraagd. Het waardegegeven behoort daar niet toe.

4.9 In de Tweede Nota van wijziging bij het wetsvoorstel Wet WOZ is de volgende toelichting gegeven:[15]

In artikel 41 is thans een bepaling opgenomen dat de aan de taxatie onderliggende gegevens aan degene te wiens aanzien de beschikking is genomen op verzoek ter inzage worden gegeven. Hiermee is het opschrift van hoofdstuk VII in overeenstemming gebracht.

4.10 Op 22 juni 1994 heeft Tweede Kamerlid G.J. Kamp het volgende amendement voorgesteld,[16] welk amendement is aangenomen:[17]

In artikel 41 [40], tweede lid, wordt «geeft uitsluitend aan degene te wiens aanzien een beschikking is genomen, op verzoek inzage in de gegevens die ten grondslag liggen aan de vastgestelde waarde» vervangen door: verstrekt uitsluitend aan degene te wiens aanzien een beschikking is genomen, op verzoek een afschrift van de gegevens die ten grondslag liggen aan de vastgestelde waarde.

Toelichting:

Met dit amendement wordt beoogd te bereiken dat degene te wiens aanzien een beschikking is genomen desgevraagd het taxatierapport in fotokopie toegestuurd krijgt.

Regelgeving

4.11 Artikel 2 van het Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet waardering onroerende zaken (hierna: Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ), luidt, voor zover in cassatie van belang:[18], [19]

Bij ministeriële regeling wordt een instructie vastgesteld waarin regels zijn neergelegd voor de onderbouwing en de uitvoering van de waardebepaling van onroerende zaken op de voet van de wet.

4.12 Artikel 6 van het Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet waardering onroerende zaken (hierna: Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ), luidt, voor zover in cassatie van belang:

De instructie bevat richtlijnen voor de onderbouwing van de taxatie ten behoeve van belanghebbenden in de vorm van een model-taxatieverslag. Het taxatieverslag bevat ten minste de objectaanduiding, de waarderelevante objectgegevens, de motivering van de individuele afwijking ten opzichte van de relevante marktgegevens en de getaxeerde waarde.

Jurisprudentie I Hoge Raad

4.13 Op 7 juni 2002 heeft de Hoge Raad geoordeeld:[20]

De stukken van het geding laten geen andere conclusie toe dan dat belanghebbende geen gebruik heeft gemaakt van het recht te worden gehoord voordat over zijn bezwaarschrift zou worden beslist, waardoor hij zichzelf de mogelijkheid van inzage in het belastingdossier welke art. 7:4 van de Algemene wet bestuursrecht biedt, heeft onthouden. Nu belanghebbende voorts in het geding voor het Hof bekend was met de bewijsvoering van de Inspecteur, maar uit 's Hofs uitspraak of de stukken van het geding niet blijkt dat hij met betrekking tot concrete stellingen van de Inspecteur heeft aangevoerd dat deze voortvloeiden of zouden kunnen voortvloeien uit onrechtmatig verkregen bewijs, valt niet in te zien in welk opzicht aan belanghebbende een eerlijk proces zou zijn onthouden. De klacht faalt derhalve.

4.14 De Hoge Raad heeft bij arrest van 21 april 2006 overwogen:[21]

Ten aanzien van belanghebbende is bij in één geschrift vervatte beschikkingen de waarde van de onroerende zaken a-straat 1 en 2 te Z voor het tijdvak 1 januari 2001 tot en met 31 december 2004 vastgesteld op ƒ 455.000 (€ 206.470) respectievelijk ƒ 365.000 (€ 165.630).

Na door belanghebbende daartegen gemaakt bezwaar heeft de heffingsambtenaar van de gemeente Winterswijk bij in één geschrift vervatte uitspraken de waarden nader vastgesteld op € 186.050 respectievelijk € 152.016.

Belanghebbende is tegen die uitspraken in beroep gekomen bij het Hof.

Het Hof heeft het beroep ongegrond verklaard. De uitspraak van het Hof is aan dit arrest gehecht.

(…)

3.1. Het door belanghebbende in de bezwaarfase gedane en voor het Hof herhaalde verzoek, om - naast de hem reeds verstrekte waardegegevens van de vergelijkingspanden - drie waardegegevens van vergelijkbare woningen uit de directe omgeving, drie waardegegevens van andere categorieën woningen en drie waardegegevens van woningen op andere locaties te verstrekken, is door de heffingsambtenaar niet gehonoreerd. In de procedure voor het Hof heeft de heffingsambtenaar in zijn verweerschrift aangegeven op het verzoek niet te zijn ingegaan omdat belanghebbendes verzoek geen door deze geselecteerde woningen bevatte.

3.2. Op grond van artikel 40 van de Wet WOZ kan de heffingsambtenaar op verzoek het waardegegeven van een bepaalde onroerende zaak verstrekken aan een ieder die kan aantonen uit hoofde van de belastingheffing te zijnen aanzien een gerechtvaardigd belang te hebben bij de verkrijging daarvan. Deze bepaling strekt ertoe dat de belastingplichtige bepaalde waardegegevens kan verkrijgen, waarover hij wenst te beschikken om te kunnen controleren of sprake is van een juiste waardevaststelling van zijn onroerende zaak. Gelet op de strekking en bewoordingen van dit artikel dient het verzoek om verstrekking van waardegegevens betrekking te hebben op bepaalde, door de belanghebbende aangewezen onroerende zaken. Dienovereenkomstig spreekt de door belanghebbende ingeroepen "instructie gerechtvaardigd belang" van de Waarderingskamer over "door de verzoeker geselecteerde woningen".

3.3. Belanghebbendes klacht te dezer zake berust uitsluitend op de stelling dat niet hijzelf maar de heffingsambtenaar een selectie van de objecten diende te maken. Die stelling faalt, gelet op het hiervoor in 3.2 overwogene.

4.15 In het arrest van 13 augustus 2010 heeft de Hoge Raad overwogen:[22]

5.1. Het eerste middel betoogt dat het Hof bij zijn hiervoor in onderdeel 4.2 weergegeven oordeel is uitgegaan van een onjuiste maatstaf voor de aard van de bewijslast en tevens de bewijsmiddelen niet of niet voldoende in hun onderlinge samenhang heeft bezien.

Het eerste onderdeel van het middel berust op een onjuiste lezing van 's Hofs uitspraak. Uit hetgeen het Hof in de onderdelen 5.1.1, tweede alinea, en 5.1.3 van zijn uitspraak overweegt, volgt dat het Hof ervan is uitgegaan dat belanghebbende aannemelijk moest maken dat de goederen op het tijdstip van de aanvaarding van de aangifte niet aanwezig waren. De

overwegingen van het Hof met betrekking tot het al dan niet voldaan hebben aan de aldus op belanghebbende rustende bewijslast, houden niet in dat het Hof in feite een zwaardere bewijslast dan 'aannemelijk maken' op belanghebbende heeft gelegd.

Met betrekking tot de klacht dat het Hof de bewijsmiddelen niet of niet voldoende in hun onderlinge samenhang heeft bezien, heeft te gelden dat het aan de rechter die over de feiten oordeelt, vrijstaat om van het beschikbare bewijsmateriaal tot het bewijs te bezigen wat hem uit een oogpunt van betrouwbaarheid dienstig voorkomt en datgene terzijde te stellen wat hij voor het bewijs van geen waarde acht. Deze beslissing behoeft geen motivering, behoudens in bijzondere gevallen, waarvan in dit geval geen sprake is. Voor het overige berusten 's Hofs oordelen op de aan het Hof voorbehouden waardering van de bewijsmiddelen.

Op grond van het vorenstaande faalt het middel.

4.16 Op 20 december 2013 heeft de Hoge Raad geoordeeld:[23]

4.2.5. Opmerking verdient nog dat in een geval als het onderhavige, waarin geen bezwaar openstaat tegen een ingevolge de belastingwet genomen besluit of een daarmee gelet op het hiervoor in 4.2.3 overwogene gelijk te stellen besluit, en daartegen niettemin bezwaar wordt gemaakt, de inspecteur dat bezwaar bij zijn uitspraak niet-ontvankelijk dient te verklaren. De belastingrechter is wel bevoegd kennis te nemen van het beroep tegen de uitspraak die de inspecteur op een zodanig bezwaar heeft gedaan. Bij zijn oordeel over de gegrondheid van een dergelijk beroep dient de belastingrechter te beoordelen of inderdaad sprake is van een ingevolge de belastingwet genomen besluit waartegen geen bezwaar openstaat. De Hoge Raad verwijst hiertoe naar zijn arrest van 1 maart 2000, nr. 35041, ECLI:NL:HR:2001:AA4984, BNB 2000/171. Dat arrest had weliswaar betrekking op de tot 1 januari 2005 geldende tekst van artikel 26 AWR, maar met de wijziging van die bepaling per 1 januari 2005 is geen inhoudelijke wijziging op dit punt beoogd (zie HR 5 maart 2010, nr. 08/01707, ECLI:NL:HR:2010:BL6423, BNB 2010/167).

4.17 De Hoge Raad heeft bij arrest van 25 november 2016 overwogen:[24]

2.5.1. De klachten betogen voorts dat het Hof zich ten onrechte onbevoegd heeft verklaard van het hoger beroep kennis te nemen.

2.5.2. Artikel 40 Wet WOZ (tekst tot 1 oktober 2016) strekt ertoe dat degene te wiens aanzien een waardebeschikking is genomen bepaalde waardegegevens kan verkrijgen, waarover hij wenst te beschikken om de juistheid van die waardebeschikking te kunnen controleren (vgl. HR 21 april 2006, nr. 41185, ECLI:NL:HR:2006:AW2326, BNB 2006/231). De beoordeling van de vraag of een heffingsambtenaar naar aanleiding van een daartoe strekkend verzoek heeft voldaan aan zijn in artikel 40 Wet WOZ neergelegde verplichting tot openbaarmaking, dient plaats te vinden in de procedure tegen die waardebeschikking (vgl. Afdeling bestuursrechtspraak van de Raad van State 11 december 2013, nr. 201208181/1/A3, ECLI:NL:RVS:2013:2326, V-N 2014/5.11). Voor die procedure is de belastingrechter de bevoegde rechter (artikel 30 Wet WOZ).

2.5.3. Gelet op deze, door de wetgever beoogde en normaliter aanwezige, samenhang tussen een besluit op een verzoek in de zin van artikel 40 Wet WOZ en de vaststelling van de waarde van een onroerende zaak op grond van die wet, moet worden aangenomen dat ook het eerstbedoelde besluit, in aansluiting op het bepaalde in artikel 30, lid 1, Wet WOZ, is gelijk te stellen met een ingevolge de belastingwet genomen besluit.

2.5.4. Dat heeft ook te gelden in een geval als het onderhavige, waarin een belanghebbende niet is opgekomen tegen een hem betreffende waardebeschikking, maar wel een verzoek heeft gedaan als bedoeld in artikel 40 Wet WOZ. Als de belanghebbende tegen het besluit op een dergelijk verzoek bezwaar maakt, dient de heffingsambtenaar dat bezwaar gelet op hetgeen hiervoor in 2.5.3 is overwogen bij uitspraak niet-ontvankelijk te verklaren. In de wet is immers niet voorgeschreven dat die beslissing door de heffingsambtenaar wordt genomen bij voor bezwaar vatbare beschikking. De belastingrechter en niet de algemene bestuursrechter is bevoegd kennis te nemen van het beroep tegen de uitspraak die de

heffingsambtenaar op een zodanig bezwaar heeft gedaan (vgl. HR 20 december 2013, nr. 12/02872, ECLI:NL:HR:2013:1797, BNB 2014/42, onderdeel 4.2.5).

2.5.5. Het Hof heeft zich gelet op het onder 2.5.4 overwogene ten onrechte onbevoegd verklaard. De klachten slagen in zoverre.

2.6. s Hofs uitspraak kan niet in stand blijven. De Hoge Raad kan de zaak afdoen. Gelet op het onder 2.5.4 overwogene had de heffingsambtenaar het bezwaar niet-ontvankelijk moeten verklaren.

Jurisprudentie I Feitenrechters

4.18 De Rechtbank Zeeland-West-Brabant heeft bij uitspraak van 4 december 2014 overwogen:[25]

2.4. De rechtbank overweegt ten aanzien van het voorgaande als volgt. Ingevolge artikel 40, tweede lid, van de Wet WOZ, verstrekt de heffingsambtenaar op verzoek een afschrift van de gegevens die ten grondslag liggen aan de vastgestelde waarde. Hieronder zijn naar het oordeel van de rechtbank te verstaan de objectieve gegevens die ten grondslag liggen aan de vastgestelde waarde van het onroerend goed, zoals de verkoopwaarden van de vergelijkingsobjecten, de inhouds- en oppervlaktematen, de ligging en kwaliteitsfactoren van de woning en de vergelijkingsobjecten. Naar het oordeel van de rechtbank valt hier niet specifiek de grondstaffel onder. De opbouw van de waarde van de grond is namelijk een hulpmiddel om tot de totale waarde van een onroerende zaak te komen, waarbij de waarde van die onroerende zaak in het kader van de Wet WOZ relevant is en niet de waardeopbouw van de grond afzonderlijk. Voorts is de rechtbank van oordeel dat met hetgeen de heffingsambtenaar in de beroepsfase heeft overgelegd, het taxatierapport en de matrix, in voldoende mate heeft voldaan aan het gestelde in het tweede lid van artikel 40 van de Wet WOZ. In de matrix zijn immers ook de grondwaarden van de vergelijkingsobjecten opgenomen en is de ligging van de woning en de vergelijkingsobjecten gekwalificeerd. Artikel 40 van de Wet WOZ, verplicht de heffingsambtenaar niet, naar het oordeel van de rechtbank, nog meer gegevens te overleggen dan hij gedaan heeft. Met name niet nu de heffingsambtenaar onweersproken heeft gesteld dat er geen grondstaffel is.

Jurisprudentie I Afdeling bestuursrechtspraak Raad van State (hierna: ABRvS)

4.19 Bij uitspraak van 11 augustus 2004 heeft de ABRvS overwogen:[26]

2.4.2. Ten aanzien van de toepasselijkheid van art. 40, tweede lid, Wet WOZ overweegt de Afdeling het volgende. Blijkens de memorie van toelichting bij de Wet WOZ (*Kamerstukken II* 1992/93, 22 885, nr. 3, blz. 27) heeft de wetgever met gegevens in de zin van art. 40, tweede lid, Wet WOZ bedoeld, de gegevens die direct verband houden met de ingevolge deze wet vastgestelde waarde, waarbij nadrukkelijk het aan het waardegegeven onderliggende taxatierapport is genoemd.

4.20 Op 5 juni 2013 heeft de ABRvS geoordeeld:[27]

3.1. [appellant] heeft in hoger beroep niet bestreden dat de door hem verzochte gegevens niet worden geregistreerd, zodat de directeur niet beschikt over een document waarin die gegevens zijn vervat. Anders dan de Afdeling heeft overwogen in de uitspraak van 20 juni 2007 in zaak nr. 200607848/1 bevat de Wob geen verplichting om gegevens te vervaardigen die niet in bestaande documenten zijn neergelegd, ongeacht de mate van inspanning. Wat betreft de stelling van [appellant] dat de gegevens op grond van wettelijke verplichtingen aan andere overheidsinstanties dienen te worden verstrekt, ziet de Afdeling geen aanleiding om te twijfelen aan de mededeling van de directeur dat deze verplichtingen zien op informatie waartoe de door [appellant] gevraagde informatie niet behoort.

De Wob is van toepassing op informatie neergelegd in bestaande documenten. Daarop is het verzoek van [appellant] niet gericht. Voorts leidt het in artikel 10, eerste lid, van het EVRM neergelegde recht om inlichtingen te ontvangen niet tot het oordeel dat op de directeur een verplichting rustte een document met de gevraagde gegevens aan te maken. Evenmin levert de weigering om de gegevens te verstrekken strijd met andere internationale verdragen of met algemene beginselen van behoorlijk bestuur op.

4.21 De ABRvS heeft op 7 augustus 2013 geoordeeld:[28]

5.1. Het verzoek dat [appellant] heeft gedaan, kan niet anders worden begrepen dan als een verzoek om verstrekking van gegevens die ten grondslag liggen aan de vastgestelde waarde voor zijn pand en de daarbij in aanmerking genomen referentiepanden. In artikel 40, tweede lid, van de Wet Woz is bepaald dat deze gegevens niet worden verstrekt aan een ander dan degene te wiens aanzien een beschikking is genomen. Door toepassing van de Wob zou afbreuk worden gedaan aan de goede werking van deze bepaling. Gelet hierop deelt de Afdeling het oordeel van de rechtbank dat in artikel 40, tweede lid, van de Wet Woz een bijzondere openbaarmakingsregeling met een uitputtend karakter is vervat, die de bepalingen van de Wob opzij zet. De rechtbank heeft dan ook met juistheid overwogen dat de directeur zich terecht op het standpunt heeft gesteld dat de Wob niet op het verzoek van toepassing is. Zij heeft het beroep van [appellant] tegen het besluit van 8 december 2011 tot afwijzing van het verzoek van 13 november 2011, terecht ongegrond verklaard.

4.22 De ABRvS heeft bij uitspraak van 11 december 2013 overwogen:[29], [30]

4. Bij de beoordeling van de betogen van [appellant] ziet de Afdeling zich ten eerste voor de vraag gesteld of de rechtbank met betrekking tot de verschillende soorten gegevens terecht heeft beoordeeld of deze onder het bereik van artikel 40 van de Wet Woz dan wel onder die van de Wob vallen, nu zij eerder heeft overwogen (uitspraak van 17 september 2003 in zaak nr. 200300659/1) dat artikel 40 van de Wet Woz een bijzondere regeling voor openbaarmaking met een uitputtend karakter bevat, die de bepalingen van de Wob opzij zet.

Uit de geschiedenis van de totstandkoming van artikel 40 van de Wet Woz (Kamerstukken II 1992/93, 22 885, nr. 3, blz. 27 e.v.) volgt dat de wetgever met deze bepaling een toegesneden regeling inzake openbaarmaking en geheimhouding heeft willen treffen ter zake van bij de waardevaststelling van woningen betrokken gegevens. Daarbij heeft de wetgever getracht een evenwicht te vinden tussen het belang van degene te wiens aanzien een beschikking is genomen om de waardevaststelling van de eigen woning te kunnen controleren en het belang van geheimhouding van gegevens over een niet aan hen toe te rekenen object.

Dat [appellant] heeft verzocht om gegevens die niet expliciet zijn benoemd in artikel 40 van de Wet Woz, laat onverlet dat zijn verzoek beoordeeld moet worden aan de hand van de in die bepaling vervatte openbaarmakingsregeling, nu hij deze gegevens heeft opgevraagd met het oog op het kunnen controleren van de waardevaststelling van zijn woning. Het verstrekken van de verzochte gegevens met toepassing van de Wob zou afbreuk doen aan de strekking van de in artikel 40 van de Wet Woz neergelegde regeling die een vorm van 'beperkte openbaarheid' beoogt. Op het verzoek van [appellant] is de Wob derhalve niet van toepassing. De beoordeling of de gemeenteambtenaar op grond van de in artikel 40 van de Wet Woz neergelegde openbaarmakingsregeling aan [appellant] de door hem verzochte gegevens had moeten verstrekken, dient plaats te vinden in de procedure tegen de waardevaststelling van zijn woning door de in die procedure bevoegde rechter (vergelijk de uitspraak van 7 augustus 2013 in zaak nr. 201208505/1/A3).

4.23 De ABRvS heeft bij uitspraak van 2 april 2014 overwogen:[31]

4. De Afdeling heeft evenwel eerder in een uitspraak van 11 december 2013 in zaak nr. 201208181/1/A3 overwogen dat artikel 40 van de Wet Woz een bijzondere regeling voor openbaarmaking met een uitputtend karakter bevat, die de bepalingen van de Wob opzij zet. Uit de geschiedenis van de totstandkoming van artikel 40 van de Wet Woz (Kamerstukken II 1992/93, 22 885, nr. 3, blz. 27 e.v.) volgt dat de wetgever met deze bepaling een toegesneden regeling inzake openbaarmaking en geheimhouding heeft willen treffen ter zake van bij de waardevaststelling van woningen betrokken gegevens. Daarbij heeft de wetgever getracht een evenwicht te vinden tussen het belang van degene te wiens aanzien een beschikking is genomen om de waardevaststelling van de eigen woning te kunnen controleren en het belang van geheimhouding van gegevens over een niet aan hem toe te rekenen object.

Dat [appellant] heeft verzocht om gegevens die niet expliciet zijn benoemd in artikel 40 van de

Wet Woz, laat onverlet dat zijn verzoek beoordeeld moet worden aan de hand van de in die bepaling vervatte openbaarmakingsregeling, nu hij deze gegevens heeft opgevraagd mede met het oog op het kunnen controleren van de waardevaststelling van zijn woning. Het verstrekken van de verzochte gegevens met toepassing van de Wob zou afbreuk doen aan de strekking van de in artikel 40 van de Wet Woz neergelegde regeling die een vorm van 'beperkte openbaarheid' beoogt. Op het verzoek van [appellant] is de Wob derhalve niet van toepassing. De beoordeling of de heffingsambtenaar op grond van de in artikel 40 van de Wet Woz neergelegde openbaarmakingsregeling aan [appellant] de door hem verzochte gegevens had moeten verstrekken, dient plaats te vinden in de procedure tegen de waardevaststelling van zijn woning door de in die procedure bevoegde rechter.

De rechtbank heeft het voorgaande niet onderkend. Dit noopt evenwel niet tot vernietiging van de aangevallen uitspraak, nu uit het vorenoverwogene volgt dat de rechtbank wel terecht tot het oordeel is gekomen dat het beroep van [appellant] ongegrond dient te worden verklaard.

4.24 Bij uitspraak van 16 november 2016 heeft de ABRvS geoordeeld:[32]

2.2. Zoals de Afdeling eerder heeft overwogen (onder meer uitspraak van 23 december 2015, ECLI:NL:RVS:2015:3918), bevat artikel 40 van de Wet woz een bijzondere regeling voor openbaarmaking met een uitputtend karakter, die de bepalingen van de Wob opzij zet.

Blijkens de memorie van toelichting bij de Wet woz heeft de wetgever met deze bepaling een toegesneden regeling willen treffen inzake openbaarmaking en geheimhouding van voor de waardebepaling van onroerende zaken van belang zijnde gegevens. Daarbij heeft de wetgever getracht een evenwicht te vinden tussen het belang van degene jegens wie een beschikking is genomen om de waardebepaling van de eigen onroerende zaak te kunnen controleren en het belang van geheimhouding van gegevens over een niet aan hem toe te rekenen object.

De rechtbank heeft met juistheid overwogen dat grondstaffels voor de waardebepaling van onroerende zaken van belang zijnde gegevens zijn en iedere grondstaffel aan een specifieke onroerende zaak is gekoppeld. De omstandigheid dat voor vergelijkbare onroerende zaken eenzelfde grondstaffel kan gelden, doet daaraan niet af.

Gelet hierop dient het verzoek van [appellant] te worden beoordeeld aan de hand van de in artikel 40 van de Wet woz vervatte openbaarmakingsregeling. De Wob is derhalve niet van toepassing.

De beoordeling of de heffingsambtenaar op grond van de in artikel 40 van de Wet woz neergelegde openbaarmakingsregeling aan [appellant] de door hem verzochte gegevens had moeten verstrekken, dient plaats te vinden in de procedure tegen de waardebepaling van zijn woning door de in die procedure bevoegde rechter. De rechtbank is terecht tot hetzelfde oordeel gekomen.

5 Beoordeling van de middelen

Inleiding

5.1 Vooropgesteld zij dat het College in de hoger beroepsfase als bijlage bij het verweerschrift de litigieuze grondstaffel heeft overgelegd aan belanghebbende, waarbij ik opmerk dat niet in geschil is dat de overgelegde grondstaffel een andere is dan die ten grondslag heeft gelegen aan de beslissing op het bezwaar.

5.2 Het belang van de vraag of de grondstaffel op grond van artikel 40, eerste dan wel tweede lid, van de Wet WOZ had moeten verstrekt, dan wel het College op grond van artikel 7:4, tweede lid, van de Awb voorafgaand aan het horen inzage had moeten geven in de grondstaffel als zijnde een op de zaak betrekking hebbend stuk, is hier gelegen in de veroordeling van de Heffingsambtenaar in de proceskosten van de beroepsfase en de hoger beroepsfase. Uit r.o. 2.4 en 2.5 van de uitspraak van het Hof is af te leiden dat in deze procedure, onder 'grondstaffel' wordt verstaan de kavelwaarde en de opbouw daarvan.[33], [34]

5.3 Ondanks dat er een verschil bestaat tussen de bezwaarfase, waarin het bestuursorgaan (in casu de heffingsambtenaar) een voor bezwaar en beroep vatbare beschikking neemt en de beroepsfase, waar de bestuursrechter/belastingrechter oordeelt, meen ik dat geen verschil in uitleg bestaat tussen de zinsnede 'de op de zaak betrekking hebbende stukken' als bedoeld in artikel 7:4, tweede lid, van de Awb en artikel 8:42, eerste lid, van de Awb.[35]

5.4 Bij de behandeling van de middelen houd ik, als gebruikelijk, de door het College gekozen volgorde aan.[36]

Eerste middel

5.5 Het eerste middel van het College luidt:

De uitspraak is onvoldoende gemotiveerd doordat het Hof niet ingaat op de grief dat de grondstaffel niet is neergelegd in een stuk dat voor verstrekking in aanmerking komt.

5.6 Het Hof heeft ten aanzien van de litigieuze grondstaffel overwogen:

> 4.12.3. Naar het oordeel van het Hof volgt uit artikel 40 van de Wet WOZ dat een belastingplichtige recht heeft op alle gegevens die van belang kunnen zijn voor het controleren van de voor zijn onroerende zaak vastgestelde waarde, een en ander met inachtneming van het belang van de privacy van anderen. Ook de grondstaffel kan in dit verband van belang zijn, en had derhalve op het verzoek van belanghebbende verstrekt dienen te worden. Met behulp van de grondstaffel kan de belastingplichtige zich een beter gefundeerd oordeel vormen omtrent de grondwaarde van zijn onroerende zaak en controleren of de waardes van de grond van de objecten waarmee hij wordt vergeleken juist is vastgesteld. Daarbij is van belang dat een onjuiste vaststelling van de grondwaardes doorgaans ook gevolgen heeft voor de waardes van de (hoofd)woningen. Uit de onderhavige casus blijkt dat voor de toepassing van de grondstaffel niet blindelings op de Heffingsambtenaar kan worden vertrouwd. Kortom, het kennisnemen van de grondstaffel kan belanghebbende helpen om inzicht te verschaffen in de vaststelling van de waarde van zijn onroerende zaak en hem daarmee in de gelegenheid stellen om op een beter gefundeerde wijze al dan niet de vastgestelde waarde aan te vechten.
>
> Het verstrekken van de grondstaffel levert naar het oordeel van het Hof geen schending van de privacy van anderen op daar de grondstaffel gedestilleerd is uit openbare verkooptransacties en/of onderdeel is van kenbaar gronduitgiftebeleid, en derhalve niet te herleiden is naar een individu.

5.7 Het College heeft toegelicht:

Het Gerechtshof gaat uit van de veronderstelling dat de grondstaffel is neergelegd in een stuk dat voor verstrekking in aanmerking komt. Deze veronderstelling is evenwel niet juist en daarnaast ook onbegrijpelijk. In de motivering van het hoger beroep, aan het Hof toegezonden op 13 januari 2016, is uitvoerig uiteengezet dat de grondstaffel niet kan worden aangemerkt als een stuk in de zin van artikel 7:4 van de Awb. In dit kader werd het volgende opgemerkt:

"De waardebepaling zoals die in het kader van de Wet Waardering onroerende zaken (hierna: WOZ) geschiedt, vindt bij de taxatie in eerste aanleg geautomatiseerd plaats. Hiertoe worden door (nagenoeg uitsluitend externe) softwareontwikkelaars taxatiemodellen ontworpen die per gemeente worden gevuld met de relevante objectgegevens en de beschikbare verkoopinformatie. Op grond van modelmatige analyse worden in het taxatiemodel onder meer staffels berekend die kunnen worden gebruikt voor het bepalen van de grondwaarde van het te taxeren object, met inachtneming van het type object en de ligging van de onroerende zaak.

De grondstaffels zijn derhalve verwerkt in het taxatiemodel, welke programmatuur is neergelegd in een softwareprogramma van de ingeschakelde externe softwareleverancier. Deze staffels zijn dan ook niet neergelegd in een stuk, noch zijn zij uit de taxatiesoftware (in het onderhavige geval betreft dit het

programma GeoTax 2.5) te destilleren.

Anders dan de rechtbank overweegt is de grondstaffel dus niet neergelegd in een stuk; van een op de zaak betrekking hebbend stuk zoals bedoeld in artikel 7:4 van de Awb is dan ook geen sprake. Evenmin is het de gemeente mogelijk deze grondstaffels uit de betreffende programmatuur te destilleren. Reeds om deze reden heeft de rechtbank niet kunnen oordelen dat sprake is van schending van dit wetsartikel. Dit geeft aanleiding om de uitspraak van de rechtbank op dit punt te vernietigen hetgeen eveneens moet gelden voor de door de rechtbank uitgesproken proceskostenveroordeling."

De vraag of sprake is van een stuk in de zin van artikel 7:4 van de Awb is van belang om te kunnen bepalen of terinzagelegging noodzakelijk is. Voornoemde grief, inhoudende dat van een voor terinzagelegging vatbaar stuk in de zin van artikel 7:4 van de Awb geen sprake is, is dus van belang voor de beantwoording van de vraag of sprake is van een verplichting tot terinzagelegging als bedoeld in dit artikel en of, wanneer van een dergelijke verplichting sprake is, men dan gehouden is om voorafgaand aan de hoorzitting alsnog niet bestaande stukken voor de terinzagelegging te vervaardigen. (Op deze laatste vraag wordt later in dit beroepschrift meer uitgebreid teruggekomen).

Het Hof is in de bestreden uitspraak echter op geen enkele manier ingegaan op de opgeworpen grief, zoals neergelegd in de motivering van het hoger beroepschrift van 13 januari 2016 (pagina's 2 en 3) en gaat zonder meer uit van de veronderstelling dat sprake is van een stuk als bedoeld in artikel 7:4 van de Awb. Gelet op de naar voren gebrachte grieven terzake had het Hof dit standpunt niet kunnen innemen zonder op deze grieven in te gaan en deze gemotiveerd te verwerpen. De omstandigheid dat het Hof in de bestreden uitspraak geen enkele overweging aan deze grieven wijdt, maakt de uitspraak naar de mening van het College onbegrijpelijk en onvoldoende gemotiveerd. Een en ander geeft grond voor vernietiging van de uitspraak.

5.8 In de toelichting op het eerste middel stelt het College dat het Hof ten onrechte niet heeft gemotiveerd of en waarom een grondstaffel een 'stuk' is als bedoeld in artikel 7:4, tweede lid, van de Awb.

5.9 Op grond van artikel 7:4, tweede lid, van de Awb legt het bestuursorgaan 'het bezwaarschrift en alle verder op de zaak betrekking hebbende stukken voorafgaand aan het horen gedurende ten minste een week voor belanghebbenden ter inzage'. Op grond van artikel 7:4, vierde lid, van de Awb kunnen belanghebbenden 'van deze stukken tegen vergoeding van ten hoogste de kosten afschriften verkrijgen'.[37]

5.10 Zoals reeds in de inleiding van dit vijfde onderdeel van de conclusie aan de orde is gekomen, meen ik dat voor de uitleg van de zinsnede 'op de zaak betrekking hebbende stukken' (en daarmee ook 'stuk') als bedoeld in artikel 7:4, tweede lid, van de Awb, kan worden aangesloten bij hetgeen heeft te gelden op grond van artikel 8:42, eerste lid, van de Awb. Wel zij opgemerkt dat een belangrijk verschil tussen deze artikelen is dat het in de bezwaarfase gaat om een verplichting van de inspecteur of de heffingsambtenaar om inzage te geven in de stukken of om een afschrift daarvan te verstrekken, terwijl het in de rechterlijke fase in beginsel gaat om het verstrekken van de stukken.

5.11 Vooropgesteld meen ik dat aan het College kan worden toegegeven dat het Hof zich niet expliciet erover heeft uitgelaten of een grondstaffel moet worden aangemerkt als een 'stuk' in de zin van artikel 7:4, tweede lid, van de Awb.

5.12 Ook uit de wetsgeschiedenis bij de Awb is niet af te leiden wat moet worden verstaan onder het begrip 'stuk'.[38]

5.13 Blijkens de hierboven opgenomen toelichting zijn de grondstaffels verwerkt in een taxatiemodel, waarvan de programmatuur is neergelegd in een softwareprogramma van een ingeschakelde,

externe softwareleverancier. Afhankelijk van de vraag of dit programma een digitaal bestand is of een applicatie,[39] heeft naar mijn mening het volgende te gelden.

5.14 De Hoge Raad heeft zich nog niet eerder over de vraag uitgelaten of digitale bestanden, databases en applicaties zijn aan te merken als 'op de zaak betrekking hebbende stukken' in de zin van artikel 7:4, tweede lid, van de Awb. Wel heeft de Hoge Raad in het arrest van 20 december 2013 overwogen dat onder stukken als bedoeld in artikel 8:42, eerste lid, van de Awb 'mede zijn verstaan afdrukken van in elektronische vorm vastgelegde gegevens'.[40]

5.15 Zoals in onderdeel 5.4 van de Bijlage aan de orde is gekomen, kunnen na de inwerkingtreding van het digitaal procederen bij de bestuursrechter, pdf-bestanden door de inspecteur digitaal ter beschikking worden gesteld aan de bestuursrechter. Ten aanzien van stukken opgenomen in andere digitale bestanden dan pdf-bestanden, bijvoorbeeld databases, heb ik in onderdeel 5.6 van de Bijlage mijn mening te kennen gegeven dat ook deze bestanden (in welke vorm dan ook) moeten worden aangemerkt als 'op de zaak betrekking hebbende stukken' als bedoeld in artikel 8:42, eerste lid, van de Awb, indien deze aan de inspecteur ter beschikking staan dan wel hebben gestaan en een rol hebben gespeeld bij de besluitvorming. Overlegging moet ook volgen indien een belanghebbende voldoende gemotiveerd heeft gesteld dat bepaalde stukken van enig belang kunnen zijn (geweest) voor de besluitvorming in zijn zaak.

5.16 De vervolgvraag is of het voorgaande ook heeft te gelden ten aanzien van 'op de zaak betrekking hebbende stukken' die zijn opgenomen in applicaties die door de inspecteur of heffingsambtenaar bij de belastingheffing zijn gebruikt. Vooropgesteld zij, dat in beginsel ook applicaties zijn aan te merken als 'digitale bestanden'. Naar mijn mening zijn ook applicaties aan te merken als 'op de zaak betrekking hebbende stukken' indien deze bij de besluitvorming aan de inspecteur ter beschikking staan dan wel hebben gestaan en daarbij een rol (kunnen) hebben gespeeld. Zie de uitspraak van de geheimhoudingskamer van gerechtshof 's-Hertogenbosch van 20 mei 2016, waarin is overwogen dat 'de applicatie op digitale wijze hetgeen [bevat dat] normaliter in een papieren dossier wordt vastgelegd'.[41] Voorts heeft dit gerechtshof overwogen dat 'de applicatie een op de zaak van belanghebbende betrekking hebbend stuk [is], in zoverre gegevens zijn opgenomen van belanghebbende'.[42]

5.17 Het belang van het voorgaande voor de onderhavige zaak lijkt evenwel beperkt, nu de belanghebbende niet om digitale inzage in het programma heeft gevraagd en uiteindelijk, na afloop van de bezwaarfase de softwareleverancier de grondstaffel uit het taxatiemodel heeft weten te herleiden, zodat deze op dat moment wél beschikbaar was en een afschrift daarvan, te zien als een op de zaak betrekking hebbend stuk, met het verweerschrift in hoger beroep aan belanghebbende is verzonden.[43] Overigens geldt dat het digitaal procederen, in het kader van de inwerkingtreding van het 'bestuursrecht 2.0' voorlopig nog niet mogelijk en verplicht is en dat artikel 8:36a van de Awb nog niet in werking is getreden.[44]

5.18 Het lijkt mij dat de digitalisering (van het bestuursprocesrecht) ook gevolgen zal hebben voor het inzagerecht als bedoeld in artikel 7:4, tweede lid, van de Awb, alsmede artikel 7:4, vierde lid, van de Awb. Gedacht zou kunnen worden aan inzage in digitale zin en/of eventueel de verstrekking door de inspecteur of heffingsambtenaar van stukken in digitale zin.

5.19 Wat hiervan ook zij, de Hoge Raad heeft bij arrest van 20 december 2013 overwogen dat onder op de zaak betrekking hebbende stukken 'mede zijn verstaan afdrukken van in elektronische vorm vastgelegde gegevens'. Nu de grondstaffel in een dergelijke elektronische vorm was opgenomen, meen ik dat reeds daarom het eerste middel van het College faalt.[45]

Tweede middel

5.20 Het tweede middel van het College luidt:

Bevestiging van de uitspraak van het hof creëert rechtsongelijkheid.

5.21 Het College heeft toegelicht:

De heersende jurisprudentie inzake de gegevensverstrekking die onder de werking van de Wet openbaarheid van bestuur (hierna: Wob) dient plaats te vinden (zie bijvoorbeeld de uitspraak van de Afdeling Bestuursrechtspraak van de Raad van State (ABRvS) van 5 juni 2013, nr. 201204362/1/A3, ECLI:NL:RVS:2013:CA2102)[46] houdt in dat de Wob geen verplichting bevat om gegevens te vervaardigen die niet in bestaande documenten zijn neergelegd. Niet valt in te zien waarom een dergelijke redenering niet ook zou moeten worden gevolgd in het geval van een verzoek om verstrekking van gegevens waarop de Wob niet van toepassing is.

Voor een analoge toepassing van deze jurisprudentie bestaat naar de mening van het College alle aanleiding. Wanneer een verzoek om gegevens in het kader van artikel 40 van de Wet WOZ wordt gedaan zonder dat tevens bezwaar wordt gemaakt of beroep wordt ingesteld tegen een belastingaanslag, is in hoger beroep de ABRvS de bevoegde rechter (Gerechtshof Arnhem-Leeuwarden, 2 februari 2016, 15/00046, ECLI:NL:GHARL:2016:639). Deze legt als maatstaf voor de vraag of de gevraagde gegevens moeten worden verstrekt (mede) aan of de gevraagde gegevens al dan niet in bestaande documenten zijn neergelegd (ABRvS, 5 juni 2013, nr. 201204362/1/A3, ECLI:NL:RVS:2013:CA2102).

Wanneer het verzoek om gegevensverstrekking op basis van artikel 40 van de Wet WOZ wordt gedaan binnen het kader van een bezwaar- of beroepsprocedure tegen de waarde van een onroerende zaak, is de Wob niet op dit verzoek van toepassing maar moet het verzoek worden beoordeeld binnen het kader van de bezwaar- of beroepsprocedure tegen de belastingaanslag of de WOZ-beschikking (vgl. ABRvS 11 december 2013, nr. 201208181/1/A3, ECLI:NL:RVS:2013:2326[47] en ABRvS 2 april 2014, nr. 201308103/1/A3, ECLI:NL:RVS:2014:1161). [48] In hoger beroep is dan het Gerechtshof bevoegd.

Wanneer de uitspraak van het Gerechtshof door Uw Raad wordt bevestigd, een uitspraak waarin het Hof meent dat gegevens moeten worden verstrekt die niet in bestaande documenten zijn neergelegd, en hiermee tot leidende jurisprudentie wordt gemaakt, wordt rechtsongelijkheid gecreëerd. In dat geval heeft immers een verzoeker die om gegevens verzoekt buiten het kader van een belastingprocedure geen recht op verstrekking van gegevens die niet in documenten zijn vastgelegd (immers; bevoegde rechter is de ABRvS), terwijl een verzoeker die om gegevens verzoekt binnen het kader van een belastingprocedure wel recht heeft op verstrekking van deze gegevens (immers: bevoegde rechter is het Gerechtshof).

Een en ander is naar de mening van het College aanleiding de bestreden uitspraak te vernietigen.

5.22 Ik wil voorop stellen dat de door het College gepercipieerde rechtsongelijkheid ten gevolge van andersluidende uitspraken van enerzijds de ABRvS en de anderzijds de Hoge Raad, geen directe grond kan zijn om de in casu bestreden uitspraak van het Hof te vernietigen. In fiscale zaken wordt gevaren op de rechtsoordelen van (de belastingkamer van) de Hoge Raad. Dat neemt overigens niet weg dat rechtseenheid in oordelen van de hoogste bestuursrechtelijke rechtscolleges in principe wenselijk is.[49]

5.23 Daarbij komt het mij voor dat het College met de geschetste rechtsongelijkheid uitgaat van een onjuiste rechtsopvatting. De Hoge Raad heeft op 25 november 2016 arrest gewezen en heeft de uitspraak van het Gerechtshof, waarop het College zich hier beroept, vernietigd.[50] Volgens de Hoge Raad moet gelet op de 'door de wetgever beoogde en normaliter aanwezige, samenhang tussen een besluit op een verzoek in de zin van artikel 40 Wet WOZ en de vaststelling van de waarde van een onroerende zaak op grond van die wet, (…) worden aangenomen dat ook het eerstbedoelde besluit, in aansluiting op het bepaalde in artikel 30, lid 1, Wet WOZ, is gelijk te stellen met een ingevolge de belastingwet genomen besluit'[51] en dat dit ook heeft te gelden in een geval 'waarin een belanghebbende niet is opgekomen tegen een hem betreffende waardebeschikking, maar wel een verzoek heeft gedaan als bedoeld in artikel 40 Wet WOZ'. Voorts heeft de Hoge Raad in dit arrest overwogen dat indien een belanghebbende 'tegen het besluit op een dergelijk verzoek bezwaar maakt, (…) de heffingsambtenaar dat bezwaar (…) bij uitspraak niet-ontvankelijk dient te verklaren. In de wet is immers niet voorgeschreven dat die

beslissing door de heffingsambtenaar wordt genomen bij voor bezwaar vatbare beschikking. De belastingrechter en niet de algemene bestuursrechter is bevoegd kennis te nemen van het beroep tegen de uitspraak die de heffingsambtenaar op een zodanig bezwaar heeft gedaan'.[52]

5.24 Kortom, van enige ongelijkheid is geen sprake, nu de belastingrechter ten aanzien van een beroep op artikel 40 van de Wet WOZ de enige bevoegde rechter is en niet de algemene bestuursrechter.

5.25 Daarmee ontvalt ook hier het belang aan de door het College genoemde uitspraak van de ABRvS van 5 juni 2013, waarin is overwogen dat 'de Wob geen verplichting [bevat] om gegevens te vervaardigen die niet in bestaande documenten zijn neergelegd, ongeacht de mate van inspanning'.[53], [54] Daarbij merk ik nog op dat het in die zaak ging om niet-geregistreerde gegevens, terwijl in de onderhavige zaak de grondstaffels juist wel waren geregistreerd.

5.26 Ook het tweede middel van het College kan niet tot cassatie leiden.

Derde middel

5.27 Het derde middel van het College luidt:

Ten onrechte overweegt het Hof dat de grondstaffel A) voorafgaand aan de hoorzitting ter inzage had moeten worden gelegd en B) valt onder de gegevens die op grond van artikel 40 van de Wet WOZ moeten worden verstrekt.

5.28 Uit de toelichting bij het derde middel zijn drie verschillende klachten af te leiden, te weten i) of aan de verplichting als bedoeld in artikel 7:4, tweede lid, van de Awb moet worden voldaan als pas tijdens de hoorzitting om inzage van een bepaald stuk is verzocht, ii) wat de reikwijdte is van artikel 40 van de Wet WOZ en iii) hoe artikel 40 van de Wet WOZ en artikel 7:4, tweede lid, van de Awb zich tot elkaar verhouden.[55]

5.29 Ad i) Blijkens de toelichting stelt het College dat 'niet [kan] worden gesteld dat de grondstaffel voorafgaand aan de hoorzitting aan belanghebbende ter beschikking had moeten worden gesteld, aangezien namens belanghebbende eerst ter hoorzitting om verstrekking van de grondstaffel is verzocht'.

5.30 Met deze klacht komt het College op tegen het volgende oordeel van het Hof:

> 4.12.5. Ingevolge artikel 7:4, lid 2, van de Algemene wet bestuursrecht (hierna: Awb) dient de Heffingsambtenaar alle op de zaak betrekking hebbende stukken voorafgaand aan het horen ter inzage te leggen voor belanghebbende gedurende ten minste een week. Ingevolge artikel 7:4, lid 4, van de Awb kan belanghebbende tegen vergoeding van ten hoogste de kosten afschriften van de ter inzage verstrekte stukken verkrijgen. Naar het oordeel van het Hof had de Heffingsambtenaar de grondstaffel, als zijnde een op de zaak betrekking hebbend stuk als bedoeld in de zin van artikel 7:4, lid 2, van de Awb ter inzage moeten leggen, en indien de belanghebbende bij de inzage om een afschrift daarvan had verzocht, hem dat moeten verstrekken. Gelet hierop had de Heffingsambtenaar het verzoek van belanghebbende (een afschrift van) de grondstaffel te verstrekken niet mogen weigeren.

5.31 Aan het College kan worden toegegeven dat artikel 7:4, tweede lid, van de Awb een verplichting voor het bestuursorgaan inhoudt om de op de zaak betrekking hebbende stukken *voorafgaand aan het horen* aan belanghebbende(n) ter inzage te leggen en dat het verzoek om inzage van de grondstaffel pas op de hoorzitting van 28 augustus 2014 door belanghebbende is gedaan.

5.32 Echter, er is wettelijk geen voorafgaand verzoek van de kant van belanghebbende nodig om inzage te mogen krijgen in de op de zaak betrekking hebbende stukken. Het bestuursorgaan dient die namelijk zelf, op eigen initiatie ter inzage te leggen, voorafgaand aan het horen in de

bezwaarfase.[56]

5.33 Wat hiervan ook zij, het Hof heeft de Heffingsambtenaar niet in de door belanghebbende in bezwaar gemaakte kosten veroordeeld, overwegende dat 'van de Heffingsambtenaar (…) immers niet verlangd [kan] worden om (ongevraagd) voorafgaande aan of gelijktijdig met het versturen van de beschikking en de aanslag de grondstaffel aan belanghebbende te verstrekken'. Daardoor heeft het College naar mijn mening geen belang bij klacht i), zodat die faalt.[57]

5.34 Ad ii) Het Hof heeft overwogen:

> 4.12.3. Naar het oordeel van het Hof volgt uit artikel 40 van de Wet WOZ dat een belastingplichtige recht heeft op alle gegevens die van belang kunnen zijn voor het controleren van de voor zijn onroerende zaak vastgestelde waarde, een en ander met inachtneming van het belang van de privacy van anderen. Ook de grondstaffel kan in dit verband van belang zijn, en had derhalve op het verzoek van belanghebbende verstrekt dienen te worden. Met behulp van de grondstaffel kan de belastingplichtige zich een beter gefundeerd oordeel vormen omtrent de grondwaarde van zijn onroerende zaak en controleren of de waardes van de grond van de objecten waarmee hij wordt vergeleken juist is vastgesteld. Daarbij is van belang dat een onjuiste vaststelling van de grondwaardes doorgaans ook gevolgen heeft voor de waardes van de (hoofd)woningen. Uit de onderhavige casus blijkt dat voor de toepassing van de grondstaffel niet blindelings op de Heffingsambtenaar kan worden vertrouwd. Kortom, het kennisnemen van de grondstaffel kan belanghebbende helpen om inzicht te verschaffen in de vaststelling van de waarde van zijn onroerende zaak en hem daarmee in de gelegenheid stellen om op een beter gefundeerde wijze al dan niet de vastgestelde waarde aan te vechten. Het verstrekken van de grondstaffel levert naar het oordeel van het Hof geen schending van de privacy van anderen op daar de grondstaffel gedestilleerd is uit openbare verkooptransacties en/of onderdeel is van kenbaar gronduitgiftebeleid, en derhalve niet te herleiden is naar een individu.

5.35 Het College betoogt dat het Hof heeft miskend 'dat artikel 40 van de Wet WOZ een uitputtende regeling geeft voor openbaarmaking van de gegevens die aan de waardebepaling ten grondslag liggen'.[58] Dat artikel beoogt volgens het College 'een beperkte openbaarmaking van gegevens (waaronder het taxatieverslag), aan de hand waarvan een belanghebbende de waarde van zijn woning kan controleren'. Volgens het College valt de grondstaffel niet onder artikel 40, eerste lid van de Wet WOZ en valt onder artikel 40, tweede lid, van de Wet WOZ, slechts het taxatieverslag.

5.36 Tevens haalt het College een uitspraak van de ABRvS van 2 april 2014 aan, waarin de ABRvS heeft overwogen dat 'uit de geschiedenis van de totstandkoming van artikel 40 van de Wet Woz[59] volgt dat de wetgever met deze bepaling een toegesneden regeling inzake openbaarmaking en geheimhouding heeft willen treffen ter zake van bij de waardevaststelling van woningen betrokken gegevens (…)',[60] evenals een uitspraak van de ABRvS van 11 augustus 2004, waarin is overwogen dat 'de wetgever met gegevens in de zin van art. 40, tweede lid, Wet WOZ (…) , de gegevens [heeft bedoeld] die direct verband houden met de ingevolge deze wet vastgestelde waarde, waarbij nadrukkelijk het aan het waardegegeven onderliggende taxatierapport is genoemd'.

5.37 Ik meen evenwel dat met het in artikel 40, tweede lid, van de Wet WOZ bedoelde *afschrift van de gegevens die ten grondslag liggen aan de vastgestelde waarde,* niet alleen wordt gedoeld op het taxatieverslag. Ingevolge de Tweede Nota van Wijziging bij het wetsvoorstel Wet WOZ gaat het om 'de aan de taxatie onderliggende gegevens', waaronder naar mijn mening ook de grondstaffel dient te worden verstaan.[61] Dat in de Memorie van Toelichting bij het wetsvoorstel Wet WOZ expliciet 'het taxatierapport' is genoemd, doet daaraan naar mijn mening niet af, omdat het mij lijkt dat dit slechts is genoemd als een onderdeel van 'de aan de taxatie onderliggende gegevens'.

5.38 Aldus komt het mij voor dat ’s Hofs oordeel uitgaat van een juiste rechtsopvatting en dat gelet op de overweging dat ‘met behulp van de grondstaffel (…) de belastingplichtige zich een beter gefundeerd oordeel [kan] vormen omtrent de grondwaarde van zijn onroerende zaak en controleren of de waardes van de grond van de objecten waarmee hij wordt vergeleken juist is vastgesteld’, overigens ook niet onbegrijpelijk te achten is.

5.39 Dat betekent dat ook klacht ii) moet falen.

5.40 Ad iii) Ten slotte betoogt het College dat de bijzondere openbaarmakingsregeling van artikel 40 van de Wet WOZ verhindert dat gegevens die op grond van ‘deze uitputtende regeling niet behoeven te worden verstrekt langs de weg van artikel 7:4 van de Awb kan worden bereikt dat deze gegevens toch aan belanghebbende bekend moeten worden gemaakt’.

5.41 Blijkens de toelichting op dit middelonderdeel,[62] stelt het College dat moet worden aangesloten bij een overweging van de ABRvS in een uitspraak van 2 april 2014, waarin is overwogen dat ‘het verstrekken van de verzochte gegevens met toepassing van de Wob (...) afbreuk (zou) doen aan de strekking van de in artikel 40 van de Wet Woz neergelegde regeling die een vorm van 'beperkte openbaarheid’ beoogt’ en ‘niet valt in te zien waarom ten aanzien van artikel 7:4 van de Awb een andere redenering zou moeten worden gevolgd’.[63]

5.42 Ik zou die benadering niet willen volgen. Vooropgesteld zij, dat blijkens een uitspraak van de ABRvS van 16 november 2016 de WOB helemaal niet van toepassing is op grondstaffels.[64] Ook ziet, anders dan waar het College vanuit lijkt te gaan, de ‘beperkte openbaarheid’ op de beperking aan *wie* een afschrift van de gegevens moet worden verstrekt en niet op een beperking van de te verstrekken stukken of gegevens. Immers de gegevens dienen ingevolge artikel 40, tweede lid, van de Awb, alleen te worden verstrekt ‘aan degene te wiens aanzien een beschikking is genomen’.

5.43 Iets dergelijks kan zich niet voordoen bij toepassing van artikel 7:4, tweede lid, van de Awb, omdat dit niet uitgaat van een verzoek, maar van een eigen verplichting van de inspecteur.

5.44 Naar mijn mening is er noch in de wettekst, noch in de wetsgeschiedenis bij artikel 40 van de Wet WOZ of bij artikel 7:4 van de Awb, steun te vinden voor de op exclusiviteit gerichte opvatting van het College.[65]

5.45 Een en ander betekent dat klacht iii) het lot van de voorgaande twee moet delen, zodat het derde middel faalt.

Vierde middel

5.46 Het vierde middel van het College luidt:

Ten onrechte overweegt het Hof dat eerst in hoger beroep voldoende inzicht is geboden in de totstandkoming van de waarde.

5.47 Dit middel is gericht tegen de volgende overweging van het Hof:

4.14. Het ten onrechte weigeren om de grondstaffel te verstrekken vormt voor het Hof geen aanleiding om de uitspraken op bezwaar te vernietigen, omdat de beschikte waarde immers niet te hoog is vastgesteld. Wel zal het Hof de Heffingsambtenaar veroordelen in de proceskosten van de beroepsfase en de hoger beroepsfase. Belanghebbende is immers door de weigerachtige houding van de Heffingsambtenaar moeten blijven procederen tot in hoger beroep om een gefundeerd oordeel te kunnen vormen van de door de Heffingsambtenaar vastgestelde waarde om aan de hand van de ontvangen gegevens in alle redelijkheid te kunnen inschatten of hij al dan niet (hoger) beroep zou instellen. Pas na ontvangst van het verweerschrift in hoger beroep, met daarbij gevoegd de nieuwe matrix met de grondstaffel en de definitieve, gewijzigde, cijfers

betreffende de grondwaardes en de waardes van de (hoofd)woningen van de onroerende zaak en de referentieobjecten, kon belanghebbende de hiervoor vermelde inschatting maken.

Voor de bezwaarfase veroordeelt het Hof de Heffingsambtenaar niet voor de door belanghebbende in bezwaar gemaakte kosten. Van de Heffingsambtenaar kan immers niet verlangd worden om (ongevraagd) voorafgaande aan of gelijktijdig met het versturen van de beschikking en de aanslag de grondstaffel aan belanghebbende te verstrekken, opdat deze voorafgaande aan het al dan niet instellen van bezwaar de voor zijn onroerende zaak vastgestelde waarde kan toetsen en controleren.

5.48 Blijkens de toelichting bestrijdt het College 's Hofs (impliciete) oordeel dat onvoldoende inzicht is geboden. Het College betoogt dat het alle gegevens heeft verstrekt als bedoeld in artikel 6 van het Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ. Tevens heeft het College blijkens de toelichting op het vierde middel herhaald, wat zij bij de toelichting op haar vierde middel reeds heeft betoogd, namelijk dat alleen het taxatieverslag geldt als 'gegevens' in de zin van artikel 40, tweede lid, van de Wet WOZ en dat daarmee 's Hofs impliciete oordeel 'dat (…) de heffingsambtenaar niet aan zijn verplichting tot het leveren van informatie heeft voldaan' onbegrijpelijk is. Dat ik dit anders zie, blijkt uit de bovenstaande behandeling van het derde middel.

5.49 Ten slotte acht het College het onbegrijpelijk dat 'het niet verstrekken van de grondstaffel (…) belanghebbende verhindert een juiste beoordeling van de juistheid van de waarde te maken', maar het ontbreken van de grondstaffel de Rechtbank en het Hof er niet van heeft weerhouden om een oordeel over de vastgestelde waarde te vellen.[66]

5.50 Bij de behandeling van dit vierde middel stel ik voorop dat uit artikel 6 van het Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ in samenhang met de Uitvoeringsregeling instructie waardebepaling Wet WOZ, is af te leiden dat het taxatieverslag ten minste moet bevatten: de objectaanduiding, de waarderelevante objectgegevens, de motivering van de individuele afwijking ten opzichte van de relevante marktgegevens en de getaxeerde waarde.[67] Anders dan het College kennelijk voorstaat, kan daaruit naar mijn mening niet worden afgeleid dat daarmee per definitie is voldaan aan de verplichtingen als bedoeld in artikel 40 van de Wet WOZ, alsmede artikel 7:4, tweede lid, van de Awb.

5.51 Ik acht het niet onbegrijpelijk dat het Hof enerzijds heeft geoordeeld dat de Heffingsambtenaar aan het verzoek van de gemachtigde om de grondstaffel te verstrekken tegemoet had moeten komen, maar anderzijds heeft geoordeeld dat het ten onrechte weigeren daarvan geen aanleiding vormt om de uitspraken op bezwaar te vernietigen, nu volgens het Hof de beschikte waarde niet te hoog is vastgesteld. Het College miskent hier dat sprake is van twee verschillende oordelen op de twee verschillende geschilpunten bij het Hof, te weten een oordeel ten aanzien van een materiële klacht (of de vastgestelde waarde van de onroerende zaak van € 373.000 per peildatum 1 januari 2013 te hoog is vastgesteld) en een oordeel ten aanzien van een formele klacht (of de Heffingsambtenaar aan het verzoek van de gemachtigde om de grondstaffel te verstrekken tegemoet had moeten komen).[68]

5.52 Aldus faalt ook het vierde middel van het College.

6 Conclusie

De conclusie strekt ertoe dat het beroep in cassatie van het college van burgemeester en wethouders van de gemeente Waalwijk ongegrond dient te worden verklaard.

De Procureur-Generaal bij de

---

[1] Gerechtshof 's-Hertogenbosch 10 februari 2017, nr. 15/01429, ECLI:NL:GHSHE:2017:501, *FutD* 2017/0886.

[2] Zie voor de integrale wettekst 4.4.

[3] De in deze conclusie opgenomen citaten uit jurisprudentie en literatuur zijn meestal zonder daarin voorkomende voetnoten opgenomen. Citaten met een tekstbewerking, zoals onderstrepingen, vet- of cursiefzettingen, zijn veelal als onbewerkt weergegeven. In citaten voorkomende witregels zijn soms weggelaten.

[4] Daarnaar wordt hierna verwezen als: Bijlage.

[5] Gerechtshof 's-Hertogenbosch 10 februari 2017, nr. 15/01429, ECLI:NL:GHSHE:2017:501, *FutD* 2017/0886.

[6] Rechtbank Zeeland-West-Brabant 27 oktober 2015, nr. AWB 14/7668, ECLI:NL:RBZWB:2015:7281, *FutD* 2015/2886.

[7] Gerechtshof 's-Hertogenbosch 10 februari 2017, nr. 15/01429, ECLI:NL:GHSHE:2017:501, *FutD* 2017/0886.

[8] Voetnoot A-G: tegen deze uitspraak is eveneens cassatie ingesteld. De zaak is thans aanhangig bij de Hoge Raad onder nr. 16/04497. Ik neem in de onderhavige zaak en de zaak met nr. 16/04497 gelijktijdig, als onderdeel van een cluster zaken, conclusie.

[9] *NTFR* 2017/1431.

[10] *Belastingblad* 2017/200.

[11] Per 1 september 1999 is hoofdstuk 8 van de Awb van overeenkomstige toepassing verklaard op het belastingrecht. Zie: Wet van 29 oktober 1998, houdende aanpassing van het fiscale procesrecht aan de Algemene wet bestuursrecht en wijziging van een aantal fiscale en andere wetten (herziening van het fiscale procesrecht), Stb. 1998, 621

[12] Met ingang van 1 oktober 2016 is artikel 40 van de Wet WOZ gewijzigd door alleen van toepassing te zijn op een bepaalde onroerende zaak die niet in hoofdzaak tot woning dient. Zie voor de verstrekking van waardegegevens ten aanzien van een bepaalde onroerende zaak die in hoofdzaak tot woning dient artikel 40a van de Wet WOZ.

[13] *Kamerstukken II* 1992/93, 22 885, nr. 3, p. 27-29.

[14] Uit onder meer de derde nota van wijziging bij de Wet WOZ is af te leiden dat het oorspronkelijke voorgestelde artikel 40 van de Wet WOZ is komen te vervallen, zodat uiteindelijk artikel 41 is vernummerd tot artikel 40. Waar in het citaat 'artikel 41' wordt genoemd, wordt aldus artikel 40 van de Wet WOZ bedoeld. Zie: *Kamerstukken II* 1993/94, 22 885, nr. 198, p. 3.

[15] *Kamerstukken II* 1993/94, 22 885, nr. 10, p. 3.

[16] *Kamerstukken II* 1993/94, 22 885, nr. 22, p. 1.

[17] Zie *Stb.* 1994/874.

[18] Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ van 23 december 2009, *Stb.* 2009,615.

[19] Deze instructie is vervat in de Uitvoeringsregeling instructie waardebepaling Wet WOZ van 30 december 2011, *Stcrt.* 2011,22974.

[20] Hoge Raad 7 juni 2002, nr. 36 801, ECLI:NL:HR:2002:AE3833, *BNB* 2002/284 met noot Spek.

[21] Hoge Raad 21 april 2006, nr. 41 185, ECLI:NL:HR:2006:AW2326, *NTFR* 2006/607 met noot Kastelein.

[22] Hoge Raad 13 augustus 2010, nr. 08/04247, ECLI:NL:HR:2010:BK8502, *BNB* 2010/323 met noot Van Casteren.

[23] Hoge Raad 20 december 2013, nr. 12/02872, ECLI:NL:HR:2013:1797, na conclusie A-G IJzerman, *BNB* 2014/42 met noot Pechler.

[24] Hoge Raad 25 november 2016, nr. 16/01414, ECLI:NL:HR:2016:2667, *BNB* 2017/41 met noot Jansen.

[25] Rechtbank Zeeland-West-Brabant 4 december 2014, nr. AWB/14/4804, ECLI:NL:RBZWB:2014:8315, *Belastingblad* 2015/108 met noot Borghols.

[26] Afdeling bestuursrechtspraak Raad van State 11 augustus 2004, nr. 200306466/1, ECLI:NL:RVS:2004:AQ6645, *Belastingblad* 2004/998.

[27] Afdeling bestuursrechtspraak Raad van State 5 juni 2013, nr. 201204362/1/A3, *V-N* 2013/32.8 met noot redactie.

[28] Afdeling bestuursrechtspraak Raad van State 7 augustus 2013, nr. 201208505/1/A3, ECLI:NL:RVS:2013:629, *V-N* 2013/54.4 met noot redactie.

[29] Afdeling bestuursrechtspraak Raad van State 11 december 2013, nr. 201208181/1A3, ECLI:NL:RVS:2013:2326, *V-N* 2014/5.11 met noot redactie.

[30] De ABRvS heeft de volgende overwogen reeds in diverse uitspraken opgenomen: 'Uit de geschiedenis van de totstandkoming van artikel 40 van de Wet WOZ (…) volgt dat de wetgever met deze bepaling een toegesneden regeling inzake openbaarmaking en geheimhouding heeft willen treffen ter zake van bij de waardevaststelling van woningen betrokken gegevens. Daarbij heeft de wetgever getracht een evenwicht te vinden tussen het belang van degene te wiens aanzien een beschikking is genomen om de waardevaststelling van de eigen woning te kunnen controleren en het belang van geheimhouding van gegevens over een niet aan hen toe te rekenen object'. Zie de volgende uitspraken: Afdeling bestuursrechtspraak Raad van State 2 april 2014, nr. 201308103/1A3, ECLI:NL:RVS:2014:1161, *V-N* 2014/50.28 met noot redactie, afdeling bestuursrechtspraak Raad van State 23 december 2015, nr. 201502882/1/A3, ECLI:NL:RVS:2015:3918, afdeling bestuursrechtspraak Raad van State 16 november 2016, nr. 201600879/1/A3, ECLI:NL:RVS:2016:3022, *V-N* 2016/62.25.11 met noot redactie.

[31] Afdeling bestuursrechtspraak Raad van State van 2 april 2014, nr. 201308103/1/A3, ECLI:NL:RVS:2014:1161, *V-N* 2014/50.28 met noot redactie.

[32] Afdeling bestuursrechtspraak Raad van State van 16 november 2016, nr. 201600879/1/A3/, ECLI:NL:RVS:2016:3022, *NLF* 2017/0555 met noot Van der Vegt.

[33] Zie voor deze overwegingen onderdeel 2.1 van deze conclusie.

[34] In zo'n staffel wordt de prijs per m² grond gegeven, waarbij geldt: hoe meer grond, des te goedkoper per m². Dat wil zeggen dat de prijs per m² afneemt naarmate de perceeloppervlakte groter wordt. Deze staffel wordt toegepast omdat bij grote percelen niet het gehele perceel hetzelfde nut oplevert. Voorbeeld: bij een vergelijkbare woning met meer grond wordt een lagere prijs per m² gehanteerd.

[35] Dat geen verschil bestaat in de te hanteren toets of sprake is van op de zaak betrekking hebbende stukken bij zowel artikel 7:4, tweede lid, van de Awb en artikel 8:42, eerste lid, van de Awb is naar mijn mening expliciet af te leiden uit r.o. 2.4.2 van het arrest van de Hoge Raad van 18 december 2015. Zie onderdeel 2.73 van de Bijlage.

[36] Hoewel ik betwijfel of daarin de juiste volgorde en opzet zijn gekozen, ten aanzien van toepassing van de artikel 40 van de Wet WOZ of artikel 7:4, tweede lid, van de Awb.

[37] Zie voor de tekst van dit artikel onderdeel 4.1.

[38] Zie voor de wetsgeschiedenis bij de Awb de onderdelen 2.23 – 2.35 van de Bijlage.

[39] Onder 'applicatie' moet volgens Van Dale Handwoordenboek Hedendaags Nederlands worden verstaan: 'computerprogramma'. Het gaat daarbij naar mijn mening om het in elektronische vorm verzamelen en ordenen van gegevens, als onderworpen aan programmaregels, welke kunnen leiden tot uitkomsten ('applications').

[40] Zie voor het arrest van de Hoge Raad van 20 december 2013 de onderdelen 2.64 en 4.30 van de

Bijlage.

[41] Zie voor de uitspraak van gerechtshof 's-Hertogenbosch van 20 mei 2016 de onderdelen 2.101 en 4.53 van de Bijlage.

[42] De CRvB oordeelde op 17 april 2003 andersluidend en overwoog dat een database en applicatie behoudens eventuele uitdraaien uit die database niet behoren tot de stukken als bedoeld in artikel 7:4, tweede lid, van de Awb. Zie onderdeel 2.113 van de Bijlage.

[43] Dit volgt uit de toelichting op het eerste middel, zoals opgenomen in onderdeel 3.4.

[44] Zie de onderdelen 3.11 en 3.14 van de Bijlage.

[45] Zie voor het arrest van de Hoge Raad van 20 december 2013 de onderdelen 2.64 en 4.30 van de Bijlage.

[46] Voetnoot A-G: zie 4.20 voor de uitspraak van de ABRvS van 5 juni 2013.

[47] Voetnoot A-G: zie 4.22 voor de uitspraak van de ABRvS van 11 december 2013.

[48] Voetnoot A-G: zie 4.23 voor de uitspraak van de ABRvS van 2 april 2014.

[49] Vgl. de tekst van artikel 81, eerste lid, van de Wet op de rechterlijke organisatie.

[50] Zie 4.17 voor het arrest van de Hoge Raad van 25 november 2016.

[51] Zie voor de tekst van artikel 30 van de Wet WOZ onderdeel 4.3.

[52] Zie daarvoor ook het arrest van de Hoge Raad van 20 december 2013 de onderdelen 2.64 en 4.30 van de Bijlage.

[53] Zie onderdeel 4.20.

[54] Ten overvloede merk ik op dat de ABRvS bij uitspraak van 23 december 2015 en 16 november 2016 heeft geoordeeld dat de regeling van de Wet WOZ (waaronder artikel 40 van de Wet WOZ) een lex specialis is ten opzichte van de WOB en daarop voorgaat. Zie onderdeel 4.24.

[55] Zie voor de toelichting op het derde middel onderdeel 3.8.

[56] Zie het standaardarrest van 25 april 2008, zoals opgenomen in 2.45 en 4.40.

[57] Vgl. hetgeen ik in onderdeel 5.2. heb opgemerkt over het belang van het College bij deze procedure.

[58] Zie onder meer de uitspraken van de ABRvS in de onderdelen 4.21 t/m 4.24.

[59] Voetnoot A-G: zie *Kamerstukken II* 1992/93, 22 885, nr. 3, p. 27, als opgenomen in onderdeel 4.8.

[60] Zie de uitspraak van de ABRvS in onderdeel 4.23.

[61] Zie onderdeel 4.9.

[62] Zie voor de toelichting op het derde middel onderdeel 3.8.

[63] Zie 4.23 voor de uitspraak van de ABRvS van 2 april 2014.

[64] Zie voor de uitspraak van de ABRvS van 16 november 2016 onderdeel 4.24.

[65] Zie voor de wetsgeschiedenis bij artikel 40 van de Wet WOZ de onderdelen 4.8 t/m 4.10.

[66] Zie voor de volledige toelichting op het vierde middel onderdeel 3.10.

[67] Zie voor de tekst van artikel 6 van het Uitvoeringsbesluit onderbouwing en uitvoering waardebepaling Wet WOZ onderdeel 4.11 en 4.12.

[68] Zie r.o. 3.1. van de uitspraak van het Hof.

Raad van State

ECLI	ECLI:NL:RVS:2017:1259
Datum uitspraak	17 mei 2017
Inhoudsindicatie	Bij verschillende besluiten van 14 december 2015 heeft het college vergunningen krachtens artikel 16 en 19d van de Natuurbeschermingswet 1998 (hierna: Nbw 1998) verleend voor het exploiteren en/of uitbreiden en wijzigen van zes verschillende agrarische bedrijven.

Volledige tekst

Bij deze uitspraak is een persbericht uitgebracht.

201600614/1/R2, 201600617/1/R2, 201600618/1/R2, 201600620/1/R2, 201600622/1/R2 en 201600630/1/R2.
Datum uitspraak: 17 mei 2017

AFDELING
BESTUURSRECHTSPRAAK

Verwijzingsuitspraak in het geding tussen:

de Stichting Werkgroep Behoud de Peel, gevestigd te Deurne (hierna: de Werkgroep),
appellante,

en

het college van gedeputeerde staten van Noord-Brabant,
verweerder

Procesverloop

Bij verschillende besluiten van 14 december 2015 heeft het college vergunningen krachtens artikel 16 en 19d van de Natuurbeschermingswet 1998 (hierna: Nbw 1998) verleend voor het exploiteren en/of uitbreiden en wijzigen van zes verschillende agrarische bedrijven.

Tegen deze besluiten heeft de Werkgroep beroep ingesteld.

Het college heeft in deze zaken een verweerschrift ingediend.

[vergunninghouder A] heeft gebruik gemaakt van de geboden gelegenheid een schriftelijke uiteenzetting te geven (zaaknr. 201600630/1/R2).

De Stichting Advisering Bestuursrechtspraak voor Milieu en Ruimtelijke Ordening heeft desverzocht een deskundigenbericht in deze zaken uitgebracht.
De Werkgroep en het college hebben hun zienswijze daarop naar voren gebracht.

Het college en de Werkgroep hebben nadere stukken ingediend.

De Afdeling heeft de bovengenoemde zaken met de zaken in nrs. 201506807/1/R2, 201506815/1/R2 en 201506818/1/R2 op 30 november 2016 en 1 december 2016 gevoegd ter zitting behandeld.
Ter zitting zijn de Werkgroep, vertegenwoordigd door W.M.M. van Opbergen, bijgestaan door ir. A.K.M. van Hoof, rechtsbijstandverlener te Gennep, en het college, vertegenwoordigd door mr. H.J.M. Besselink, advocaat te Den Haag, en bijgestaan door onder meer mr. M. Heerings, ir. E.J. Maltha-Nix, ir. B.J.L. Clabbers, ir. D. Bal en ir. S.J.M. Breukel, verschenen.
Voorts zijn daar namens de Stichting Advisering Bestuursrechtspraak voor Milieu en Ruimtelijke Ordening, ir. V.C.A. Bogaardt, ing. J.H. Grit, drs. J.F. Schuurman en ing. P. Stroeken, als deskundigen gehoord. Verder zijn [vergunninghouder B], vertegenwoordigd door [gemachtigde], bijgestaan door mr. J.J.J. de Rooij, advocaat te Tilburg (zaaknr. 201600620/1/R2) en [vergunninghouder A], vertegenwoordigd door mr. J.J.J. de Rooij, advocaat te Tilburg (zaaknr. 201600630/1/R2), beide vergunninghouders, gehoord.

Na de zitting zijn de zaken gesplitst.

Na het sluiten van het onderzoek ter zitting heeft de Afdeling het onderzoek heropend en partijen medegedeeld dat zij voornemens is het Hof van Justitie van de Europese Unie (hierna: het Hof van Justitie) te verzoeken bij wijze van prejudiciële beslissing uitspraak te doen op een aantal vragen. Deze vragen zijn vervolgens in concept aan partijen gezonden.

De Werkgroep en het college hebben een reactie gegeven op deze vragen.

INHOUDSOPGAVE

Keuzes, gegevens en aannames over de omvang van de depositieruimte
Conclusie onderdeel H
I. SLOT

Overwegingen

1. Op 1 januari 2017 is de Wet natuurbescherming (hierna: Wnb) in werking getreden en is de Nbw 1998 ingetrokken. Omdat de bestreden besluiten zijn genomen voor 1 januari 2017 volgt uit artikel 9.10 van de Wnb dat deze geschillen moeten worden beoordeeld aan de hand van het voor die datum geldende recht.

A. INLEIDING EN OPZET UITSPRAAK

2. Deze verwijzingsuitspraak hangt samen met de verwijzingsuitspraak van heden, ECLI:NL:RVS:2017:1260. In beide uitspraken zijn toestemmingsregimes voor activiteiten die stikstofdepositie veroorzaken op Natura 2000-gebieden aan de orde.

2.1. In onderhavige uitspraak worden de beroepen van de Werkgroep tegen zes vergunningen voor verschillende agrarische bedrijven in de provincie Noord-Brabant behandeld. Deze vergunningen zijn verleend met toepassing van het Programma Aanpak Stikstof 2015-2021 (hierna: het PAS) en de daarbij behorende regelgeving die vanaf 1 juli 2015 van kracht is.

2.2. In het PAS staat dat in 118 van de 162 Nederlandse Natura 2000-gebieden sprake is van een overbelasting van stikstofdepositie op stikstofgevoelige habitattypen en leefgebieden van soorten. De belangrijkste nationale bron van uitstoot van stikstof is de veehouderij. Daarnaast dragen verkeer, scheepvaart, industrie en consumenten (bijvoorbeeld woningen, recreatie) bij aan de stikstofbelasting. De bijdrage van bronnen in het buitenland aan de depositie op de Nederlandse Natura 2000-gebieden is substantieel: zij bedraagt gemiddeld over alle Natura 2000-gebieden circa 35% van de totale depositie (zie p. 14-15 van het PAS).

2.3. De overbelasting vormt een probleem voor zowel de verwezenlijking van de instandhoudingsdoelstellingen voor de stikstofgevoelige natuurwaarden in de Natura 2000-gebieden als voor het mogelijk maken van economische ontwikkelingen die stikstofdepositie veroorzaken. Omdat stikstof tot op grote afstand van de bron neerslaat en de 118 Natura 2000-gebieden met overbelaste stikstofgevoelige habitats en leefgebieden verspreid over Nederland liggen, is voor veel projecten, zoals woningbouw, de aanleg van wegen, industrie en veehouderij, nabij en op grote afstand van Natura 2000-gebieden een vergunning vereist waarbij de gevolgen van de daardoor veroorzaakte stikstofdepositie op verschillende Natura 2000-gebieden dienen te worden beoordeeld. De vergunningverlening voor deze projecten stagneerde omdat de beoordeling complex is en kostbaar voor initiatiefnemers.

2.4. Het probleem van de overbelasting van de natuurwaarden en de stagnatie van economische ontwikkelingen die stikstofdepositie veroorzaken is aanleiding geweest voor de ontwikkeling van een programmatische aanpak van de stikstofproblematiek. Een belangrijk onderdeel daarvan vormt het PAS. Met het PAS wordt beoogd de verslechtering van de stikstofgevoelige natuurwaarden in de Natura 2000-gebieden te voorkomen en op termijn de instandhoudingsdoelstellingen daarvoor te realiseren. Daarnaast voorziet het PAS en de daarbij behorende regelgeving in een beoordelingskader voor ontwikkelingen die

stikstofdepositie veroorzaken.

2.5. In onderhavige verwijzingsuitspraak staat de vraag centraal of het beoordelingskader voor stikstofveroorzakende projecten en andere handelingen verenigbaar is met artikel 6, tweede en derde lid, van Richtlijn 92/43/EEG van de Raad van de Europese Gemeenschappen van 21 mei 1992 inzake de instandhouding van de natuurlijke habitats en de wilde flora en fauna (PbEG1992 L206; hierna Habitatrichtlijn). Ten aanzien van de passende beoordeling die ten grondslag ligt aan het PAS is de vraag aan de orde of en onder welke voorwaarden instandhoudingsmaatregelen, passende maatregelen en beschermingsmaatregelen daarin mogen worden betrokken. De vragen over de maatregelen in de passende beoordeling worden ook voorgelegd in de verwijzingsuitspraak van heden, ECLI:NL:RVS:2017:1260.

2.6. De maatschappelijke gevolgen van de verwijzing van deze zaken naar het Hof van Justitie zijn groot. De beantwoording van de prejudiciële vragen is van belang voor veel ontwikkelingen in Nederland. In de periode vanaf de inwerkingtreding van het PAS op 1 juli 2015 tot 31 december 2016 zijn 3103 meldingen gedaan en 4299 vergunningen aangevraagd voor activiteiten die stikstofdepositie veroorzaken, zoals de realisering van woningbouwlocaties, de aanleg van wegen, uitbreiding van industriële activiteiten en ontwikkelingen in de veehouderij. Omdat de depositie ver van de bron neerslaat en de Natura 2000-gebieden verspreid over Nederland liggen, verkeren initiatiefnemers van dergelijke projecten in heel Nederland thans in onzekerheid of een vergunning voor hun project kan worden verleend en of die, indien daartegen beroep wordt ingesteld, onherroepelijk zal worden. De Afdeling acht van belang dat de onzekerheid of dergelijke economische ontwikkelingen doorgang kunnen vinden zo kort mogelijk duurt.

2.7. Beide verwijzingsuitspraken zijn omvangrijk. Dat komt door de complexiteit van de materie, de regelgeving en het programma. Een uitgebreide beschrijving daarvan is nodig voor het verkrijgen van inzicht in de instrumenten die in de Nbw 1998 zijn gekozen om aan de verplichtingen van artikel 6 van de Habitatrichtlijn te voldoen. In beide verwijzingsuitspraken heeft de Afdeling voorts een aanzet tot beantwoording van de prejudiciële vragen gegeven.

2.8. De uitspraak is als volgt opgebouwd. In de onderdelen B en C worden achtereenvolgens de betrokken veehouderijen, de bestreden besluiten en de beroepen daartegen, beschreven. Deze onderdelen zijn voor de prejudiciële en nationale procedure van belang. In onderdeel D wordt een formele beroepsgrond behandeld die alleen voor de nationale procedure van belang is. Daarna volgt in de onderdelen E en F de beschrijving van de programmatische aanpak en de verhouding van het PAS tot artikel 6 van de Habitatrichtlijn. Deze twee onderdelen zijn van belang voor de prejudiciële procedure. Onderdeel G bevat het verzoek aan het de president van het Hof van Justitie om de zaak met voorrang te behandelen. Enkele beroepsgronden tegen de keuzes, gegevens en aannames die ten grondslag liggen aan het PAS en de daarbij behorende onderzoeken en AERIUS worden besproken in onderdeel H. Onderdeel I bevat de slotoverweging en gaat in op het verzoek om hangende de prejudiciële verwijzing een voorlopige voorziening te treffen. Deze twee onderdelen (H en I) zijn alleen voor de nationale procedure van belang.

B. DE BETROKKEN VEEHOUDERIJEN

3. De bestreden besluiten betreffen Nbw-vergunningen voor zes verschillende agrarische

bedrijven in de provincie Noord-Brabant. De bedrijven veroorzaken stikstofdepositie op onder meer de Natura 2000-gebieden Groote Peel en Deurnsche Peel & Mariapeel. Beide gebieden zijn aangewezen voor het stikstofgevoelige habitattype herstellend hoogveen.

De vergunning voor de [vergunninghouder C] heeft betrekking op de oprichting van een melkvee- en varkenshouderij aan de [locatie 1] in Deurne. Op grond van de vergunning mogen 300 melkkoeien, 70 stuks jongvee en 320 vleesvarkens met een totale emissie van 3572 NH3 kg/jr worden gehouden (zaaknr. 201600614/1/R2). De vergunde situatie leidt tot een toename van stikstofdepositie van maximaal 2,99 mol N/ha/jr op de Deurnsche Peel & Mariapeel.

[vergunninghouder D] exploiteert aan de [locatie 2] in Someren een melkveebedrijf. De vergunning is verleend voor de exploitatie en uitbreiding van het bedrijf. Het bedrijf kan op grond van de vergunning groeien tot 200 stuks melkvee en 140 stuks jongvee, met een totale emissie van 3086 NH3 kg/jr (zaaknr. 201600617/1/R2). De hoogste depositie van de bestaande activiteit inclusief de uitbreiding bedraagt 1,51 mol N/ha/jr op de Groote Peel en 0,55 mol N/ha/jr op de Deurnsche Peel & Mariapeel. De maximale toename ten opzichte van de feitelijk veroorzaakte emissie in 2014 bedraagt op deze gebieden 0,18 respectievelijk 0,06 mol N/ha/jr.

[vergunninghouder E] is gevestigd aan de [locatie 3] te Someren. Aan dit bedrijf is een vergunning verleend voor de exploitatie en uitbreiding van de pluimveehouderij. Het bedrijf mag op grond van de vergunning 30.152 ouderdieren van vleeskuikens houden, met een totale emissie van 7688 NH3 kg/jr (zaaknr. 201600618/1/R2). De hoogste depositie van de bestaande activiteit inclusief de uitbreiding bedraagt 2,26 mol N/ha/jr op de Groote Peel en 1,29 mol N/ha/jr op de Deurnsche Peel & Mariapeel. De maximale toename ten opzichte van de feitelijk veroorzaakte emissie in 2014 bedraagt op deze gebieden 1,85 respectievelijk 1,05 mol N/ha/jr.

[vergunninghouder B] exploiteert een pluimveebedrijf aan de [locatie 4] te Someren. Aan dit bedrijf is een vergunning verleend voor de exploitatie van de pluimveehouderij. Het bedrijf mag op grond van de vergunning 50.000 ouderdieren van vleeskuikens houden met een totale emissie van 12.500 NH3 kg/jr (zaaknr. 201600620/1/R2). De hoogste depositie van de bestaande activiteit bedraagt 2,18 mol N/ha/jr op de Groote Peel en 1,91 mol N/ha/jr op de Deurnsche Peel & Mariapeel. Deze depostie werd feitelijk ook veroorzaakt in 2014, zodat ten opzichte van die situatie de depositie niet toeneemt.

Varkenshouderij Limburglaan B.V. exploiteert een varkenshouderij aan de Limburglaan 7 in Someren. Aan het bedrijf is een vergunning verleend voor de exploitatie en uitbreiding van de varkenshouderij met 7980 vleesvarkens, 3584 gespeende biggen, 750 guste/dragende zeugen, 240 kraamzeugen en 2 dekberen, met een totale emissie van 6057,04 NH3 kg/jr (zaaknr. 201600622/1/R2). De hoogste depositie van de bestaande activiteit inclusief de uitbreiding bedraagt 1,05 mol N/ha/jr op de Groote Peel en 0,76 mol N/ha/jr op de Deurnsche Peel & Mariapeel. De maximale toename ten opzichte van de feitelijk veroorzaakte emissie in 2014 bedraagt op deze gebieden 0,24 respectievelijk 0,18 mol N/ha/jr.

[vergunninghouder A] exploiteert een varkenshouderij aan de [locatie 5] te Someren. Aan het bedrijf is een vergunning verleend voor de verandering van de varkenshouderij met in totaal 2996 vleesvarkens en 20132 gespeende biggen, met een totale emissie van 2650 NH3 kg/jr

(zaaknr. 201600630/1/R2). Aan dit bedrijf is eerder een Nbw-vergunning verleend. De thans vergunde situatie leidt ten opzichte van de eerder Nbw-vergunde situatie tot een toename van 0,27 mol N/ha/jr op de Deurnsche Peel & Mariapeel.

De Werkgroep, die beroep heeft ingesteld tegen deze vergunningen, zet zich in voor het behoud en herstel van het Peelgebied als een hoogveen(achtig) landschap met de daaraan verbonden historische waarden en natuurkwaliteiten. De Werkgroep geeft voorlichting over het Peelgebied en is betrokken bij het opstellen van overheidsplannen zoals het landinrichtingsplan Peelvenen, waterhuishoudingsplannen, en beheerplannen voor Natura 2000-gebieden. Daarnaast verzorgt de Werkgroep met een groep vrijwilligers het beheerwerk in de Deurnsche Peel en Heidsche Peel. Dat werk bestaat onder meer uit het afdammen van de oude afwateringssloten en de wijken waardoor het gebied vernat, en het zagen en kappen van berkenbomen, die er als gevolg van de verdroging zijn gaan groeien. Sommige delen van het gebied worden begraasd door schapen of koeien, om te proberen boom- en grasgroei weg te houden en heide terug te krijgen. De Werkgroep heeft beroep ingesteld omdat zij vreest dat de stikstofdepositie die de agrarische bedrijven veroorzaken op de natuurwaarden in de Natura 2000-gebieden Groote Peel en Deurnsche Peel & Mariapeel leiden tot aantasting van die waarden.

C. KORTE DUIDING BESTREDEN BESLUITEN EN BEROEP

4. Het college heeft bij de bestreden besluiten vergunningen verleend voor de exploitatie en/of uitbreiding van agrarische bedrijven die stikstofdepositie veroorzaken op stikstofgevoelige natuurwaarden in Natura 2000-gebieden. Het heeft daarbij toepassing gegeven aan het PAS en de daarbij behorende regelgeving die vanaf 1 juli 2015 in de Nbw 1998, het Besluit grenswaarden programmatische aanpak stikstof (hierna: het Besluit grenswaarden) en de Regeling programmatische aanpak stikstof (hierna: de Regeling) is opgenomen. Deze regelgeving biedt een beoordelingskader voor vergunningen voor projecten en andere handelingen die stikstof uitstoten en daardoor significante gevolgen kunnen hebben voor stikstofgevoelige natuurwaarden in Natura 2000-gebieden.

In één zaak (201600620/1/R2) heeft het college de vergunning verleend omdat de stikstofdepositie van het bedrijf niet toeneemt ten opzichte van de feitelijk veroorzaakte depositie voorafgaand aan het PAS. De depositie van bestaande activiteiten is voor het PAS passend beoordeeld als onderdeel van de achtergronddepositie. De vergunning is verleend onder verwijzing naar de passende beoordeling die voor het PAS is gemaakt.

In de overige zaken zijn vergunningen aan de orde voor bedrijfsactiviteiten die leiden tot een toename van stikstofdepositie ten opzichte van de feitelijk veroorzaakte of de vergunde depositie voorafgaand aan het PAS. Voor zover deze projecten stikstofdepositie veroorzaken die de voor het desbetreffende Natura 2000-gebied geldende drempel- of grenswaarde (0,05 of 1 mol N/ha/jr) overschrijdt, is voor de toename van stikstofdepositie ontwikkelingsruimte toegedeeld. De depositie die kan ontstaan door benutting van de (totale) ontwikkelingsruimte die in het PAS per hectare van een stikstofgevoelig habitattype in een Natura 2000-gebied is berekend, is voor het PAS passend beoordeeld. De vergunningen zijn verleend onder verwijzing naar de passende beoordeling die voor het PAS is gemaakt.

Voor zover deze projecten stikstofdepositie veroorzaken die onder de drempel- en grenswaarde blijft die voor het desbetreffende Natura 2000-gebied geldt, heeft bij de

vergunningverlening geen beoordeling plaatsgevonden van de toename van stikstofdepositie op de stikstofgevoelige natuurwaarden in die Natura 2000-gebieden. Voor dergelijke toenames is in het PAS zogenoemde depositieruimte gereserveerd, die voor het PAS passend is beoordeeld. Voor de projecten is, voor zover het deposities tussen 0,05 mol en 1 mol N/ha/jr betreft, wel een meldingsplicht van toepassing. Hieraan is door het college uitvoering gegeven.

4.1. De beroepen van de Werkgroep strekken ertoe dat het college de vergunningen niet kon verlenen met toepassing van het PAS en de regelgeving over de programmatische aanpak stikstof. Het PAS en de regelgeving waarborgen volgens de Werkgroep niet dat op een juiste wijze uitvoering wordt gegeven aan artikel 6 van de Habitatrichtlijn. Het ambitieniveau van het PAS, dat inzet op het voorkomen van verslechtering in de eerste PAS-periode van zes jaar en het bijdragen aan herstel in de tweede en derde PAS-periode is volgens de Werkgroep te mager. Het bieden van enig uitzicht op herstel van de kwaliteit van de habitats ondanks een blijvende zeer hoge overschrijding van de kritische depositiewaarde na drie PAS-periodes, biedt volgens de Werkgroep geen uitzicht op het bereiken van een gunstige staat van instandhouding. Het PAS biedt daarom niet de zekerheid dat projecten die stikstofdepositie veroorzaken en op basis van het PAS worden toegestaan, de kwaliteit van de natuurlijke habitats niet zal aantasten.

Verder meent de Werkgroep dat een programma niet de plaats van een individuele beoordeling kan innemen, die artikel 6, derde lid, van de Habitatrichtlijn vereist voor projecten die significante gevolgen kunnen hebben voor Natura 2000-gebieden. Verder stelt zij dat de passende beoordeling die aan het PAS ten grondslag ligt niet voldoet aan de eisen die de Habitatrichtlijn stelt, omdat daarin instandhoudings- en passende maatregelen zijn betrokken die op grond van artikel 6, eerste en tweede lid, van de Habitatrichtlijn, moeten worden getroffen. Daarnaast zijn in de passende beoordeling bron- en herstelmaatregelen betrokken die volgens de Werkgroep het karakter van compensatie hebben. De Werkgroep vindt het in het licht van artikel 6 van de Habitatrichtlijn voorts bezwaarlijk dat de ontwikkelingsruimte kan worden toegedeeld voordat de positieve gevolgen van de maatregelen zich hebben voorgedaan.

Behalve deze principiële juridische bezwaren stelt de Werkgroep verschillende inhoudelijke bezwaren tegen het PAS aan de orde. In algemene zin stelt zij dat het voor belanghebbenden nauwelijks mogelijk is inzicht te krijgen in de uitgangspunten, de prognoses en de berekeningen die aan het PAS en het rekeninstrument AERIUS ten grondslag liggen. De uitgangspunten en berekeningen van de omvang van de depositiedaling door de te treffen PAS-bronmaatregelen, de afdwingbaarheid van deze maatregelen en het onverklaarde verschil tussen de gemeten en berekende concentratie van ammoniak in de lucht, bevatten volgens de Werkgroep te veel onzekerheden, om daar een bepaalde omvang van depositie-/ontwikkelingsruimte voor nieuwe ontwikkelingen aan te ontlenen. De effectiviteit van bepaalde herstelmaatregelen staat volgens de Werkgroep niet vast en onzeker is of de herstelmaatregelen daadwerkelijk en tijdig worden getroffen. Het instellen van een drempelwaarde acht de Werkgroep bezwaarlijk. Door de overgangsregeling voor extern salderen is bovendien niet gewaarborgd dat een deel van de depositie van stoppende agrarische bedrijven, die als ontwikkelingsruimte beschikbaar wordt gesteld, al is benut voor extern salderen. De Werkgroep betwijfelt de mogelijkheden van bijsturing in het geval dat de afname van depositie achterblijft bij de prognose.

Tot slot stelt de Werkgroep dat de bestreden besluiten in strijd met artikel 2, vijfde lid, van de Nbw 1998, niet in overeenstemming met het college van gedeputeerde staten van Limburg zijn genomen.

D. INSTEMMING LIMBURG MET VERGUNNINGVERLENING

(slechts relevant voor de nationale procedure)

5. De Werkgroep betoogt dat de vergunningen in afwijking van artikel 2, vijfde lid, van de Nbw 1998 niet in overeenstemming met het college van gedeputeerde staten van Limburg zijn verleend. Het uitblijven van een reactie kan volgens de Werkgroep niet gelijk worden gesteld met instemming.

5.1. Artikel 2, vijfde lid, van de Nbw 1998 luidde ten tijde van belang: "Gedeputeerde staten beslissen niet op een aanvraag voor een vergunning als bedoeld in artikel 19d, eerste lid, dan in overeenstemming met gedeputeerde staten van de andere provincies waarin Natura 2000-gebieden, of delen daarvan, zijn gelegen waarvoor het project of de andere handeling waarvoor vergunning wordt verleend gevolgen kan hebben".

5.2. Vaststaat dat het college van gedeputeerde staten van Noord-Brabant bevoegd gezag is en dat de besluiten in overeenstemming met het college van gedeputeerde staten van Limburg dienden te worden genomen.

5.3. Het college van gedeputeerde staten van Limburg heeft bij brief van 28 juli 2015 aan de andere provincies en het ministerie van Economische Zaken zijn werkwijze ten aanzien van artikel 2, vijfde lid, van de Nbw 1998 bekendgemaakt. Hierin is vermeld dat een vergunning in overeenstemming met het college van gedeputeerde staten van Limburg is verleend indien:

"- overeenkomstig uw beleidsregels wordt beschikt zonder gebruikmaking van de hardheidsclausule;

- de effecten van de activiteit op een in Limburg gelegen Natura 2000-gebied beperkt blijven tot effecten als gevolg van stikstofdepositie;

- na bekendmaking van het ontwerpbesluit vier weken zijn verstreken en u van ons geen reactie heeft ontvangen.

Indien niet voldaan wordt aan bovengenoemde voorwaarden, dient u ons expliciet om instemming te verzoeken. De termijn van vier weken is dan niet van toepassing".

5.4. In de aan de orde zijnde vergunningzaken heeft het college van gedeputeerde staten van Noord-Brabant het ontwerpbesluit tot verlening van de vergunning toegezonden aan het college van gedeputeerde staten van Limburg met het verzoek in te stemmen met het ontwerpbesluit. Daarbij is vermeld dat het college de instemming binnen vier weken na verzending van de brief tegemoet ziet en indien niet binnen vier weken een reactie wordt ontvangen, dat dan conform het door alle provincies vastgestelde beleid automatisch wordt ingestemd met het besluit.

De Afdeling is van oordeel dat artikel 2, vijfde lid, van de Nbw 1998 een dergelijke

handelwijze, waarbij uitdrukkelijk om instemming wordt verzocht en overeenkomstig de afspraak met het college dat om instemming wordt gevraagd, is aangegeven dat de instemming geacht wordt te zijn gegeven als niet na een bepaalde termijn een reactie is ontvangen, niet uitsluit.

Het betoog faalt.

E. BESCHRIJVING VAN DE PROGRAMMATISCHE AANPAK STIKSTOF

Terminologie die in het PAS wordt gebruikt

6. Voorafgaand aan een inhoudelijke bespreking van het PAS worden de centrale begrippen uit dit systeem weergegeven en kort gedefinieerd. Deze weergave beoogt niet de beschrijving van de begrippen in de overwegingen hierna te vervangen, maar dient uitsluitend als hulpmiddel bij het lezen van deze uitspraak.

Stikstofgevoelige natuurwaarden:

habitattypen in Habitatrichtlijngebieden en habitats van soorten in Habitatrichtlijngebieden en Vogelrichtlijngebieden die aangetast kunnen worden door een overmaat aan depositie van stikstof. De mogelijke aantasting bestaat in dat geval uit een vermestende en/of een verzurende invloed van het stikstof.

Kritische depositiewaarde:

de grens waarboven het risico bestaat dat de kwaliteit van het habitat significant wordt aangetast door de verzurende en/of vermestende invloed van atmosferische stikstofdepositie.

Vergunning:

toestemming door het bevoegde gezag, zonder welke geen projecten mogen worden gerealiseerd of andere handelingen mogen worden verricht die de kwaliteit van de natuurlijke habitats of de habitats van soorten in een Natura 2000-gebied kunnen verslechteren of een significant verstorend effect kunnen hebben op de soorten waarvoor dat gebied is aangewezen.

Melding:

een bericht aan het bevoegd gezag dat wordt ingediend met gebruikmaking van AERIUS. Hierbij wordt een omschrijving gegeven van de aard en de omvang van een voorgenomen project of andere handeling dat stikstofdepositie tot gevolg kan hebben. De melding wordt ten minste vier weken maar ten hoogste twee jaar voor de aanvang gedaan.

Achtergronddepositie:

de heersende depositie als gevolg van het totaal van bijdragen van alle emissiebronnen. Bij de aanvraag van een vergunning of de melding van een activiteit wordt deze achtergronddepositie als uitgangspunt genomen. De stikstofdepositie als gevolg van de

activiteit waarvoor de vergunning wordt aangevraagd of de melding wordt gedaan komt hier bovenop.

Depositieruimte:

de totale hoeveelheid stikstofdepositie die in het PAS tijdens één tijdvak van 6 jaren beschikbaar is voor de groei van bestaande activiteiten en voor het uitvoeren van nieuwe projecten of andere handelingen. Deze ruimte is onderverdeeld per in het programma opgenomen Natura 2000-gebied. De depositieruimte bestaat uit vier segmenten: ruimte voor autonome ontwikkeling (groei), ruimte voor deposities onder de grenswaarde, ontwikkelingsruimte voor prioritaire projecten en vrije ontwikkelingsruimte.

Autonome ontwikkeling:

dit is de verandering van stikstofdepositie als gevolg van ontwikkelingen die los staan van het PAS, zoals de groei van stikstofdepositie als gevolg van economische groei en de daling van stikstofdepositie als gevolg van technische ontwikkelingen.

Grenswaarde:

een waarde die is uitgedrukt in mol stikstofdepositie per hectare per jaar (mol N/ha/jr). Voor zover een project of andere handeling stikstofdepositie tot gevolg heeft, maar beneden deze waarde blijft, is deze binnen het PAS uitgezonderd van de vergunningplicht.

Drempelwaarde:

een waarde die is uitgedrukt in mol stikstofdepositie per hectare per jaar (mol N/ha/jr).Voor zover een project of andere handeling stikstofdepositie tot gevolg heeft, maar beneden deze waarde blijft, is deze binnen het PAS uitgezonderd van de vergunningplicht en de meldingsplicht.

Ontwikkelingsruimte:

stikstofdepositie die kan worden toegedeeld of gereserveerd voor toestemmingsbesluiten van projecten of andere handelingen. Deze maakt deel uit van de depositieruimte. De ontwikkelingsruimte bestaat uit twee segmenten: segment 1 bevat prioritaire projecten. Dit is een lijst met specifiek benoemde projecten die het rijk of de provincies aanmerken als projecten van nationaal of provinciaal maatschappelijk belang; segment 2 bestaat uit vrije ontwikkelingsruimte voor overige projecten en handelingen.

Gebiedsanalyse:

een ecologische beoordeling van een Natura 2000-gebied waarin de gevolgen van het PAS zijn geanalyseerd. De gebiedsanalyses vormen in samenhang met het algemene deel van de passende beoordeling van het programma, op gebiedsniveau de passende beoordeling van het PAS.

PAS-gebied:

Natura 2000-gebied met stikstofgevoelige habitats dat in het PAS is opgenomen (momenteel zijn dit 118 van de 162 Natura 2000-gebieden in Nederland).

PAS-herstelmaatregelen:

maatregelen binnen PAS-gebieden die gericht zijn op het bestendiger maken van de natuur tegen een overbelasting van stikstof. Deze maatregelen worden in aanvulling op het regulier beheer getroffen.

PAS-bronmaatregelen:

maatregelen waarmee beoogd wordt de emissie van stikstofbronnen te verminderen. Deze bestaan uit stalmaatregelen om de emissie van stallen te verminderen, maatregelen voor emissiearme bemesting en voer- en managementmaatregelen. Deze maatregelen worden ook geduid als generieke PAS-maatregelen.

AERIUS:

softwaresysteem waarin de depositieontwikkeling per PAS-gebied wordt gemonitord, waarin de prognoses voor de depositieontwikkeling worden weergegeven, dat gebruikt wordt om te bepalen of vergunningen voor stikstofemitterende activiteiten nodig zijn en verleend kunnen worden, dat gebruikt wordt om te bepalen of een melding is vereist voor een stikstofemitterende activiteit en dat voortdurend registreert hoeveel depositieruimte en ontwikkelingsruimte aanwezig is per PAS-gebied.

Bestuurlijke keuzes en ambitieniveau

Aanleiding voor en doelen van het PAS

6.1. Aanleiding voor het PAS is het feit dat in Nederland overbelasting met stikstofdepositie een probleem vormt voor zowel de verwezenlijking van de instandhoudingsdoelstellingen voor de stikstofgevoelige habitats in veel Natura 2000-gebieden als voor het mogelijk maken van economische ontwikkelingen die stikstofdepositie veroorzaken op deze gebieden. De hoge achtergronddepositie zorgt voor een zogenoemde stikstofdeken die tot gevolg heeft dat in veel gebieden de zogenoemde kritische depositiewaarden voor de aangewezen habitattypen ruim worden overschreden. Een overschrijding van de kritische depositiewaarde betekent dat niet langer op voorhand kan worden uitgesloten dat er een risico bestaat dat de kwaliteit van habitattypen wordt aangetast als gevolg van de verzurende en/of vermestende invloed van stikstofdepositie.

Het PAS kent een zogenoemde dubbeldoelstelling, te weten enerzijds het behoud en, waar nodig, het herstel van de in het PAS opgenomen Natura 2000-gebieden om op landelijk niveau een gunstige staat van instandhouding te bereiken en anderzijds het mogelijk maken van economische ontwikkelingen die stikstofdepositie veroorzaken op deze gebieden. Door het treffen van gebiedsspecifieke herstelmaatregelen en bronmaatregelen in het PAS wordt een verbetering van de draagkracht van de natuur en een extra daling van de stikstofdepositie verwacht ten opzichte van de reeds ingezette daling van stikstofdepositie op grond van buiten het PAS getroffen maatregelen. De daling van de stikstofdepositie wordt deels ingezet voor depositie- en ontwikkelingsruimte voor het mogelijk maken van

economische ontwikkelingen. De combinatie van deze twee doelen heeft een ambitieniveau tot gevolg waarbij het eerste tijdvak van zes jaar (2015-2021) is gericht op het behoud van de stikstofgevoelige habitattypen en leefgebieden van soorten en het voorkomen van aantasting van de natuurlijke kenmerken van de gebieden. Verbetering van de kwaliteit of uitbreiding van de oppervlakte van de habitattypen of leefgebieden kan in de gevallen waarin een verbeter- en/of uitbreidingsdoelstelling geldt in een tweede of derde tijdvak van dit programma aanvangen.

Bron- en herstelmaatregelen

6.2. Stikstofdepositie is afkomstig uit verschillende buitenlandse en binnenlandse bronnen, waarbij de veehouderij de belangrijkste binnenlandse bron is. Om op termijn een gunstige staat van instandhouding van de habitattypen te bereiken, is het ondanks de reeds ingezette dalende trend in de stikstofdepositie in de afgelopen decennia, noodzakelijk dat de stikstofdepositie verder daalt. Om een verdergaande daling van de stikstofdepositie te bewerkstelligen zijn in het PAS extra bronmaatregelen opgenomen. Het gaat hierbij om stalmaatregelen, maatregelen voor emissiearme bemesting en voer- en managementmaatregelen. De effecten van deze maatregelen zijn berekend en de conclusie hiervan is dat de ammoniakemissie als gevolg van het PAS in 2020 met 13,4 kiloton per jaar zal dalen ten opzichte van de situatie zonder het PAS. Om met onzekerheden, zoals mogelijk tegenvallende resultaten, rekening te houden zijn in het PAS per maatregel marges aangehouden en is in het PAS rekening gehouden met een totale daling door deze maatregelen van 6,4 kiloton per jaar in 2021, het einde van het eerste tijdvak van 6 jaar. Daarnaast voorziet het PAS voor de daarin opgenomen gebieden in gebiedsspecifieke herstelmaatregelen die tot doel hebben de stikstofgevoelige habitats, zoals hoogveen, te versterken. De herstelmaatregelen betreffen onder meer hydrologische maatregelen en extra vegetatiemaatregelen, in aanvulling op het reguliere beheer van de Natura 2000-gebieden.

Depositie- en ontwikkelingsruimte

6.3. De berekende daling van de stikstofdepositie wordt deels ingezet als depositieruimte. Bij de bronmaatregelen die depositieruimte moeten opleveren zijn op grond van het PAS diverse marges aangehouden om eventuele tegenvallende resultaten op te kunnen vangen en te kunnen verzekeren dat de natuurlijke kenmerken niet worden aangetast.

De totale depositieruimte bestaat uit depositieruimte voor autonome ontwikkelingen en activiteiten onder de grenswaarde en uit ontwikkelingsruimte die wordt toegedeeld aan nieuwe activiteiten waarvoor voorafgaande toestemming is vereist. In onderstaande afbeelding (in kleur) uit het PAS wordt de opbouw van de totale depositieruimte geïllustreerd.

Zowel bij de ontwikkelingsruimte als bij de depositieruimte voor autonome ontwikkelingen en activiteiten onder de grenswaarde is rekening gehouden met een scenario met een economische groei van 2,5%. Voor dit scenario is gekozen om maximaal ruimte te kunnen bieden aan (nieuwe) economische ontwikkelingen en als extra buffer voor onzekerheden in de ontwikkeling van de stikstofdepositie. In onderstaande afbeelding uit het PAS wordt door middel van de blauwe en lichtblauwe delen in de buizen geïllustreerd hoe de totale depositieruimte is opgebouwd.

In het eerste tijdvak wordt de depositiedaling door de PAS-bronmaatregelen (buis 3) gedeeltelijk ongedaan gemaakt door de uitgifte van 50% hiervan als ontwikkelingsruimte, zoals verbeeld met het teruggieten in buis 4.

Gebiedsanalyses

6.4. Voor ieder in het PAS opgenomen Natura 2000-gebied is een afzonderlijke gebiedsanalyse gemaakt. De gebiedsanalyses vormen samen met het algemene deel van de passende beoordeling van het PAS op gebiedsniveau de passende beoordeling. De conclusie van deze passende beoordelingen luidt dat de depositie die in 2014 plaatsvond en de depositie die gedurende de PAS-periode van zes jaar kan gaan plaatsvinden door benutting van de depositie- en ontwikkelingsruimte, rekening houdend met de autonome daling van stikstofdepositie door bestaande, toekomstige en in het kader van het programma te treffen (extra) bronmaatregelen en bestaande en te treffen herstelmaatregelen niet zullen leiden tot een aantasting van de natuurlijke kenmerken van het betreffende Natura 2000-gebied. Volgens deze analyses kunnen op termijn de instandhoudingsdoelstellingen worden gehaald.

De gebiedsanalyses geven voor elk gebied aan wat de omvang is van de stikstofdepositie aan het begin van het eerste tijdvak van het PAS en hoe de stikstofdepositie zich zal gaan ontwikkelen bij uitvoering van het PAS. Daarnaast bevatten de gebiedsanalyses per stikstofgevoelig habitattype en leefgebied een uitwerking van de bestaande en te treffen herstelmaatregelen, en een ecologische beoordeling van de realisatie van de instandhoudingsdoelstellingen.

De rol van het PAS bij toestemmingsbesluiten

6.5. Een belangrijk doel van het PAS is het vereenvoudigen van de toestemmingverlening voor stikstofveroorzakende activiteiten. Vanaf de inwerkingtreding van het programma kan bij de verlening van toestemming voor activiteiten die stikstofdepositie veroorzaken voor het aspect stikstof gebruik worden gemaakt van het programma en de daaraan ten grondslag liggende passende beoordeling. De uitgifte van de ontwikkelingsruimte vindt plaats bij het toestemmingsbesluit voor de activiteit. Omdat de ontwikkelingsruimte passend beoordeeld is hoeft de initiatiefnemer geen aanvullende onderbouwing aan te leveren. Van de vrije ontwikkelingsruimte (segment 2) mag in de eerste helft van het PAS-tijdvak van zes jaar maximaal 60% worden toegedeeld, in de tweede helft 40%. Daarmee is beoogd te voorkomen dat reeds bij aanvang van het PAS met toestemmingsbesluiten alle ontwikkelingsruimte zou worden uitgegeven voordat de extra daling van de stikstofdepositie en het herstel van de gebieden is ingezet.

Voor activiteiten die in geringe mate bijdragen aan stikstofdepositie op een Natura 2000-gebied is na de inwerkingtreding van het PAS geen voorafgaande toestemming meer nodig. De stikstofdepositie door activiteiten die onder de drempelwaarde van 0,05 mol N/ha/jr vallen wordt opgevangen in de depositieruimte voor de autonome groei. Voor activiteiten die onder de grenswaarde van 1 mol N/ha/jr vallen is binnen de depositieruimte ruimte gereserveerd.

Monitoring en bijsturing

6.6. Om de voortgang van het PAS te volgen en de doelen op termijn te kunnen bereiken is gekozen voor een systeem van monitoring en bijsturing. Monitoring beoogt zicht te geven en te houden op de ontwikkeling van de stikstofdepositie, de beschikbare en uitgegeven depositie- en ontwikkelingsruimte, de voortgang van de uitvoering van de maatregelen in het PAS en de ontwikkeling van de stikstofgevoelige habitats. Jaarlijks zal worden voorzien in een monitorings- en bijsturingsrapportage. Als uit deze rapportage blijkt dat de doelen van het PAS in gevaar komen en de aangehouden marges worden overschreden, zoals de marge tussen 6,4 kiloton daling van ammoniakemissie waarmee het PAS in 2021 rekening houdt en 13,4 kiloton daling van ammoniakemissie die is berekend in 2020 (zie 6.2), is voorzien in de mogelijkheid van bijsturing. Deze bijsturing bestaat eruit dat bron- en/of herstelmaatregelen in het PAS kunnen worden gewijzigd, vervangen of toegevoegd en dat er minder of geen ontwikkelingsruimte voor segment 2 voor een gebied beschikbaar zal worden gesteld. Met monitoring en mogelijke bijsturing is beoogd te verzekeren dat de doelen van het PAS worden behaald binnen de bedoelde tijdvakken.

Juridische vormgeving PAS

Doelen en ambitieniveau van het PAS

6.7. De ministers van Economische Zaken en van Infrastructuur en Milieu zijn op grond van artikel 19kg, eerste lid, van de Nbw 1998 verplicht een programma vast te stellen voor de daarin opgenomen Natura 2000-gebieden ter vermindering van de stikstofdepositie in die gebieden en ter verwezenlijking van de instandhoudingsdoelstellingen voor de stikstofgevoelige habitats in die gebieden binnen afzienbare termijn. Dit programma is het hier aan de orde zijnde PAS. In het tweede lid van genoemde bepaling staat dat het programma een ambitieuze en realistische vermindering beoogt van de stikstofdepositie, afkomstig van in Nederland aanwezige bronnen. Het PAS wordt ten minste eenmaal in de zes jaar vastgesteld en geldt voor een tijdvak van zes jaar (artikel 19kg, vijfde lid, van de Nbw 1998). Op de bestuursorganen die het aangaat rust de verplichting om zorg te dragen voor een tijdige uitvoering van de bronmaatregelen en de herstelmaatregelen die in het PAS zijn opgenomen (artikel 19kj van de Nbw).

Natura 2000-gebieden mogen alleen worden opgenomen in het PAS wanneer voor deze gebieden in het programma wordt beschreven dat uit de passende beoordeling blijkt in hoeverre met maatregelen wordt voorkomen dat door ontwikkelingsruimte een verslechtering optreedt van de kwaliteit van de natuurlijke habitats en de habitats van soorten in het gebied. Voorts deint uit de passende beoordeling te blijken in hoeverre met maatregelen wordt voorkomen dat door ontwikkelingsruimte storende factoren optreden voor de soorten waarvoor het gebied is aangewezen voor zover die factoren, gelet op de instandhoudingsdoelstellingen van het gebied, een significant effect kunnen hebben (artikel 19kh, eerste lid, aanhef en onder h, sub 2 en 3 van de Nbw 1998).

Bron- en herstelmaatregelen

6.8. In het PAS moeten zowel bron- als herstelmaatregelen worden beschreven die zijn of worden getroffen om een vermindering van de stikstofdepositie te bewerkstelligen onderscheidenlijk om de instandhoudingsdoelstellingen te verwezenlijken voor de stikstofgevoelige habitats in de Natura 2000-gebieden die in het programma zijn opgenomen

(artikel 19kh, eerste lid, aanhef en onder c en g, van de Nbw 1998).

Om de beoogde vermindering van de stikstofdepositie te behalen zijn in het kader van het PAS drie bronmaatregelen voorzien. Het Besluit emissiearme huisvesting voorziet in maximale emissiewaarden voor huisvestingssystemen van agrarische bedrijven met landbouwhuisdieren. Voorts zullen de normen in het Besluit gebruik meststoffen worden aangescherpt voor de aanwending van dierlijke mest. Ten slotte zijn in de "Overeenkomst generieke maatregelen in verband met het Programma Aanpak Stikstof" voer- en managementmaatregelen opgenomen, die tot een vermindering van stikstofemissie moeten leiden. Deze overeenkomst is een convenant tussen Land- en Tuinbouw Organisatie Nederland, de Nederlandse Zuivel Organisatie, de Nederlandse Vereniging Diervoederindustrie, Cumela Nederland, de Nederlandse Melkveehouders Vakbond, de Nederlandse Vakbond Varkenshouders, de Nederlandse Vakbond Pluimveehouders en de Staatssecretaris van Economische Zaken.

In het PAS moet worden beschreven wat de verwachte autonome ontwikkelingen zijn ten aanzien van de stikstofemissie en de effecten daarvan op de omvang van de stikstofdepositie in de betrokken Natura 2000-gebieden (artikel 19kh, eerste lid, onder b, van de Nbw 1998).

Toestemmingverlening met het PAS

6.9. De vaststelling van het PAS heeft tot gevolg dat de bepalingen uit hoofdstuk III, titel 2, paragraaf 2a 'Programmatische aanpak stikstof' van de Nbw 1998, het Besluit grenswaarden en de Regeling, zoals deze luidden tot 1 januari 2017, van toepassing zijn op de beoordeling van projecten en andere handelingen die stikstofdepositie veroorzaken op stikstofgevoelige natuurwaarden in Natura 2000-gebieden. Deze bepalingen zijn op 1 juli 2015 in werking getreden. Genoemde projecten en andere handelingen zijn op grond van artikel 19d, eerste lid, van de Nbw 1998 vergunningplichtig omdat ze kunnen leiden tot een verslechtering van de kwaliteit van de natuurlijke habitats en de habitats van soorten die stikstofgevoelig zijn.

Uit artikel 19kh, zevende lid, van de Nbw 1998 gelezen in samenhang met artikel 2 van het Besluit grenswaarden volgt dat een uitzondering op de vergunningplicht in artikel 19d, eerste lid, van de Nbw 1998 geldt voor projecten en andere handelingen die geen andere negatieve gevolgen voor een Natura 2000-gebied hebben dan stikstofdepositie, terwijl die depositie een bepaalde drempel- (0,05 mol N/ha/jr) of grenswaarde (1 mol N/ha/jr) niet overschrijdt of het project of andere handeling op een grotere afstand gerekend tot het Natura 2000-gebied wordt gerealiseerd dan is vastgesteld voor hoofdwegen (3 km) of hoofdvaarwegen (5 km). Voor projecten en andere handelingen die onder de hiervoor genoemde grenswaarde vallen, maar een stikstofdepositie op een stikstofgevoelig habitat in een Natura 2000-gebied veroorzaken die hoger is dan 0,05 mol N/ha/jr geldt wel een meldingsplicht (artikel 8, eerste lid, van de Regeling). De grenswaarde wordt verlaagd naar 0,05 mol N/ha/jr wanneer uit AERIUS Register (zie hierna) blijkt dat ten aanzien van een hectare van een stikstofgevoelige habitat in het betreffende Natura 2000-gebied 5% of minder van de depositieruimte voor grenswaarden beschikbaar is (artikel 2, derde lid, van het Besluit grenswaarden).

Voor projecten en andere handelingen die de grenswaarde overschrijden geldt de vergunningplicht onverkort. Voor de beoordeling van de vergunningaanvraag dient te worden bezien of de aangevraagde activiteit tot een toename van stikstofdepositie leidt. Bepalend daarvoor is of de aangevraagde situatie leidt tot een hogere depositie dan de depositie

waarvoor eerder een Nbw-vergunning is verleend, of tot een hogere depositie dan de feitelijk veroorzaakte hoogste depositie in de periode 1 januari 2012 - 31 december 2014 (artikel 5 van de Regeling). Als het project of de andere handeling niet leidt tot een toename van stikstofdepositie, kan de vergunning onder verwijzing naar de passende beoordeling bij het PAS worden verleend. De depositie van de aangevraagde activiteit maakt in een dergelijk geval deel uit van de achtergronddepositie in 2014 die in de passende beoordeling van het PAS is betrokken.

Leidt het project of de andere handeling tot een toename van stikstofdepositie dan kan de vergunning worden verleend als het bevoegd gezag daarvoor ontwikkelingsruimte toedeelt (artikel 19km, eerste lid, van de Nbw 1998). Bij dat besluit kan worden verwezen naar de passende beoordeling van het PAS. De depositie die kan ontstaan door benutting van de (totale) ontwikkelingsruimte die in het PAS per hectare van een stikstofgevoelig habitattype in een Natura 2000-gebied is vastgesteld, is voor het PAS passend beoordeeld.

Uit artikel 19kh, negende lid, van de Nbw 1998 gelezen in samenhang met artikel 2 van het Besluit grenswaarden volgt dat bij de beoordeling van de vergunning niet wordt betrokken de depositie op stikstofgevoelige natuurwaarden in Natura 2000-gebieden die onder de hiervoor bedoelde drempel- en grenswaarde blijft. Voor bepaalde handelingen en projecten geldt dat een aanvraag voor een toestemmingsbesluit tevens geldt als als een melding (artikel 8, zevende lid, van de Regeling).

AERIUS

6.10. Het softwareprogramma AERIUS Calculator moet worden gebruikt als rekeninstrument om vast te stellen of een project of andere handeling door het veroorzaken van stikstofdepositie op een voor stikstof gevoelig habitat in een Natura 2000-gebied een verslechterend of significant verstorend effect kan hebben (artikel 2 van de Regeling). Ook voor de vaststelling van de omvang van de toe te delen ontwikkelingsruimte moet AERIUS Calculator worden gebruikt (artikel 5, eerste lid, van de Regeling). Uit artikel 5, derde lid, van de Regeling volgt dat een besluit waarbij ontwikkelingsruimte wordt toebedeeld in beginsel geldig is voor onbepaalde tijd.

Het bestuursorgaan dat ontwikkelingsruimte toebedeelt of intrekt, dient de afschrijving van de toegedeelde ontwikkelingsruimte of de bijschrijving van ontwikkelingsruimte na intrekking of vervallen van een besluit waarbij ontwikkelingsruimte is toebedeeld te registreren (artikel 19ko, eerste en tweede lid, van de Nbw 1998). Het softwareprogramma AERIUS Register is een registratie-instrument voor gegevens over de afschrijving, bijschrijving en reservering van ontwikkelingsruimte en gegevens over meldingplichtige projecten of andere handelingen (artikel 7, eerste lid, van de Regeling).

Monitoring en bijsturing

6.11. Uit artikel 19kh, eerste lid, aanhef en onder f, van de Nbw 1998 volgt dat in het PAS wordt beschreven op welke wijze de voortgang en uitvoering van de PAS-bronmaatregelen en de effecten op de stikstofdepositie worden gemonitord (zie 6.6). Drie jaar na de vaststelling van het PAS dient inzichtelijk te worden gemaakt welke ontwikkelingsruimte in de tweede helft van het PAS beschikbaar zal zijn en welke ontwikkelingsruimte naar verwachting in het volgende PAS beschikbaar zal zijn (artikel 19kha, van de Nbw 1998).

Indien uit de monitoring blijkt dat dit noodzakelijk is, kunnen op grond van artikel 19ki, eerste lid, bron- en herstelmaatregelen worden vervangen of toegevoegd aan het PAS en kan de uit te geven ontwikkelingsruimte worden bijgesteld.

Beschrijving AERIUS

6.12. Om de hierboven beschreven doelen te bereiken en om te voldoen aan de eisen die zijn opgenomen in de hierboven beschreven wettelijke regeling zijn instrumenten ontwikkeld die gebruikt worden om de huidige depositie van stikstof te bepalen, de ontwikkeling van de stikstofdepositie te monitoren, prognoses te maken voor de verwachte ontwikkeling van de depositie en om een beoordeling te maken van de vergunningplicht en vergunningverlening voor activiteiten die stikstofdepositie kunnen veroorzaken. Dit betreft het softwarepakket AERIUS dat via de website www.aerius.nl openbaar toegankelijk is. Drie modules van dit pakket zijn van belang in deze zaak.

6.13. Ten eerste is dit AERIUS Calculator. Dit rekeninstrument maakt een deels geautomatiseerde besluitvorming mogelijk. Het berekent de depositiebijdrage van emissiebronnen die een gebruiker in het systeem invoert of importeert en wordt op grond van artikel 2, eerste lid, van de Regeling gebruikt voor de vaststelling of een project of een andere handeling door het veroorzaken van stikstofdepositie op een voor stikstof gevoelig habitat in een Natura 2000-gebied een verslechterend of significant verstorend effect kan hebben. Het resultaat van de berekening geeft inzicht in de depositiebijdrage van de ingevoerde bronnen op vaste rekenpunten binnen Natura 2000-gebieden of op rekenpunten die de gebruiker zelf heeft gedefinieerd. AERIUS Calculator biedt op deze wijze inzicht of een ingevoerde ontwikkeling stikstofdepositie veroorzaakt die de drempel- of grenswaarde niet overschrijdt en daardoor is uitgezonderd van de vergunningplicht, of de grenswaarde wel overschrijdt, waardoor deze vergunningplichtig is. AERIUS Calculator kan voor een ingevoerde ontwikkeling, zoals de vestiging of uitbreiding van een veehouderij, weergeven of nog voldoende ontwikkelingsruimte aanwezig is in de Natura 2000-gebieden waarop de ingevoerde emissiebron stikstofdepositie veroorzaakt. Calculator biedt verder de mogelijkheid om de rekenresultaten te exporteren als een document dat kan worden gebruikt als bijlage bij de vergunningaanvraag.

6.14. Ten tweede is dit AERIUS Register. Dit wordt door het bevoegd gezag gebruikt bij het beheer van de depositieruimte die is berekend met AERIUS Monitor. Zie rechtsoverweging 6.3 hierboven voor de rol die de depositieruimte speelt binnen het PAS.

6.15. Ten derde is dit AERIUS Monitor. Hiermee wordt de uitvoering van het PAS gevolgd. Monitor is een rekeninstrument dat op hectareniveau inzicht geeft in:

- de depositietrend: de verwachte depositieontwikkeling in de tijd. Deze kan gevarieerd worden, afhangende van de gekozen beleidsscenario's.
- de extra daling die bereikt wordt met het PAS: Monitor geeft het effect van het PAS-beleid weer op de emissies en depositie van stikstof.
- de depositieruimte en ontwikkelingsruimte: Monitor kan gebruikt worden om inzicht te krijgen in het deel van de totale depositie dat beschikbaar is voor nieuwe ontwikkelingen.
- een confrontatie tussen depositieruimte en ontwikkelingsbehoefte: de verwachte overschotten of tekorten aan ontwikkelingsruimte kunnen met Monitor inzichtelijk gemaakt worden.

6.16. Bij het softwarepakket AERIUS is een onderbouwing gevoegd van de wijze waarop de gepresenteerde resultaten tot stand komen en op welke gegevens deze resultaten zijn gebaseerd. Deze onderbouwing staat in 190 factsheets die bij het softwareprogramma op de website beschikbaar zijn gesteld.

6.17. Op onderstaande afbeeldingen is bij wijze van voorbeeld te zien hoe AERIUS Monitor de gegevens over stikstofdepositie weergeeft voor een deel van het Natura 2000-gebied Groote Peel, één van de Natura 2000-gebieden die thans in geding zijn. Dit gebied is verdeeld in hexagonen die ieder een hectare binnen het gebied beslaan. AERIUS Monitor is in staat om per hexagoon de stikstofdepositie weer te geven. De betekenis van de kleuren van de hexagonen in de eerste afbeelding is de totale stikstofdepositie die in 2014 feitelijk plaatsvond en in de tweede afbeelding de totale stikstofdepositie die in 2020 wordt verwacht. Hoe donkerder de kleur is, hoe hoger de verwachte depositie. Rondom de geselecteerde hexagoon in de kaart (weergegeven in de kleur paars met een "i" erbij) is dit een depositie van 1300-1600 mol N/ha/jr in 2014. In 2020 wordt in AERIUS Monitor bij vijf van de omliggende hexagonen een depositie van 1000-1300 mol/ha/jr verwacht en bij één hexagoon 1300-1600 mol N/ha/jr.

F. VERHOUDING PAS TOT ARTIKEL 6 VAN DE HABITATRICHTLIJN

Het toepasselijke recht

Recht van de Europese Unie

7. Habitatrichtlijn

Artikel 2:

"1. Deze richtlijn heeft tot doel bij te dragen tot het waarborgen van de biologische diversiteit door het instandhouden van de natuurlijke habitats en de wilde flora en fauna op het Europese grondgebied van de Lid-Staten waarop het Verdrag van toepassing is.

2. De op grond van deze richtlijn genomen maatregelen beogen de natuurlijke habitats en de wilde dier- en plantensoorten van communautair belang in een gunstige staat van instandhouding te behouden of te herstellen.

3. In de op grond van deze richtlijn genomen maatregelen wordt rekening gehouden met de vereisten op economisch, sociaal en cultureel gebied, en met de regionale en lokale bijzonderheden".

Artikel 6:

"1. De Lid-Staten treffen voor de speciale beschermingszones de nodige instandhoudingsmaatregelen; deze behelzen zo nodig passende specifieke of van ruimtelijke-ordeningsplannen deel uitmakende beheersplannen en passende wettelijke,

bestuursrechtelijke of op een overeenkomst berustende maatregelen, die beantwoorden aan de ecologische vereisten van de typen natuurlijke habitats van bijlage I en de soorten van bijlage II die in die gebieden voorkomen.

2. De Lid-Staten treffen passende maatregelen om ervoor te zorgen dat de kwaliteit van de natuurlijke habitats en de habitats van soorten in de speciale beschermingszones niet verslechtert en er geen storende factoren optreden voor de soorten waarvoor de zones zijn aangewezen voor zover die factoren, gelet op de doelstellingen van deze richtlijn een significant effect zouden kunnen hebben.

3. Voor elk plan of project dat niet direct verband houdt met of nodig is voor het beheer van het gebied, maar afzonderlijk of in combinatie met andere plannen of projecten significante gevolgen kan hebben voor zo'n gebied, wordt een passende beoordeling gemaakt van de gevolgen voor het gebied, rekening houdend met de instandhoudingsdoelstellingen van dat gebied. Gelet op de conclusies van de beoordeling van de gevolgen voor het gebied en onder voorbehoud van het bepaalde in lid 4, geven de bevoegde nationale instanties slechts toestemming voor dat plan of project nadat zij de zekerheid hebben verkregen dat het de natuurlijke kenmerken van het betrokken gebied niet zal aantasten en nadat zij in voorkomend geval inspraakmogelijkheden hebben geboden.

4. Indien een plan of project, ondanks negatieve conclusies van de beoordeling van de gevolgen voor het gebied, bij ontstentenis van alternatieve oplossingen, om dwingende redenen van groot openbaar belang, met inbegrip van redenen van sociale of economische aard, toch moet worden gerealiseerd, neemt de Lid-Staat alle nodige compenserende maatregelen om te waarborgen dat de algehele samenhang van Natura 2000 bewaard blijft. De Lid-Staat stelt de Commissie op de hoogte van de genomen compenserende maatregelen.

Wanneer het betrokken gebied een gebied met een prioritair type natuurlijke habitat en/of een prioritaire soort is, kunnen alleen argumenten die verband houden met de menselijke gezondheid, de openbare veiligheid of met voor het milieu wezenlijke gunstige effecten dan wel, na advies van de Commissie, andere dwingende redenen van groot openbaar belang worden aangevoerd".

Nationaal recht

7.1. Voor de beschrijving van het in de Nbw 1998, het Besluit grenswaarden en de Regeling geregelde toestemmingsregime wordt verwezen naar 6.9. De relevante nationale bepalingen luiden als volgt.

7.2. Natuurbeschermingswet 1998 (Nbw 1998)

Artikel 19d, eerste lid:

"Het is verboden zonder vergunning […] van gedeputeerde staten […] projecten of andere handelingen te realiseren onderscheidenlijk te verrichten die gelet op de instandhoudingsdoelstelling […] de kwaliteit van de natuurlijke habitats en de habitats van soorten in een Natura 2000-gebied kunnen verslechteren of een significant verstorend effect kunnen hebben op de soorten waarvoor het gebied is aangewezen. […]".

Artikel 19e:

"Gedeputeerde staten houden bij het verlenen van een vergunning als bedoeld in artikel 19d, eerste lid, rekening (a) met de gevolgen die een project of andere handeling, waarop de vergunningaanvraag betrekking heeft, gelet op de instandhoudingsdoelstelling […] kan hebben voor een Natura 2000-gebied".

Artikel 19f, eerste lid:

"Voor projecten waarover gedeputeerde staten een besluit op een aanvraag voor een vergunning als bedoeld in artikel 19d, eerste lid, nemen, en die niet direct verband houden met of nodig zijn voor het beheer van een Natura 2000-gebied maar die afzonderlijk of in combinatie met andere projecten of plannen significante gevolgen kunnen hebben voor het desbetreffende gebied, maakt de initiatiefnemer alvorens gedeputeerde staten een besluit nemen, een passende beoordeling van de gevolgen voor het gebied waarbij rekening wordt gehouden met de instandhoudingsdoelstelling […], van dat gebied".

Artikel 19g, eerste lid:

"Indien een passende beoordeling is voorgeschreven op grond van artikel 19f, eerste lid, kan een vergunning als bedoeld in artikel 19d, eerste lid, slechts worden verleend indien gedeputeerde staten zich op grond van de passende beoordeling ervan hebben verzekerd dat de natuurlijke kenmerken van het gebied niet zullen worden aangetast".

Artikel 19kg, eerste, tweede en vijfde lid:

"1. Onze Minister en Onze Minister van Infrastructuur en Milieu stellen een programma vast voor de daarin opgenomen Natura 2000-gebieden ter vermindering van de stikstofdepositie in die gebieden en ter verwezenlijking van de instandhoudingsdoelstellingen voor de voor stikstof gevoelige habitats in die gebieden binnen afzienbare termijn.

2. Het programma beoogt een ambitieuze en realistische vermindering van de stikstofdepositie, afkomstig van in Nederland aanwezige bronnen.

[…]

5.Het programma wordt ten minste eenmaal in de zes jaar vastgesteld en geldt voor een tijdvak van zes jaar".

Artikel 19kh, eerste, vierde, zevende, achtste en negende lid:

"1. In een programma als bedoeld in artikel 19kg worden voor de betrokken Natura 2000-gebieden in elk geval beschreven of genoemd:

a. de omvang van de stikstofdepositie aan het begin van het tijdvak van het programma […]

b. de verwachte autonome ontwikkelingen ten aanzien van de stikstofemissie door de factoren, bedoeld in onderdeel a, en de effecten daarvan op de omvang van stikstofdepositie

in de gebieden;

c. de getroffen of te treffen maatregelen die bijdragen aan een vermindering van de stikstofdepositie, of die op een andere wijze bijdragen aan het bereiken van een goede staat van instandhouding van de voor stikstof gevoelige habitats, en de verwachte effecten van die maatregelen op de omvang van de depositie, onderscheidenlijk het bereiken van een goede staat van instandhouding in de gebieden;

d. de sociaal-economische evaluatie en weging van haalbaarheid en betaalbaarheid van maatregelen als bedoeld in de onderdelen c en g;

e. de doelstellingen ten aanzien van de omvang van de stikstofdepositie, al dan niet met tussendoelstellingen, of de indicatoren waaruit kan worden afgeleid of een doelstelling al dan niet is behaald welke noodzakelijk zijn met het oog op het bereiken van een gunstige staat van instandhouding;

f. de wijze waarop en frequentie waarmee de rapportage plaatsvindt over de voortgang en uitvoering van de getroffen of te treffen in het programma beschreven en genoemde maatregelen en de effecten daarvan op de depositie;

g. de getroffen of te treffen maatregelen ter verwezenlijking van de instandhoudingsdoelstellingen voor de voor stikstof gevoelige habitats in de Natura 2000-gebieden die in het programma zijn opgenomen;

h. de resultaten van een beoordeling voor elk Natura 2000-gebied dat in het programma is opgenomen, in hoeverre de maatregelen, bedoeld in de onderdelen c en g, rekening houdend met de verwachte algemene ontwikkeling van de stikstofdepositie, in het bijzonder het totaal van de stikstofdeposities, bedoeld in het zevende en negende lid, en de ontwikkelingsruimte:

1°. bijdragen aan de verwezenlijking van de instandhoudingsdoelstellingen voor de voor stikstof gevoelige habitats in het gebied;

2°. voorkomen dat verslechtering optreedt van de kwaliteit van de natuurlijke habitats en de habitats van soorten in het gebied;

3°. voorkomen dat storende factoren optreden voor de soorten waarvoor het gebied is aangewezen voor zover die factoren, gelet op de instandhoudingsdoelstellingen van het gebied, een significant effect kunnen hebben, en

4°. de verwezenlijking van de instandhoudingsdoelstellingen van het gebied die geen betrekking hebben op voor stikstof gevoelige habitats, niet in gevaar brengen.

4. In het programma worden de uitgangspunten opgenomen voor de bepaling van de ontwikkelingsruimte en voor de toedeling en reservering van ontwikkelingsruimte. In het programma wordt de op het tijdstip van vaststelling van het programma beschikbare ontwikkelingsruimte vermeld.

7. Het verbod, bedoeld in artikel 19d, eerste lid, is met betrekking tot een Natura 2000-gebied niet van toepassing op een project of andere handeling dat voldoet aan elk van de volgende

voorwaarden:

a. het project of de handeling:

1°. veroorzaakt een stikstofdepositie op voor stikstof gevoelige habitats in het Natura 2000-gebied die afzonderlijk en, ingeval het project of de handeling betrekking heeft op een inrichting als bedoeld in artikel 1.1, derde lid, van de Wet milieubeheer, in cumulatie met andere projecten of handelingen met betrekking tot dezelfde inrichting in de periode waarvoor het programma geldt, niet een waarde die is vastgesteld bij algemene maatregel van bestuur overschrijdt, of […]

b. het project of de handeling kan voor het desbetreffende Natura 2000-gebied geen andere gevolgen veroorzaken dan stikstofdepositie die, gelet op de instandhoudingsdoelstellingen, de kwaliteit van de natuurlijke habitats en de habitats van soorten in een Natura 2000-gebied kunnen verslechteren of een significant verstorend effect kunnen hebben op de soorten waarvoor het gebied is aangewezen.

8. De waarde […] bedoeld in het zevende lid, onderdeel a, kan voor de onderscheiden Natura 2000-gebieden verschillend worden vastgesteld. De waarde […] wordt zodanig vastgesteld dat:

[…]

2°. op voorhand op grond van objectieve gegevens kan worden uitgesloten dat projecten of andere handelingen als bedoeld in het zevende lid afzonderlijk of in combinatie met andere plannen of projecten de natuurlijke kenmerken van een Natura 2000-gebied zullen aantasten.

9. Bij het nemen van een besluit als bedoeld in artikel 19km, eerste lid, betrekt het bevoegd gezag niet de stikstofdepositie die het project of de andere handeling veroorzaakt op voor stikstof gevoelige habitats in een Natura 2000-gebied, indien de stikstofdepositie de in het zevende lid, onderdeel a, bedoelde waarde niet overschrijdt, onderscheidenlijk indien het project of de handeling wordt gerealiseerd, onderscheidenlijk verricht op een grotere afstand dan de op grond van het zevende lid, onderdeel a, vastgestelde afstand".

Artikel 19kha:

"In het programma wordt, ter uitvoering van artikel 19kh, eerste lid, onderdeel f, in elk geval beschreven dat drie jaar nadat het programma is vastgesteld, Onze Minister en Onze Minister van Infrastructuur en Milieu, na overleg met de bestuursorganen, bedoeld in artikel 19kg, derde lid, de ontwikkelingsruimte inzichtelijk maken die:

a. de tweede helft van het programma beschikbaar zal zijn;

b. naar verwachting in het daarop volgende programma beschikbaar zal zijn, in het bijzonder in de eerste helft van het tijdvak van dat programma".

Artikel 19ki,eerste lid:

"1. Onze Minister en Onze Minister van Infrastructuur en Milieu kunnen in het programma

maatregelen als bedoeld in artikel 19kh, eerste lid, onderdeel c of g, wijzigen of door andere maatregelen vervangen, dan wel maatregelen toevoegen. Zij stellen voor de Natura 2000-gebieden waarop de maatregelen betrekking hebben in het programma vast wat daarvan de gevolgen zijn voor de ontwikkelingsruimte en voor de beoordeling, bedoeld in artikel 19kh, eerste lid, onderdeel h. Ingeval het wijzigen, vervangen of toevoegen van maatregelen leidt tot minder ontwikkelingsruimte met betrekking tot een Natura 2000-gebied, geschiedt dit in overeenstemming met de bestuursorganen die het beheerplan, bedoeld in artikel 19a of 19b, voor dat gebied vaststellen".

Artikel 19kj:

"De bestuursorganen die het aangaat dragen zorg voor een tijdige uitvoering van de in het programma opgenomen maatregelen, bedoeld in artikel 19kh, eerste lid, onderdelen c en g".

Artikel 19km, eerste lid:

"De ontwikkelingsruimte voor een in het programma opgenomen Natura 2000-gebied, kan […] overeenkomstig de uitgangspunten, bedoeld in artikel 19kh, vierde lid, door het bestuursorgaan dat bevoegd is tot het nemen van het desbetreffende besluit, worden toegedeeld in:

[…]

b. een vergunning als bedoeld in artikel 19d, eerste lid;"

Artikel 19ko, eerste lid:

"1. Een bestuursorgaan dat ontwikkelingsruimte toedeelt in een besluit als bedoeld in artikel 19km, eerste lid, […], draagt tijdig zorg voor een nauwkeurige en volledige registratie van de afschrijving van de toegedeelde ontwikkelingsruimte van de ontwikkelingsruimte die beschikbaar is voor projecten en andere handelingen die stikstofdepositie veroorzaken in de Natura 2000-gebieden waarop het besluit betrekking heeft.[…]".

7.3. Besluit grenswaarden programmatische aanpak (Besluit grenswaarden)

Artikel 2

"1. De waarde, bedoeld in artikel 19kh, zevende lid, onderdeel a, onder 1°, van de wet, is 1 mol per hectare per jaar".

[…]

3. In afwijking van het eerste lid is de waarde, bedoeld in artikel 19kh, zevende lid, onderdeel a, onder 1°, van de wet, voor een project of andere handeling, niet zijnde een project of andere handeling als bedoeld in artikel 19kn, eerste lid, van de wet, 0,05 mol per hectare per jaar, zolang uit het krachtens artikel 19kb voorgeschreven rekenmodel blijkt dat ten aanzien van een hectare van een voor stikstof gevoelige habitat in het desbetreffende Natura 2000-gebied 5% of minder van de depositieruimte voor grenswaarden beschikbaar is".

7.4. Regeling programmatische aanpak stikstof (Regeling)

Artikel 2, eerste lid:

"1. Voor de vaststelling of een project of een andere handeling als bedoeld in artikel 19d, eerste lid, van de wet, of een bestemmingsplan als bedoeld in artikel 19db, eerste lid, van de wet, door het veroorzaken van stikstofdepositie op een voor stikstof gevoelig habitat in een Natura 2000-gebied een verslechterend of significant verstorend effect kan hebben, wordt de stikstofdepositie berekend met gebruikmaking van AERIUS Calculator".

Artikel 5, eerste, tweede en derde lid:

"1. Het bevoegd gezag stelt de omvang van de in een toestemmingsbesluit toe te delen ontwikkelingsruimte vast met gebruikmaking van AERIUS Calculator.

2 De ontwikkelingsruimte die het bevoegd gezag toedeelt in een toestemmingsbesluit is gelijk aan de toename van de stikstofdepositie op een hectare van een voor stikstof gevoelig habitat in een Natura 2000-gebied die een project of andere handeling per kalenderjaar kan veroorzaken, uitgaande van het jaar waarin de depositie als gevolg van dat project of die andere handeling het hoogst is.

3. In een toestemmingsbesluit dat geldig is voor onbepaalde tijd kent het bevoegd gezag ontwikkelingsruimte eenmalig toe voor onbepaalde tijd".

Artikel 7, eerste, tweede en derde lid:

"1. Er is een registratie-instrument waarin gegevens worden opgenomen die betrekking hebben op de afschrijving, bijschrijving en reservering van ontwikkelingsruimte en op meldingen als bedoeld in artikel 8.

2. Bij aanvang van het programma draagt de minister er zorg voor dat de beschikbare ontwikkelingsruimte in AERIUS Register wordt opgenomen. Dat gebeurt ook bij wijzigingen van het programma.

3. De registraties, bedoeld in artikel 19ko, eerste, tweede, derde en vierde lid, van de wet, geschieden in AERIUS Register, terstond nadat een toestemmingsbesluit is genomen, ingetrokken of vervallen of terstond nadat ontwikkelingsruimte in deze regeling is gereserveerd of een reservering van ontwikkelingsruimte in deze regeling is gewijzigd of vervallen".

Artikel 8, eerste, zevende en achtste lid

"1. Degene die voornemens is een project te realiseren of een andere handeling te verrichten waarop artikel 19kh, zevende lid, onderdeel a, onder 1°, van de wet van toepassing is doet ten minste vier weken maar ten hoogste twee jaar voor de aanvang daarvan een melding, indien is voldaan aan elk van de volgende voorwaarden:

a.1°. het project of de andere handeling heeft betrekking op de oprichting, verandering of uitbreiding van een inrichting als bedoeld in artikel 1.1, derde lid, van de Wet milieubeheer

bestemd voor landbouw, industrie of het gebruik van gemotoriseerde voertuigen voor wedstrijden,

[…]

b. het project of de andere handeling veroorzaakt stikstofdepositie op een voor stikstof gevoelig habitat in een Natura 2000-gebied die hoger is dan 0,05 mol per hectare per jaar.

7. Een aanvraag voor een toestemmingsbesluit waarin gedurende het tijdvak waarvoor het programma is vastgesteld ontwikkelingsruimte is toegedeeld aan een project of een andere handeling, geldt tevens als een melding als bedoeld in het eerste lid ten aanzien van Natura 2000-gebieden waarop het project of de andere handeling stikstofdepositie veroorzaakt die lager is dan of gelijk is aan de waarde, bedoeld in artikel 2, eerste lid, van het Besluit grenswaarden programmatische aanpak stikstof.

8. Het bestuursorgaan waarbij een melding is gedaan registreert de melding terstond na de ontvangst daarvan in AERIUS Register".

Relatie ambitieniveau tot artikel 6 van de Habitatrichtlijn

Beroepsgrond en standpunt college

8. De Werkgroep stelt dat met het PAS niet kan worden voldaan aan de verplichtingen die uit artikel 6 van de Habitatrichtlijn voortvloeien omdat het ambitieniveau van het PAS, dat inzet op het voorkomen van verslechtering in de eerste PAS-periode van zes jaar en het bijdragen aan herstel in de tweede en/of derde PAS-periode daarvoor te mager is. Het bieden van enig uitzicht op herstel van de kwaliteit van de habitats ondanks een blijvende zeer hoge overschrijding van de kritische depositiewaarde na drie PAS-periodes in de Natura 2000-gebieden Groote Peel en Deurnsche Peel & Mariapeel, biedt volgens de Werkgroep geen uitzicht op het bereiken van een gunstige staat van instandhouding.

8.1. Het college stelt dat uit de gebiedsanalyses volgt dat in het eerste tijdvak van het PAS behoud van natuurwaarden is geborgd en dat de kwaliteit van de natuurwaarden na het eerste tijdvak zodanig is dat verbetering of uitbreiding daarvan in de volgende tijdvakken mogelijk is. De volgende twee tijdvakken van zes jaar zijn gericht op behoud en waar nodig (verder) herstel van de habitattypen en leefgebieden van soorten, mede met het oog op het bereiken van een landelijk goede staat van instandhouding. Om ervoor te zorgen dat dit mogelijk is, is in de gebiedsanalyse een doorkijk gegeven naar de toekomst. Het is dus niet zo dat slechts tot 2021 een waarborg bestaat dat geen verslechtering optreedt. Ook na die periode, tijdens het tweede en derde tijdvak, is minimaal behoud gegarandeerd. Verder kijken dan drie tijdvakken is in ecologisch opzicht te speculatief, aldus het college.

Het college wijst er verder op dat de stikstofdepositie in de Natura 2000-gebieden gedurende de verschillende tijdvakken van het PAS zal dalen door de autonome daling en door de PAS-bronmaatregelen. Het PAS zorgt er volgens het college voor dat er in ieder geval nergens een toename zal plaatsvinden op stikstofgevoelige natuur ten opzichte van de autonome situatie. Een deel van de extra afname komt ten goede aan de natuur. De overige extra afname van de stikstofemissie wordt ingezet om projecten en andere handelingen mogelijk te maken. Door de uitgifte van depositieruimte voor grenswaarden en

ontwikkelingsruimte voor projecten en andere handelingen neemt de stikstofdepositie langzamer af dan het geval zou zijn zonder de uitgifte van depositieruimte voor grenswaarden en ontwikkelingsruimte. Dit leidt evenwel niet tot onevenredige vertraging in het bereiken van de instandhoudingsdoelstellingen, aldus het college.

Beoordeling door de Afdeling

8.2. De Afdeling is van oordeel dat het ambitieniveau van het PAS verenigbaar is met de verplichtingen die voortvloeien uit de Habitatrichtlijn. Daartoe overweegt zij het volgende.

In de considerans van de Habitatrichtlijn is aangegeven dat de Habitatrichtlijn tot hoofddoel heeft, met inachtneming van de vereisten op economisch, sociaal, cultureel en regionaal gebied, het behoud van de biologische diversiteit te bevorderen. Dit is uitgewerkt in artikel 2, derde lid, van de Habitatrichtlijn waarin is bepaald dat in de op grond van de Habitatrichtlijn genomen maatregelen rekening wordt gehouden met de vereisten op economisch en sociaal gebied. De Afdeling leidt hieruit af dat de Habitatrichtlijn de ruimte biedt voor de keuze om een deel van de autonome daling van de stikstofdepositie en 50% van de daling van de stikstofdepositie door de PAS-bronmaatregelen in te zetten voor economische activiteiten, waardoor de stikstofdepositie langzamer afneemt dan het geval zou zijn zonder de uitgifte van depositie- en ontwikkelingsruimte. Deze keuze houdt immers verband met de in artikel 2, derde lid, van de Habitatrichtlijn genoemde vereisten op economisch en sociaal gebied.

De verplichting die uit artikel 6 van de Habitatrichtlijn voortvloeit om de gunstige staat van instandhouding van de habitattypen en soorten te herstellen is niet aan een termijn gebonden. Hoewel artikel 6, eerste lid, van de Habitatrichtlijn een resultaatsverplichting inhoudt, is het aan de lidstaten te bepalen op welke wijze en in welk tempo hieraan uitvoering wordt gegeven, waarbij ook het hiervoor vermelde geldt, dat bij de maatregelen die met het oog daarop worden getroffen rekening wordt gehouden met de vereisten op economisch, sociaal en cultureel gebied. De Afdeling is van oordeel dat met het ambitieniveau van het PAS de grenzen van de aan de lidstaat op dit punt gegeven beoordelingsruimte niet worden overschreden.

Het ambitieniveau komt voorts niet in strijd met de uit artikel 6, tweede lid, van de Habitatrichtlijn voortvloeiende verplichting om passende maatregelen te treffen ter voorkoming van verslechtering en significante verstoring van de waarden waarvoor een Natura 2000-gebied is aangewezen. Deze verplichting strekt immers niet tot herstel van de gunstige staat van instandhouding. Datzelfde geldt ook voor zover binnen het ambitieniveau ruimte wordt geboden voor nieuwe economische activiteiten. De passende beoordeling van plannen en projecten dient de zekerheid te bieden dat het plan of project de natuurlijke kenmerken van een Natura 2000-gebied niet zal aantasten. Een plan of project hoeft niet bij te dragen aan het herstel van natuurwaarden van een Natura 2000-gebied.

Het betoog faalt.

Het vereiste van een individuele toestemming of individuele beoordeling

Beroepsgronden en standpunt college

9. De Werkgroep betoogt dat de vergunningen voor de zes agrarische bedrijven niet verleend

hadden kunnen worden onder verwijzing naar het PAS, omdat artikel 6, derde lid, van de Habitatrichtlijn een individuele beoordeling van projecten die significante gevolgen kunnen hebben vereist. De passende beoordeling van het PAS heeft geen betrekking op specifieke projecten, maar op een programma. De beoordeling van herstelmaatregelen en ingecalculeerde vervuilende handelingen die in het PAS heeft plaatsgevonden, kan volgens de Werkgroep niet ten grondslag worden gelegd aan de vergunning voor een individueel project, omdat dat project op zich zelf dient te voldoen aan de vereisten die artikel 6, derde lid, van de Habitatrichtlijn stelt. In dit verband wijst de Werkgroep op het arrest van het Hof van Justitie van 1 juli 2015, Wezer, ECLI:EU:C:2015:433.

De Werkgroep betoogt verder dat deposities die onder de drempel- of grenswaarde blijven van 0,05 onderscheidenlijk 1 mol N/ha/jr significante gevolgen kunnen hebben voor stikstofgevoelige natuurwaarden. Projecten die dergelijke deposities veroorzaken dienen passend beoordeeld te worden. De passende beoordeling van het PAS die geen betrekking heeft op specifieke projecten kan volgens de Werkgroep niet ten grondslag worden gelegd aan de regeling die erin voorziet dat dergelijke deposities bij de vergunningverlening niet worden beoordeeld, zoals is gebeurd in de aan de orde zijnde vergunningzaken, en in sommige gevallen zonder vergunning zijn toegestaan. Deze regeling waarborgt volgens de Werkgroep niet een juiste uitvoering van artikel 6 van de Habitatrichtlijn.

9.1. Het college stelt dat het PAS en de daarbij behorende regelgeving op een juiste wijze uitvoering geven aan artikel 6 van de Habitatrichtlijn. In het PAS heeft een toetsing plaatsgevonden van een bepaalde belasting van stikstofdepositie in relatie tot de instandhoudingsdoelstellingen in alle afzonderlijke Natura 2000-gebieden met stikstofgevoelige natuurwaarden, zoals artikel 6, derde lid, van de Habitatrichtlijn, vereist. In de gebiedsanalyses is voor alle locaties met stikstofgevoelige natuurwaarden beoordeeld of er wetenschappelijk gezien redelijkerwijs geen twijfel is dat met het beschikbaar stellen van depositie- en ontwikkelingsruimte voor projecten en andere handelingen, rekening houdend met de bron- en herstelmaatregelen van het programma, de instandhoudingsdoelstellingen voor de stikstofgevoelige natuurwaarden op termijn worden gehaald en tevens dat behoud is geborgd. Indien uitbreiding van oppervlakte of verbetering van kwaliteit een doelstelling is, is beoordeeld of dit kan aanvangen in het huidige tijdvak van het programma, dan wel in een volgend tijdvak. De hoeveelheid depositie- en ontwikkelingsruimte die beschikbaar is gesteld voor alle projecten en andere handelingen die op grond van het PAS mogelijk worden gemaakt is dus passend beoordeeld. Uit de passende beoordeling volgt dat de kwaliteit van de habitattypen niet zal verslechteren en dat de natuurlijke kenmerken van de Natura 2000-gebieden niet zullen worden aangetast.

Verder stelt het college dat het Wezerarrest van het Hof van Justitie, waarnaar de Werkgroep verwijst, geen betrekking heeft op de Habitatrichtlijn, maar op Richtlijn 2000/60/EG van het Europees Parlement en de Raad van 23 oktober 2000 tot vaststelling van een kader voor communautaire maatregelen betreffende het waterbeleid (PbEG2000 L327; hierna: Kaderrichtlijn Water). Uit dit arrest kan niet worden afgeleid dat een programmatische aanpak op grond van de Habitatrichtlijn niet is toegestaan.

Het PAS-beoordelingskader

9.2. Het PAS en de bijbehorende regelgeving in de Nbw 1998, het Besluit grenswaarden en de Regeling bieden een beoordelingskader voor stikstofveroorzakende projecten en andere

handelingen. De beschrijving hiervan is opgenomen in 6.9. Het beoordelingskader houdt in dat:

(a) projecten en andere handelingen die stikstofdepositie veroorzaken die de drempelwaarde van 0,05 mol N/ha/jr niet overschrijden zonder voorafgaande toestemming zijn toegestaan;

(b) projecten en andere handelingen die stikstofdepositie veroorzaken die de grenswaarde van 0,05 - 1 mol N/ha/jr niet overschrijden, zonder voorafgaande toestemming zijn toegestaan, zij het dat in bepaalde gevallen wel een meldingsplicht geldt;

(c) projecten en andere handelingen die stikstofdepositie veroorzaken boven de grenswaarde vergunningplichtig zijn. De vergunning kan worden verleend als deze projecten en andere handelingen niet leiden tot een toename van stikstofdepositie, of, indien de projecten en andere handelingen wel tot een toename van stikstofdepositie leiden, als voor die toename door het bevoegd gezag ontwikkelingsruimte wordt toegedeeld.

De stikstofdepositie die veroorzaakt wordt door een project of andere handeling wordt berekend met het verplicht voorgeschreven rekeninstrument AERIUS. De rol van AERIUS in het beoordelingskader is beschreven in 6.13.

9.3. Aan het beoordelingskader ligt ten grondslag dat voor het PAS voor elk Natura 2000-gebied een passende beoordeling is gemaakt waarin is onderzocht of de stikstofdepositie die in 2014 plaatsvond en de depositie die gedurende de PAS-periode van zes jaar kan gaan plaatsvinden na benutting van de depositie- en ontwikkelingsruimte de natuurlijke kenmerken van het betrokken Natura 2000-gebied niet zullen aantasten. Daarbij is rekening gehouden met de autonome daling van de stikstofdepositie en de daling van de stikstofdepositie door de PAS-bronmaatregelen, alsmede met de herstelmaatregelen. Voor de depositie door activiteiten die de drempel- of grenswaarde niet overschrijden (categorie a en b) is in het PAS depositieruimte gereserveerd. Voor activiteiten waarvoor een vergunning is vereist (categorie c) geldt dat deze kan worden verleend als de depositie past binnen de totale hoeveelheid stikstofdepositie die voor het PAS passend is beoordeeld.

Aanleiding prejudiciële vragen

9.4. Het beoordelingskader voor stikstofveroorzakende activiteiten dat met het PAS van kracht is geworden strekt ertoe dat voor projecten en andere handelingen die stikstofdepositie veroorzaken die de drempel- of grenswaarde niet overschrijden, geen individuele toestemming is vereist. De gevolgen van alle projecten en andere handelingen tezamen die gebruik kunnen maken van de uitzondering op de vergunningplicht is in de passende beoordeling voor het PAS betrokken.

Voor projecten en andere handelingen die stikstofdepositie veroorzaken die de grenswaarde overschrijden is wel een individuele toestemming vereist (vergunning), maar hoeft geen individuele passende beoordeling te worden gemaakt. De passende beoordeling die voor het PAS is gemaakt wordt ten grondslag gelegd aan de verlening van de vergunning.

9.5. Voor de activiteiten die uitgezonderd zijn van de vergunningplicht (categorie a en b), rijst de vraag of artikel 6 van de Habitatrichtlijn zich verzet tegen een wettelijke regeling waarin projecten en andere handelingen zonder individuele toestemming worden toegestaan, ervan

uitgaande dat aan die wettelijke regeling een passende beoordeling ten grondslag ligt waarin de gevolgen van de projecten en andere handelingen die gebruik kunnen maken van die regeling zijn onderzocht.

Voor de vergunningplichtige activiteiten (categorie c) rijst de vraag of artikel 6 van de Habitatrichtlijn zich ertegen verzet dat de passende beoordeling van een programma, in dit geval het PAS, ten grondslag wordt gelegd aan de verlening van een vergunning voor een project of andere handeling die stikstofdepositie veroorzaakt die past binnen de totale hoeveelheid stikstofdepositie die in het programma passend is beoordeeld.

Hoewel de Afdeling aannemelijk acht dat artikel 6 van de Habitatrichtlijn ruimte biedt voor het beoordelingskader dat met het PAS van kracht is geworden, kan zij aan het toepasselijke EU-recht en de rechtspraak van het Hof van Justitie, geen zekerheid ontlenen voor de beantwoording van bovenstaande vragen. Zij ziet daarom aanleiding hierover prejudiciële vragen aan het Hof van Justitie te stellen.

Hieronder komt eerst de uitzondering op de vergunningplicht voor projecten en andere handelingen die de drempel- en grenswaarde niet overschrijden (categorieën a en b) en daarna het gebruik van de passende beoordeling van het PAS voor de vergunningverlening (categorie c) aan de orde.

De uitzondering op de vergunningplicht

9.6. Op grond van artikel 19d, eerste lid, van de Nbw 1998, is het realiseren van een project of het verrichten van een andere handeling onder meer vergunningplichtig wanneer daardoor de kwaliteit van de natuurlijke habitats en de habitats van soorten kan verslechteren. De vergunningplicht voor projecten is een implementatie van artikel 6, derde lid, van de Habitatrichtlijn. De vergunningplicht voor andere handelingen is een implementatie van artikel 6, tweede lid, van de Habitatrichtlijn.

Stikstofdepositie kan verslechterende gevolgen hebben voor stikstofgevoelige habitattypen of leefgebieden waarvoor een Natura 2000-gebied is aangewezen. Deze gevolgen kunnen significant zijn wanneer een project of andere handeling leidt tot een toename van stikstofdepositie op stikstofgevoelige habitattypen of leefgebieden die overbelast zijn. De stikstofgevoelige habitattypen en leefgebieden die in de Natura 2000-gebieden voorkomen die in het PAS zijn opgenomen zijn overbelast. Uit de rechtspraak van de Afdeling volgt dat iedere toename, hoe gering ook (bijv. 0,02 mol N/ha/jr), leidt tot de conclusie dat een project significante gevolgen kan hebben, zodat daarvoor een passende beoordeling moet worden gemaakt.

De uitzondering op de vergunningplicht voor projecten en andere handelingen die stikstofdepositie veroorzaken die de drempelwaarde van 0,05 mol N/ha/jr en de grenswaarde van 1 mol N/ha/jr niet overschrijden heeft derhalve betrekking op:

- projecten die significante gevolgen kunnen hebben voor een Natura 2000-gebied als bedoeld in artikel 6, derde lid, van de Habitatrichtlijn, en

- andere handelingen, die niet als project zijn te duiden of projecten waarvan op voorhand is uitgesloten dat deze significante gevolgen hebben (hierna samen aan te duiden als andere

handelingen). Voor deze andere handelingen is artikel 6, tweede lid, van de Habitatrichtlijn het relevante beschermingsregime.

9.7. De Afdeling constateert dat artikel 6, tweede en derde lid, van de Habitatrichtlijn, de lidstaten vrij laten in de keuze voor de instrumenten ter implementatie van de verplichtingen, zolang die instrumenten geschikt zijn om de richtlijnverplichtingen na te komen. Artikel 6, tweede en derde lid, van de Habitatrichtlijn sluiten niet uit dat projecten en andere handelingen worden uitgezonderd van de vergunningplicht, waardoor deze zonder individuele toestemming zijn toegestaan, ervan uitgaande dat de wettelijke regeling waarin de uitzondering op de vergunningplicht is voorzien waarborgt dat aan de verplichtingen van die bepalingen wordt voldaan. Artikel 6, tweede lid, van de Habitatrichtlijn verplicht de lidstaten ervoor zorg te dragen dat verslechteringen en significante verstoringen worden voorkomen. Artikel 6, derde lid, van de Habitatrichtlijn, verplicht tot het maken van een voorafgaande passende beoordeling van plannen en projecten die significante gevolgen kunnen hebben. Van belang daarbij is dat beide onderdelen van artikel 6 van de Habitatrichtlijn een preventief karakter hebben en dat, zoals uit vaste rechtspraak van het Hof van Justitie volgt, beide bepalingen hetzelfde beschermingsniveau beogen te garanderen (zie bijvoorbeeld Hof van Justitie 4 maart 2010, Frankrijk II, ECLI:EU:C:2010:114).

9.8. Hoewel een situatie als hier aan de orde, waarin aan de wettelijke regeling waarbij voorzien is in de uitzondering op de vergunningplicht een passende beoordeling ten grondslag is gelegd, in de rechtspraak van het Hof van Justitie niet aan de orde is geweest, is er wel rechtspraak over wettelijke regelingen waarin activiteiten zonder meer waren uitgesloten van een beoordeling op grond van artikel 6, tweede en derde lid, van de Habitatrichtlijn. Die rechtspraak, die hierna wordt besproken, biedt naar het oordeel van de Afdeling enkele aanknopingspunten onder welke voorwaarden op basis van een wettelijke regeling een project of andere handeling zonder individuele toestemming kan worden toegestaan.

9.9. In de arresten van het Hof van Justitie van 6 april 2000, Frankrijk I, ECLI:EU:C:2000:192, 10 januari 2006, Duitsland, ECLI:EU:C:2006:3, 4 maart 2010, Frankrijk II, ECLI:EU:C:2010:114, 26 mei 2011, België, ECLI:EU:C:2011:349, en 16 februari 2012, Solvay, ECLI:EU:C:2010:82 is de vraag aan de orde of een lidstaat een wettelijke regeling kan treffen op basis waarvan bepaalde (categorieën van) plannen en projecten zonder meer zijn uitgezonderd van een voorafgaande beoordeling van de gevolgen voor een Natura 2000-gebied. In het Solvay-arrest stelt het Hof dat "artikel 6, lid 3, van de habitatrichtlijn aldus moet worden uitgelegd dat deze bepaling een nationale instantie, ook wanneer het daarbij om een wetgevende instantie gaat, niet in staat stelt toestemming te geven voor een plan of een project zonder de zekerheid te hebben verkregen dat het de natuurlijke kenmerken van het betrokken gebied niet zal aantasten". Deze overweging is in lijn met de daaraan voorafgaande rechtspraak van het Hof van Justitie die in het België-arrest, als volgt werd samengevat:

"41. Bovendien staat de voorwaarde waarvan de beoordeling van de gevolgen van een plan of project voor een bepaald gebied afhangt, welke impliceert dat er bij twijfel over het ontbreken van significante gevolgen een dergelijke beoordeling moet plaatsvinden, niet toe om bepaalde categorieën van projecten daaraan te onttrekken op basis van criteria die niet kunnen waarborgen dat deze projecten geen significante gevolgen kunnen hebben voor beschermde gebieden (zie in die zin arrest van 10 januari 2006, Commissie/Duitsland, C-98/03, Jurispr. blz. I-53, punt 41).

42. De mogelijkheid om bepaalde activiteiten, in overeenstemming met de geldende regels, algemeen uit te sluiten van een verplichte beoordeling van de gevolgen voor het betrokken gebied, waarborgt immers niet dat deze activiteiten de natuurlijke kenmerken van het beschermde gebied niet aantasten (zie in die zin arrest Commissie/Duitsland, reeds aangehaald, punten 43 en 44, en arrest van 4 maart 2010, Commissie/Frankrijk, C-241/08, nog niet gepubliceerd in de Jurisprudentie, punt 31).

43. Artikel 6, lid 3, van de habitatrichtlijn kan een lidstaat dan ook niet machtigen nationale regels uit te vaardigen waardoor ruimtelijkeordeningsplannen op algemene wijze aan de verplichte beoordeling van de gevolgen daarvan voor het gebied zouden worden onttrokken, hetzij op grond van het geringe bedrag van de geplande uitgaven, hetzij vanwege de specifieke in geding zijnde werkterreinen (zie arrest van 6 april 2000, Commissie/Frankrijk, C-256/98, Jurispr. blz. I-2487, punt 39).

44. Evenzo komt een lidstaat, door de programma's en projecten voor bouw- of ontwikkelingswerkzaamheden waarvoor een aanmeldingsregeling geldt, systematisch vrij te stellen van de procedure van beoordeling van de gevolgen voor het gebied, de krachtens artikel 6, lid 3, van de habitatrichtlijn op hem rustende verplichtingen niet na (zie in die zin arrest van 4 maart 2010, Commissie/Frankrijk, reeds aangehaald, punt 62).

45. Bijgevolg volgt uit de rechtspraak van het Hof dat een lidstaat in beginsel overeenkomstig artikel 6, lid 3, van de habitatrichtlijn bepaalde categorieën van plannen of projecten niet systematisch en algemeen aan de verplichte beoordeling van de gevolgen ervan voor Natura 2000-gebieden kan onttrekken, op basis van het werkterrein of door de invoering van een aanmeldingsregeling".

9.10. In het arrest Frankrijk II is de vraag aan de orde of een lidstaat een wettelijke regeling kan treffen waarin wordt verklaard dat bepaalde activiteiten niet verstorend zijn als bedoeld in artikel 6, tweede lid, van de Habitatrichtlijn. Het Hof van Justitie overwoog:

"30 In de eerste plaats zij eraan herinnerd dat volgens de rechtspraak van het Hof lid 2 van artikel 6 van de habitatrichtlijn en lid 3 van datzelfde artikel hetzelfde beschermingsniveau beogen te garanderen […].

31 In de tweede plaats zij opgemerkt dat, wat artikel 6, lid 3, van de habitatrichtlijn betreft, het Hof reeds heeft geoordeeld dat de mogelijkheid om bepaalde activiteiten, in overeenstemming met de geldende regels, algemeen uit te sluiten van een verplichte beoordeling van de gevolgen voor het betrokken gebied, niet in overeenstemming is met die bepaling. Een dergelijke uitsluiting kan immers niet garanderen dat deze activiteiten de natuurlijke kenmerken van het beschermde gebied niet zullen aantasten […].

32 Bijgevolg kan, gelet op het feit dat artikel 6, lid 2, van de habitatrichtlijn en lid 3 van datzelfde artikel hetzelfde beschermingsniveau beogen te garanderen, artikel L. 414-1, lid V, derde alinea, vierde zin, van het milieuwetboek, door in het algemeen te verklaren dat bepaalde activiteiten, zoals de jacht of de visvangst, niet verstorend zijn, enkel in overeenstemming worden geacht met artikel 6, lid 2, van diezelfde richtlijn, indien is gegarandeerd dat deze activiteiten niet leiden tot een verstoring die significante gevolgen kan hebben voor de doeleinden van de richtlijn. […]

34 Derhalve moet worden onderzocht of dergelijke maatregelen of regels daadwerkelijk kunnen verzekeren dat de betrokken activiteiten geen verstoringen veroorzaken die significante effecten kunnen hebben. […]

36 Hieruit volgt dat dit doelstellingendocument niet systematisch en in alle gevallen kan garanderen dat de betrokken activiteiten geen gevolgen hebben die significant zouden kunnen zijn voor deze instandhoudingsdoelstellingen. […]

39 Uit het voorgaande volgt dat de Franse Republiek, door in het algemeen te bepalen dat visvangst, aquacultuur, jacht en andere weidelijke activiteiten die onder de in de wet- en regelgeving geldende voorwaarden en in de in die wet- en regelgeving toegestane gebieden worden bedreven, geen activiteiten zijn die storend zijn of storende gevolgen hebben, de krachtens artikel 6, lid 2, van de habitatrichtlijn op haar rustende verplichtingen niet is nagekomen".

9.11. De rechtspraak van het Hof van Justitie heeft betrekking op wettelijke regelingen die bepaalde categorieën van activiteiten zonder meer uitsluiten van een beoordeling van de gevolgen voor Natura 2000-gebieden. Uit de arresten van het Hof van Justitie leidt de Afdeling af dat in de daar aan de orde zijnde situaties in het geheel geen beoordeling van de gevolgen van de plannen en projecten voor de Natura 2000-gebieden had plaatsgevonden of zou plaatsvinden. Een dergelijke uitzondering acht het Hof van Justitie niet aanvaardbaar.

De uitzondering op de vergunningplicht die in artikel 19kh, zevende en negende lid, van de Nbw 1998 is opgenomen voor projecten en andere handelingen die de drempel- en grenswaarde van 0,05 respectievelijk 1 mol N/ha/jr op een stikstofgevoelig habitat niet overschrijden is naar het oordeel van de Afdeling niet zonder meer te vergelijken met de situaties die aan de orde waren in de arresten van het Hof van Justitie. De gevolgen van de depositie die veroorzaakt wordt door projecten en andere handelingen die de drempel- of grenswaarde niet overschrijden zijn immers in het kader van het PAS als onderdeel van de depositie in 2014 en als onderdeel van de depositieruimte passend beoordeeld.

Een ander relevant verschil met de situaties die bij het Hof van Justitie aan de orde zijn geweest is dat de regeling in de Nbw 1998, het Besluit grenswaarden en de Regeling niet een bepaalde categorie van projecten uitzondert van de vergunningplicht waarvan op voorhand niet duidelijk is welke gevolgen deze hebben voor de natuurwaarden in Natura 2000-gebieden. De uitzondering ziet in dit geval uitsluitend op de gevolgen van stikstofdepositie door projecten en andere handelingen. De (cumulatieve) gevolgen van de totale hoeveelheid stikstofdepositie die in 2014 plaatsvond en die na benutting van de depositie- en ontwikkelingsruimte zal kunnen plaatsvinden op stikstofgevoelige natuurwaarden in Natura 2000-gebieden is voor het PAS, in samenhang met de bron- en herstelmaatregelen, passend beoordeeld. Binnen de depositieruimte die in de passende boordeling is betrokken is ruimte gereserveerd voor projecten en andere handelingen die onder de drempel- en grenswaarde vallen.

De deposities die onder de drempelwaarde vallen worden in het PAS beschouwd als deposities door autonome ontwikkelingen. De toename door deze deposities moet worden opgevangen binnen het deel van de depositieruimte dat voor de autonome ontwikkelingen is gereserveerd (zie de afbeelding in 6.3).

Voor een beoordeling van de uitzondering op de vergunningplicht voor activiteiten die stikstofdepositie veroorzaken die een bepaalde grenswaarde niet overschrijden is van belang dat voor elk Natura 2000-gebied in het PAS is bepaald hoeveel depositieruimte er per hectare van een stikstofgevoelige natuurwaarde beschikbaar is voor activiteiten die de grenswaarde van 1 mol N/ha/jr niet overschrijden. Een initiatiefnemer die een stikstofveroorzakende activiteit wil realiseren dient met het verplicht voorgeschreven rekeninstrument AERIUS Calculator een berekening te maken van de depositie die zijn activiteit zal veroorzaken. Wanneer uit die berekening volgt dat de depositie onder de grenswaarde valt is de activiteit zonder vergunning toegestaan, zij het dat wel een meldingsplicht kan gelden. De melding leidt in AERIUS Register tot de registratie en de afboeking van de toename van de depositie die de activiteit veroorzaakt. Als uit deze registratie volgt dat nog 5% van de totale beschikbaar gestelde depositieruimte voor activiteiten die de grenswaarde niet overschrijden resteert, dan wordt de grenswaarde automatisch verlaagd naar 0,05 mol N/ha/jr. De uitzondering op de vergunningplicht voor activiteiten die de grenswaarde niet overschrijden is derhalve beperkt tot een bepaalde hoeveelheid depositie per hectare van een stikstofgevoelige natuurwaarde in een Natura 2000-gebied. De depositieruimte voor activiteiten die de grenswaarde niet overschrijden maakt onderdeel uit van de totale depositieruimte waarvoor voor het PAS een passende beoordeling is gemaakt.

9.12. De Afdeling ziet in de hiervoor besproken rechtspraak van het Hof van Justitie geen aanknopingspunten dat artikel 6, tweede en derde lid, van de Habitatrichtlijn in de weg staan aan een wettelijke regeling die ertoe strekt dat projecten en andere handelingen die stikstofdepositie veroorzaken die een bepaalde drempel- of grenswaarde niet overschrijden zijn uitgezonderd van de vergunningplicht en derhalve zonder individuele toestemming zijn toegestaan, ervan uitgaande dat de gevolgen van alle projecten en andere handelingen tezamen die gebruik kunnen maken van de wettelijke regeling voor de vaststelling van die wettelijke regeling passend zijn beoordeeld. De passende beoordeling van een programma, zoals het PAS, waarin de cumulatieve gevolgen van een bepaalde hoeveelheid stikstofdepositie die veroorzaakt kan worden door activiteiten die zijn uitgezonderd van de vergunningplicht en door vergunningplichtige activiteiten, is beoordeeld, kan naar het oordeel van de Afdeling aan zo'n wettelijke regeling ten grondslag worden gelegd. Artikel 6, tweede en derde lid, van de Habitatrichtlijn vereisen in dat geval geen individuele toestemming voor een project dat of andere handeling die stikstofdepositie veroorzaakt die past binnen de totale hoeveelheid stikstofdepositie die in het kader van het programma passend is beoordeeld. Dit betekent dat de Afdeling aannemelijk acht dat artikel 19kh, zevende en negende lid, van de Nbw 1998 gelezen in samenhang met artikel 2 van het Besluit grenswaarden, waarin een uitzondering op de vergunningplicht is opgenomen voor projecten of andere handelingen die de drempel- en grenswaarde van 0,05 respectievelijk 1 mol N/ha/jr op een stikstofgevoelig habitat niet overschrijden, niet in strijd is met artikel 6 van de Habitatrichtlijn, ervan uitgaande dat de passende beoordeling die aan het PAS ten grondslag is gelegd voldoet aan de eisen die artikel 6, derde lid, van de Habitatrichtlijn daaraan stelt. Als de passende beoordeling voldoet aan de eisen die artikel 6, derde lid, van de Habitatrichtlijn stelt, dan voldoet deze beoordeling ook aan de eisen die artikel 6, tweede lid, van de Habitatrichtlijn stelt. Artikel 6, tweede en derde lid, van de Habitatrichtlijn beogen immers hetzelfde beschermingsniveau te garanderen.

Prejudiciële vraag 1

9.13. Gelet op het in 9.12 overwogene acht de Afdeling aannemelijk dat artikel 6, tweede en derde lid, van de Habitatrichtlijn, niet in de weg staan aan een wettelijke regeling die ertoe strekt dat projecten en andere handelingen die significante gevolgen kunnen hebben zijn uitgezonderd van de vergunningplicht, waardoor deze zonder individuele toestemming zijn toegestaan, ervan uitgaande dat de gevolgen van alle projecten en andere handelingen tezamen die gebruik kunnen maken van de wettelijke regeling voor de vaststelling van de wettelijke regeling passend zijn beoordeeld. De Afdeling kan echter aan het toepasselijke EU-recht en de rechtspraak van het Hof van Justitie geen zekerheid ontlenen voor haar oordeel. Zij ziet daarom aanleiding aan het Hof van Justitie de volgende vraag voor te leggen:

1. Staat artikel 6, tweede en derde lid, van de Habitatrichtlijn in de weg aan een wettelijke regeling die ertoe strekt dat projecten en andere handelingen die stikstofdepositie veroorzaken die een drempel- of grenswaarde niet overschrijden zijn uitgezonderd van de vergunningplicht en daardoor zonder individuele toestemming zijn toegestaan, ervan uitgaande dat de gevolgen van alle projecten en andere handelingen tezamen die gebruik kunnen maken van de wettelijke regeling voor de vaststelling van die wettelijke regeling passend zijn beoordeeld?

Het gebruik van een passende beoordeling van een programma bij de verlening van de vergunning voor een individueel project

9.14. Projecten en andere handelingen die de grenswaarde van 1 mol N/ha/jr overschrijden zijn vergunningplichtig (categorie c). Voor dergelijke projecten en andere handelingen is derhalve een individuele toestemming vereist.

Uit het beoordelingskader dat op grond van het PAS en de bijbehorende regelgeving voor vergunningplichtige activiteiten geldt volgt dat de initiatiefnemer bij de vergunningaanvraag geen passende beoordeling voor het aspect stikstof hoeft te overleggen. Bij de vergunningaanvraag wordt een berekening van de stikstofdepositie gevoegd die met AERIUS Calculator is gemaakt. Uit die berekening volgt of de aangevraagde activiteit past binnen de hoeveelheid stikstofdepositie die voor het PAS passend is beoordeeld. Het bevoegd gezag verleent de vergunning op basis van die berekening en onder verwijzing naar de passende beoordeling die voor het PAS is gemaakt.

Het PAS-beoordelingskader strekt er derhalve toe dat in de gevallen waarin een individuele toestemming is vereist, geen individuele passende beoordeling is vereist. Dit beoordelingskader heeft betrekking op vergunningaanvragen voor projecten die significante gevolgen kunnen hebben voor een Natura 2000-gebied als bedoeld in artikel 6, derde lid, van de Habitatrichtlijn, en op vergunningaanvragen voor andere handelingen die onder het beschermingsregime van artikel 6, tweede lid, van de Habitatrichtlijn vallen.

9.15. Artikel 6, tweede lid, van de Habitatrichtlijn verplicht de lidstaten ervoor zorg te dragen dat verslechteringen en significante verstoringen worden voorkomen. Artikel 6, derde lid, van de Habitatrichtlijn, verplicht tot het maken van een voorafgaande passende beoordeling van plannen en projecten die significante gevolgen kunnen hebben. Van belang daarbij is dat beide onderdelen van artikel 6 van de Habitatrichtlijn een preventief karakter hebben en dat, zoals uit vaste rechtspraak van het Hof van Justitie volgt, beide bepalingen hetzelfde beschermingsniveau beogen te garanderen (zie bijvoorbeeld Hof van Justitie 4 maart 2010, Frankrijk II, ECLI:EU:C:2010:114).

9.16. Ten aanzien van het betoog dat de passende beoordeling van een programma zoals het PAS niet ten grondslag kan worden gelegd aan de verlening van een vergunning voor een individueel project of andere handeling, omdat dat project op zich zelf passend beoordeeld moet worden, wordt het volgende overwogen. Voor het PAS is voor elk Natura 2000-gebied een passende beoordeling gemaakt waarin is onderzocht of de stikstofdepositie die in 2014 plaatsvond en de depositie die gedurende de PAS-periode van zes jaar kan gaan plaatsvinden na benutting van de depositie- en ontwikkelingsruimte de natuurlijke kenmerken van het betrokken Natura 2000-gebied niet zullen aantasten. Daarbij is rekening gehouden met de autonome daling van de stikstofdepositie en de daling van de stikstofdepositie door de PAS-bronmaatregelen, en met de herstelmaatregelen. De depositie- en ontwikkelingsruimte die per hectare van een stikstofgevoelig habitattype in een Natura 2000-gebied beschikbaar is gesteld is opgenomen in AERIUS.

De initiatiefnemer dient bij de vergunningaanvraag een berekening van de stikstofdepositie te voegen die met AERIUS Calculator is gemaakt. AERIUS Calculator berekent de stikstofdepositie van de aangevraagde activiteit per hectare van een stikstofgevoelig habitattype en geeft aan of deze activiteit past binnen de hoeveelheid stikstofdepositie die voor het PAS passend is beoordeeld. Vergunningaanvragen die betrekking hebben op een situatie die in 2014 feitelijk plaatsvond kunnen onder verwijzing naar de passende beoordeling voor het PAS worden verleend omdat de depositie van deze activiteiten is opgenomen in de achtergronddepositie (de uitgangssituatie) van het PAS.

Leidt de aangevraagde activiteit tot een toename van stikstofdepositie ten opzichte van 2014 dan geeft AERIUS Calculator aan of die toename past binnen de ontwikkelingsruimte die in het PAS per hectare van een stikstofgevoelig habitattype is vastgesteld. Wanneer de activiteit binnen de beschikbare ontwikkelingsruimte past en het bevoegd gezag de ontwikkelingsruimte aan de vergunning toedeelt, dan kan het bevoegd gezag de vergunning verlenen onder verwijzing naar de passende beoordeling voor het PAS. De ontwikkelingsruimte is in de passende beoordeling voor het PAS betrokken als onderdeel van de depositieruimte. De ontwikkelingsruimte die aan de vergunning wordt toegedeeld wordt geregistreerd en afgeboekt in AERIUS.

In de beide hiervoor beschreven situaties zijn de gevolgen van stikstofdepositie derhalve voorafgaand aan de verlening van een vergunning voor een stikstofveroorzakende activiteit beoordeeld, als de stikstofdepositie van de aangevraagde activiteit past binnen de hoeveelheid stikstofdepositie die in de passende beoordeling voor het PAS is betrokken.

9.17. De Afdeling acht aannemelijk dat artikel 6, tweede en derde lid, van de Habitatrichtlijn niet eraan in de weg staan dat de passende beoordeling voor het PAS, waarin de cumulatieve gevolgen van een bepaalde hoeveelheid stikstofdepositie zijn beoordeeld, zonder dat daaraan specifieke projecten of andere handelingen zijn verbonden, ten grondslag kan worden gelegd aan de verlening van een vergunning voor een project dat of andere handeling die stikstofdepositie veroorzaakt die past binnen de in het PAS beoordeelde totale hoeveelheid stikstofdepositie, ervan uitgaande dat de passende beoordeling die voor het PAS is gemaakt voldoet aan de vereisten die de Habitatrichtlijn daaraan stelt. Als de passende beoordeling van het PAS voldoet aan de eisen die de Habitatrichtlijn stelt, is verzekerd dat het project waarvoor een vergunning wordt verleend waaraan die passende beoordeling ten grondslag is gelegd, de natuurlijke kenmerken van het Natura 2000-gebied niet zal aantasten. Als de

passende beoordeling voldoet aan de eisen die artikel 6, derde lid, van de Habitatrichtlijn stelt, dan voldoet deze beoordeling ook aan de eisen die artikel 6, tweede lid, van de Habitatrichtlijn stelt. Artikel 6, tweede en derde lid, van de Habitatrichtlijn beogen immers hetzelfde beschermingsniveau te garanderen.

9.18. Het betoog van de Werkgroep over het Wezerarrest strekt ertoe dat uit hetgeen het Hof van Justitie over artikel 4 van de Kaderrichtlijn Water heeft geoordeeld zou moeten worden afgeleid dat ook op grond van artikel 6, derde lid, van de Habitatrichtlijn ieder specifiek project afzonderlijk passend moet worden beoordeeld voordat hiervoor toestemming wordt verleend. Ter zitting heeft de Werkgroep desgevraagd toegelicht dat het systeem van de Kaderrichtlijn Water en het systeem van de Habitatrichtlijn in zoverre vergelijkbaar is. De Afdeling volgt dit betoog niet. In het Wezerarrest heeft het Hof van Justitie uitleg gegeven aan de verplichtingen uit artikel 4 van de Kaderrichtlijn Water. Die uitleg strekte ertoe dat artikel 4 van de Kaderrichtlijn Water niet slechts beginselverplichtingen kent, maar ook betrekking heeft op specifieke projecten en dat de lidstaat zijn goedkeuring voor een project moet weigeren wanneer dat project de toestand van het desbetreffende waterlichaam kan verslechteren of het bereiken van een goede toestand van oppervlaktewaterlichamen in gevaar kan brengen, tenzij voor het project een afwijking op grond van artikel 4, zevende lid, van de Kaderrichtlijn Water geldt. De Kaderrichtlijn Water voorziet in een verplicht stroomgebiedsbeheersplan en het daarin vervatte maatregelenprogramma dat ten uitvoer moet worden gelegd. Volgens het Hof van Justitie kunnen projecten niet los worden gezien van genoemd beheersplan. Artikel 6 van de Habitatrichtlijn kent een dergelijk systeem van tenuitvoerlegging van een verplicht voorgeschreven plan niet. Artikel 6 van de Habitatrichtlijn verplicht de lidstaten tot het treffen van de nodige instandhoudings- en passende maatregelen voor speciale beschermingszones en brengt daarnaast en los daarvan met zich dat plannen of projecten die afzonderlijk of in combinatie significante gevolgen kunnen hebben voor Natura 2000-gebieden passend beoordeeld moeten worden en in beginsel alleen mogen worden toegestaan als de zekerheid is verkregen dat de natuurlijke kenmerken niet zullen worden aangetast. Artikel 4 van de Kaderrichtlijn Water en artikel 6 van de Habitatrichtlijn zijn in zoverre dan ook niet vergelijkbaar. De uitleg van het Hof van Justitie van de verplichtingen uit artikel 4 van de Kaderrichtlijn Water is derhalve naar het oordeel van de Afdeling niet op voorhand doorslaggevend voor de uitleg van de verplichtingen die voortvloeien uit artikel 6, derde lid, van de Habitatrichtlijn.

Het betoog faalt.

Prejudiciële vraag 2

9.19. Het in 9.17 overwogene strekt ertoe dat de Afdeling aannemelijk acht dat artikel 6, tweede en derde lid, van de Habitatrichtlijn er niet aan in de weg staan dat de passende beoordeling voor een programma waarin een bepaalde totale hoeveelheid stikstofdepositie is beoordeeld ten grondslag wordt gelegd aan de verlening van een vergunning voor een individueel project of andere handeling, die stikstofdepositie veroorzaakt die binnen de in het kader van het programma beoordeelde depositieruimte past. De Afdeling kan echter aan het toepasselijke EU-recht en de rechtspraak van het Hof van Justitie geen zekerheid ontlenen voor haar oordeel. Zij ziet daarom aanleiding aan het Hof van Justitie de volgende vraag voor te leggen:

2. Staat artikel 6, tweede en derde lid, van de Habitatrichtlijn eraan in de weg dat een passende beoordeling voor een programma waarin een bepaalde totale hoeveelheid stikstofdepositie is beoordeeld ten grondslag wordt gelegd aan de verlening van een vergunning (individuele toestemming) voor een project of andere handeling, die stikstofdepositie veroorzaakt die binnen de in het kader van het programma beoordeelde depositieruimte past?

De passende beoordeling in het licht van artikel 6 van de Habitatrichtlijn

Beroepsgronden en standpunt college

10. De Werkgroep stelt dat de passende beoordeling die aan het PAS ten grondslag ligt niet voldoet aan de eisen die de Habitatrichtlijn stelt. Bij de beoordeling van de gevolgen vanwege de depositie die in 2014 plaatsvond en die na benutting van de depositieruimte kan gaan plaatsvinden, zijn volgens de Werkgroep ten onrechte instandhoudingsmaatregelen en passende maatregelen betrokken. Deze maatregelen, die op grond van artikel 6, eerste en tweede lid, van de Habitatrichtlijn, zonder meer moeten worden getroffen, mogen niet ingezet worden om bestaande, maar nog niet vergunde, en om nieuwe stikstofveroorzakende activiteiten toe te staan. Daarnaast stelt de Werkgroep dat uit de arresten van het Hof van Justitie van 15 mei 2014, Briels, ECLI:EU:C:2014:330 en van 21 juli 2016, Orleans, ECLI:EU:C:2016:583 volgt dat maatregelen die betrokken worden in een passende beoordeling niet het karakter van een instandhoudings- of passende maatregel kunnen hebben.

De herstelmaatregelen die in de passende beoordeling zijn betrokken zijn volgens de Werkgroep maatregelen die direct verband houden met de benutting van de depositieruimte voor projecten die stikstofdepositie veroorzaken. De Werkgroep leidt uit de gebiedsanalyses voor de Natura 2000-gebieden Groote Peel en Deurnsche Peel & Mariapeel af dat indien de herstelmaatregelen niet worden getroffen de hoge stikstofbelasting zal leiden tot een aantasting van de habitattypen droge heide, actief hoogveen en herstellend hoogveen. De Werkgroep vindt de herstelmaatregelen die in het kader van het PAS worden genomen vergelijkbaar met de natuurmaatregelen die aan de orde waren in de arresten Briels en Orleans. In de eerste plaats omdat in die beide zaken eveneens vaststond dat de projecten die daar waren voorzien zonder het treffen van de natuurmaatregelen zouden leiden tot een aantasting van arealen van habitattypen. In de tweede plaats omdat het maatregelen betreft waarvan de positieve gevolgen zich nog moeten gaan manifesteren. De positieve effecten van de herstelmaatregelen die in de passende beoordeling zijn betrokken, zijn volgens de Werkgroep onzeker en zullen pas na enkele jaren zichtbaar worden. De Werkgroep meent dat de maatregelen geen mitigerend maar een compenserend karakter hebben. Het zijn maatregelen die relevant zijn bij de toepassing van artikel 6, vierde lid, van de Habitatrichtlijn.

Ook de effecten van de bronmaatregelen zijn naar de mening van de Werkgroep onzeker, terwijl de ontwikkelingsruimte al toegedeeld kan worden voordat de positieve gevolgen van de maatregelen zich hebben voorgedaan. Zij stelt voorts dat een programma met een looptijd van zes jaar niet ten grondslag mag worden gelegd aan vergunningen voor onbepaalde tijd, omdat de in het PAS voorziene herstelmaatregelen slechts zijn verzekerd voor zes jaar. Verder stelt de Werkgroep dat de maatregelen evenmin als feitelijke ontwikkeling kunnen worden betrokken in de passende beoordeling, zoals aan de orde was in de uitspraken van de Afdeling van 27 januari 2016, ECLI:NL:RVS:2016:170. Anders dan in die zaak aan de orde

was zijn de maatregelen die in het kader van de uitvoering van het PAS worden getroffen nadrukkelijk bedoeld om vergunningverlening mogelijk te maken.

10.1. Het college wijst er allereerst op dat het PAS en de zaken Briels en Orleans wezenlijk van elkaar verschillen omdat in de zaken Briels en Orleans vaststond dat de plannen of projecten zouden leiden tot een aantasting of verlies van arealen van habitattypen en in de passende beoordeling de ontwikkeling van nieuwe natuur op een andere locatie binnen het Natura 2000-gebied was betrokken. Een dergelijke maatregel is volgens het Hof van Justitie geen beschermingsmaatregel die betrokken kan worden in een passende beoordeling, aldus het college.

Het PAS zal volgens het college nergens leiden tot een onmiddellijke fysieke aantasting van een type natuurlijke habitat, die wordt gecompenseerd door ontwikkeling van nieuwe habitats elders in het gebied. Het PAS stelt volgens het college zeker dat op geen van de onderscheiden locaties van de stikstofgevoelige habitats in de betrokken Natura 2000-gebieden een toename van stikstofdepositie plaatsvindt, ook niet bij toedeling van de beschikbaar gestelde depositie- en ontwikkelingsruimte aan nieuwe projecten die stikstofdepositie veroorzaken. Per saldo leiden de in het PAS voorziene maatregelen voor alle gebieden en habitats tot een afname van stikstofdepositie. In de passende beoordeling is speciaal aandacht besteed aan de mogelijkheid van een eventuele tijdelijke toename van stikstofdepositie gedurende de eerste helft van het tijdvak van het programma, als de uitgifte van ontwikkelingsruimte en het gebruik daarvan sneller zou verlopen dan de daling van de stikstofdepositie. Daarbij is, zo stelt het college, onderbouwd dat deze toename altijd tijdelijk van aard is en nog gedurende de programmaperiode op dezelfde locatie zal worden gevolgd door een afname die verzekert dat per saldo de stikstofdepositie afneemt en dat de aantasting van de natuurlijke kenmerken ook in deze situatie is uit te sluiten. Uit de gebiedsanalyses blijkt voorts dat het samenstel van maatregelen van het programma in elk van de betrokken Natura 2000-gebieden ook verslechtering van de kwaliteit van de habitats in de zin van artikel 6, tweede lid, van de Habitatrichtlijn uitsluit en bijdraagt aan de realisatie van de instandhoudingsdoelstellingen van de gebieden, overeenkomstig artikel 6, eerste lid, van de Habitatrichtlijn, aldus het college.

De PAS-bronmaatregelen en de gebiedsgerichte herstelmaatregelen uit het programma zijn volgens het college deels aan te merken als instandhoudingsmaatregelen en passende maatregelen in de zin van artikel 6, eerste en tweede lid, van de Habitatrichtlijn en, in relatie tot nieuwe projecten, deels als maatregelen waardoor ontwikkelingsruimte kan worden toegedeeld. Met de maatregelen wordt op het niveau van het programma gewaarborgd dat overal een langjarige afname van stikstofdepositie wordt gerealiseerd, de kwaliteit van de stikstof gevoelige habitattypen en leefgebieden behouden blijft en de instandhoudingsdoelstellingen voor de betrokken habitattypen en soorten met een stikstofgevoelig leefgebied niet in gevaar komen en zonder onevenredige vertraging worden gehaald.

De uitvoering van de maatregelen is volgens het college voldoende verzekerd. Artikel 19kj van de Nbw 1998 bepaalt dat de bestuursorganen die het aangaat zorg dragen voor een tijdige uitvoering van de PAS-maatregelen.

Het college stelt voorts onder verwijzing naar de uitspraak van de Afdeling van 27 januari 2016, ECLI:NL:RVS:2016:170, dat instandhoudingsmaatregelen als feitelijke ontwikkeling in

een passende beoordeling mogen worden betrokken.

De passende beoordeling

10.2. In de Nbw 1998 is gekozen voor een programmatische aanpak van de stikstofproblematiek in de Natura 2000-gebieden. Het PAS is gericht op het verminderen van stikstofdepositie in de Natura 2000-gebieden die in het programma zijn opgenomen, het op termijn realiseren van de instandhoudingsdoelstellingen voor de stikstofgevoelige natuurwaarden waarvoor deze gebieden zijn aangewezen, en op het scheppen van ruimte voor nieuwe economische activiteiten die stikstofdepositie veroorzaken in deze gebieden. Het programma is gericht op het nakomen van de verplichtingen die voortvloeien uit artikel 6, eerste en tweede lid, van de Habitatrichtlijn, en voor zover het is gericht op het scheppen van ruimte voor economische ontwikkelingen, dient het programma te voldoen aan artikel 6, derde lid, van de Habitatrichtlijn. Dat betekent in dit geval dat de totale depositie in 2014 en de depositieruimte die in de programmaperiode van zes jaar beschikbaar wordt gesteld passend beoordeeld moet worden.

10.3. In de passende beoordeling die aan het PAS ten grondslag is gelegd is per Natura 2000-gebied dat in het programma is opgenomen onderzocht of de depositie die in 2014 plaatsvond en de depositie die gedurende de PAS-periode van zes jaar kan gaan plaatsvinden met benutting van de depositieruimte tot een aantasting van de natuurlijke kenmerken van het betreffende Natura 2000-gebieden zullen leiden. Bij die beoordeling is rekening gehouden met de autonome daling van de stikstofdepositie door bestaand en toekomstig beleid. Verder is rekening gehouden met de verwachte gevolgen van de daling van stikstofdepositie door de PAS-bronmaatregelen. Het gaat hierbij om stalmaatregelen, maatregelen voor emissiearme bemesting en voer- en managementmaatregelen. In de passende beoordeling is tot slot rekening gehouden met het verwachte positieve effect van getroffen en nog te treffen herstelmaatregelen. De herstelmaatregelen zijn voor elk Natura 2000-gebied in de gebiedsanalyse uitgewerkt. De maatregelen betreffen in veel gevallen hydrologische maatregelen en aanvullende beheermaatregelen. Verder is rekening gehouden met het feit dat een deel van de ontwikkelingsruimte direct na de vaststelling van het PAS kan worden toegedeeld.

In de passende beoordeling wordt geconcludeerd dat de stikstofdepositie die veroorzaakt kan worden door de projecten en andere handelingen die op grond van het PAS-beoordelingskader zijn toegestaan, gelet op de autonome daling van de stikstofdepositie en de te treffen bron- en herstelmaatregelen, niet zullen leiden tot verslechtering van de kwaliteit van habitats, tot aantasting van de natuurlijke kenmerken van de gebieden, of tot onevenredige vertraging in het halen van de instandhoudingsdoelstellingen van de betrokken Natura 2000-gebieden.

Aanleiding prejudiciële vragen

10.4. De Afdeling acht een programmatische aanpak zoals het PAS, dat een integrale aanpak beoogt van de stikstofproblematiek in de Natura 2000-gebieden, niet op voorhand een ongeschikt instrument om uitvoering te geven aan de verplichtingen die uit artikel 6 van de Habitatrichtlijn voortvloeien. Aan een programma zoals het PAS, dat enerzijds gericht is op het behoud en waar mogelijk herstel van natuurwaarden en anderzijds op het scheppen van depositieruimte voor bestaande en toekomstige activiteiten die in samenhang worden

beoordeeld, is inherent dat de gevolgen van het benutten van de depositieruimte worden beoordeeld in samenhang met alle maatregelen en autonome ontwikkelingen die zich tijdens de programmaperiode in het Natura 2000-gebied zullen voordoen. Indien, zoals de Werkgroep betoogt, een dergelijke integrale beoordeling van de maatregelen, autonome ontwikkelingen en depositieruimte voor economische activiteiten binnen de systematiek van artikel 6 van de Habitatrichtlijn niet is toegestaan, is een programmatische aanpak gericht op het behoud en waar mogelijk herstel van natuurwaarden en op het scheppen van depositieruimte voor economische activiteiten niet mogelijk.

10.5. In de rechtspraak van het Hof van Justitie is nog niet eerder de vraag aan de orde geweest of een programma dat zowel gericht is op het behoud en waar mogelijk het herstel van natuurwaarden als op het scheppen van depositieruimte voor economische activiteiten, een geschikt instrument is om de verplichtingen uit artikel 6, eerste, tweede en derde lid, van de Habitatrichtlijn na te komen. Evenmin is in die rechtspraak de vraag aan de orde geweest welke maatregelen, in het geval een dergelijk programma kan worden gebruikt, mogen worden betrokken in de passende beoordeling van een programma met een dergelijke dubbeldoelstelling. Wel bestaat er rechtspraak van het Hof van Justitie, namelijk de eerder genoemde arresten Briels en Orleans, over de vraag hoe beoordeeld moet worden of een individueel plan of project de natuurlijke kenmerken van een Natura 2000-gebied niet zal aantasten en over het betrekken van maatregelen in een passende beoordeling voor individuele projecten en plannen. Ook het recente arrest van 26 april 2017, Kolencentrale Moorburg, ECLI:EU:C:2017:301, gaat over de vraag welke maatregelen in een passende beoordeling mogen worden betrokken.

10.6. In de arresten Briels en Orleans is de vraag aan de orde of in de passende beoordeling van een plan of project dat zal leiden tot de aantasting van bestaande arealen van habitattypen rekening mag worden gehouden met de maatregel dat op een andere locatie in hetzelfde Natura 2000-gebied nieuwe natuur zal worden ontwikkeld. De aanleg van nieuwe natuur is volgens het Hof van Justitie geen beschermingsmaatregel die in een passende beoordeling kan worden betrokken, omdat de maatregel niet gericht is op het voorkomen van de rechtstreekse gevolgen van de projecten voor de bestaande arealen van habitattypen. Het Hof van Justitie wijst er in dat verband op dat bij de beoordeling van de gevolgen van een plan of project niet mag worden uitgegaan van de aanname dat toekomstige voordelen van de aanleg van nieuwe natuur de significante effecten op het gebied zullen mitigeren, terwijl het resultaat van de ontwikkeling van de nieuwe natuur onzeker is.

10.7. De maatregelen die in de passende beoordeling van het PAS zijn betrokken zijn gericht op het voorkomen van verslechtering van de bestaande arealen van habitattypen en leefgebieden van soorten door stikstofdepositie. Anders dan in de zaken Briels en Orleans is in deze zaak dan ook de vraag aan de orde of, en zo ja onder welke voorwaarden in een passende beoordeling rekening mag worden gehouden met instandhoudingsmaatregelen, passende maatregelen en maatregelen die specifiek voor het programma worden genomen en met autonome ontwikkelingen die de negatieve gevolgen van stikstofdepositie voor bestaande arealen van habitattypen en leefgebieden van soorten zodanig voorkomen of verminderen dat verslechtering van de bestaande arealen van de habitattypen en de leefgebieden wordt voorkomen en herstel, waar nodig, mogelijk blijft. Verder is de vraag aan de orde of voor deze maatregelen geldt dat ze uitsluitend in een passende beoordeling mogen worden betrokken als ze op dat moment genomen zijn en resultaat hebben gehad.

Hoewel de Afdeling aannemelijk acht dat getroffen en nog te treffen maatregelen die gericht zijn op het voorkomen van verslechtering van bestaande arealen van habitattypen en leefgebieden en autonome ontwikkelingen in een passende beoordeling kunnen worden betrokken, kan zij aan het toepasselijke EU-recht en de rechtspraak van het Hof van Justitie, waarin een geval als dit nog niet aan de orde is geweest, geen zekerheid ontlenen voor de beantwoording van bovenstaande vragen. Zij ziet daarom aanleiding hierover prejudiciële vragen aan het Hof van Justitie te stellen.

10.8. In de hiernavolgende overwegingen wordt eerst de relevante rechtspraak van het Hof van Justitie en van de Afdeling uiteengezet. Daarna worden het PAS en de maatregelen die in de passende beoordeling daarvoor zijn betrokken geduid in het licht van die jurisprudentie. Vervolgens worden de beroepsgronden besproken. Die bespreking wordt afgesloten met de prejudiciële vragen.

Rechtspraak Hof van Justitie

10.9. In de arresten Briels en Orleans is de vraag aan de orde hoe beoordeeld moet worden dat een individueel plan of project de natuurlijke kenmerken van een Natura 2000-gebied niet zal aantasten en welke maatregelen betrokken mogen worden in een passende beoordeling voor een project of plan.

10.10. In het arrest Briels van het Hof van Justitie is de vraag aan de orde of de verbreding van Rijksweg A2 zal leiden tot een aantasting van de natuurlijke kenmerken van een Natura 2000-gebied. De verbreding van de weg zal leiden tot kwaliteitsverlies van een areaal van het habitattype blauwgrasland. In de passende beoordeling is geconcludeerd dat dat niet leidt tot een aantasting van de natuurlijke kenmerken van het gebied omdat in het kader van de realisering van het project tevens is voorzien in de aanleg van een areaal van dit habitattype op een andere locatie in hetzelfde Natura 2000-gebied. Het Hof van Justitie oordeelde:

"21. Het Hof heeft aldus geoordeeld dat een ingreep geen aantasting van de natuurlijke kenmerken van een gebied, te weten een natuurlijke habitat, in de zin van artikel 6, lid 3, tweede volzin, van de habitatrichtlijn meebrengt, indien dat gebied wordt bewaard in een gunstige staat van instandhouding, hetgeen neerkomt op het duurzame behoud van de bepalende kenmerken van het betrokken gebied die verband houden met de aanwezigheid van een type natuurlijke habitat waarvan de instandhoudingsdoelstelling rechtvaardigde dat dit gebied in de lijst van GCB's in de zin van die richtlijn werd opgenomen (arrest Sweetman e.a., EU:C:2013:220, punt 39).

22 In het hoofdgeding staat vast dat het betrokken Natura 2000-gebied door de Commissie als GCB en door het Koninkrijk der Nederlanden als speciale beschermingszone is aangewezen, met name wegens de aanwezigheid in dit gebied van het natuurlijke habitattype „blauwgraslanden", waarvan de instandhoudingsdoelstelling ziet op de uitbreiding van de oppervlakte van deze habitat en de verhoging van de kwaliteit ervan.

23 Bovendien blijkt uit de aan het Hof overgelegde stukken dat het tracéproject Rijksweg A2 significante negatieve gevolgen voor de habitattypen en beschermde soorten in dit gebied zal hebben, inzonderheid voor het bestaande areaal en voor de kwaliteit van het beschermde natuurlijke habitattype „blauwgraslanden", wegens de uitdroging en de verzuring van de bodem door stikstofdepositie.

24 Een dergelijk project dreigt het duurzame behoud van de wezenlijke kenmerken van het betrokken Natura 2000-gebied in gevaar te brengen en kan bijgevolg, zoals de advocaat-generaal in punt 41 van haar conclusie heeft opgemerkt, de natuurlijke kenmerken van het gebied aantasten in de zin van artikel 6, lid 3 van de habitatrichtlijn.

25 Anders dan de Nederlandse regering stelt, hierin ondersteund door de regering van het Verenigd Koninkrijk, doen de in het tracéproject Rijksweg A2 voorgestelde beschermingsmaatregelen niet af aan die vaststelling.

26 Ten eerste moet immers in herinnering worden geroepen dat in het bij artikel 6, lid 3, tweede volzin, van de habitatrichtlijn vastgestelde toestemmingscriterium het voorzorgsbeginsel ligt besloten, aangezien de bevoegde nationale instantie de toestemming voor het voorgelegde plan of project moet weigeren wanneer zij nog niet de zekerheid heeft verkregen dat het plan of project geen effecten heeft die de natuurlijke kenmerken van dat gebied zullen aantasten. Zo kan op efficiënte wijze worden voorkomen dat de natuurlijke kenmerken van de beschermde gebieden worden aangetast als gevolg van plannen of projecten. Met een minder streng toestemmingscriterium zou de verwezenlijking van de doelstelling van bescherming van de gebieden waartoe deze bepaling strekt, niet even goed kunnen worden gegarandeerd (arresten Waddenvereniging en Vogelbeschermingsvereniging, C-127/02, EU:C:2004:482, punten 57 en 58, en Sweetman e.a., EU:C:2013:220, punt 41).

27 Een overeenkomstig artikel 6, lid 3, van de habitatrichtlijn uitgevoerde beoordeling mag dus geen leemten vertonen en moet volledige, precieze en definitieve constateringen en conclusies bevatten die elke redelijke wetenschappelijke twijfel over de gevolgen van de geplande werkzaamheden voor het betrokken beschermde gebied kunnen wegnemen (zie in die zin arrest Sweetman e.a., EU:C:2013:220, punt 44 en aldaar aangehaalde rechtspraak).

28 Bijgevolg verlangt het voorzorgsbeginsel van de bevoegde nationale instantie dat zij bij de toepassing van artikel 6, lid 3, van de habitatrichtlijn de gevolgen van het project voor het betrokken Natura 2000-gebied beoordeelt in het perspectief van de instandhoudingsdoelstellingen van dit gebied, rekening houdend met de in dit project vastgestelde beschermingsmaatregelen waarmee wordt beoogd de eventuele schadelijke gevolgen die rechtstreeks uit dit project voortvloeien, te voorkomen of te verminderen, teneinde ervoor te zorgen dat het betrokken project de natuurlijke kenmerken van dat gebied niet aantast.

29 De beschermingsmaatregelen die in een project worden opgenomen om de schadelijke gevolgen van dit project voor een Natura 2000-gebied te compenseren, kunnen daarentegen bij de door artikel 6, lid 3, opgelegde beoordeling van de gevolgen van dit project niet in aanmerking worden genomen.

30 Dat zou echter het geval zijn voor de maatregelen die in het hoofdgeding aan de orde zijn die, in een situatie waarin de bevoegde nationale instantie daadwerkelijk heeft vastgesteld dat het tracéproject Rijksweg A2 significante negatieve - eventueel zelfs blijvende - gevolgen voor het beschermde habitattype van het betrokken Natura 2000-gebied kan hebben, voorzien in de toekomstige ontwikkeling van een nieuw areaal van dezelfde of een grotere omvang van dit habitattype in een ander deel van dit gebied, dat niet rechtstreeks door dit project zou worden aangetast.

31 Geconstateerd moet immers worden dat deze maatregelen er niet toe strekken om de significante negatieve gevolgen die voor dit habitattype rechtstreeks uit het tracéproject Rijksweg A2 voortvloeien, te voorkomen of te verminderen, maar beogen deze gevolgen nadien te compenseren. In die omstandigheden kunnen die maatregelen niet garanderen dat het project de natuurlijke kenmerken van dit gebied niet zal aantasten in de zin van artikel 6, lid 3, van de habitatrichtlijn.

32 Bovendien dient erop te worden gewezen dat de eventuele positieve gevolgen van het achteraf tot ontwikkeling brengen van een nieuwe habitat waarmee het verlies aan oppervlakte en kwaliteit van ditzelfde type habitat in een beschermd gebied dient te worden gecompenseerd - ook al zou het om een groter areaal van een hogere kwaliteit gaan - in de regel onzeker zijn, en dat deze gevolgen hoe dan ook slechts binnen enkele jaren zichtbaar zullen worden, zoals uit punt 87 van de verwijzingsbeslissing blijkt. Bijgevolg kan daarmee in het kader van de bij die bepaling vastgestelde procedure geen rekening worden gehouden".

10.11. Het arrest Orleans gaat over de vraag of in een passende beoordeling van een plan voor de uitbreiding van de haven van Antwerpen die zal leiden tot het verlies van enkele arealen van habitattypen, rekening mag worden gehouden met de maatregel dat de ontwikkeling van de haven (en daarmee de aantasting van de habitattypen) pas mogelijk is na de duurzame inrichting van habitats en leefgebieden van soorten in natuurkerngebieden. Het Hof overwoog:

"33 Artikel 6 van de richtlijn deelt maatregelen dus in drie categorieën in, namelijk instandhoudingsmaatregelen, preventieve maatregelen en compenserende maatregelen, als respectievelijk bedoeld in de leden 1, 2 en 4 van dit artikel.

34 In de hoofdgedingen zijn het Gemeentelijk Havenbedrijf Antwerpen en de Belgische regering van mening dat de stedenbouwkundige voorschriften van het GRUP instandhoudingsmaatregelen in de zin van artikel 6, lid 1, van de habitatrichtlijn zijn. Volgens deze regering is het ook mogelijk dat dergelijke maatregelen onder artikel 6, lid 2, vallen.

[…]

36 Het Hof heeft in dit verband reeds geoordeeld dat de lidstaten volgens de richtlijn passende beschermingsmaatregelen dienen te nemen ter instandhouding van de ecologische kenmerken van gebieden met typen natuurlijke habitats (arrest van 11 april 2013, Sweetman e.a., C-258/11, EU:C:2013:220, punt 38 en aldaar aangehaalde rechtspraak).

37 In de onderhavige zaken heeft de verwijzende rechter geconstateerd dat het GRUP met name zou leiden tot het verlies van 20 hectare schorren en slikken in het betrokken Natura 2000-gebied.

38 Vastgesteld moet dus worden dat uit de door deze rechter gedane constateringen blijkt dat de in de hoofdgedingen aan de orde zijnde maatregelen met name erin voorzien dat een deel van het gebied verloren gaat. Hieruit volgt dat dergelijke maatregelen geen maatregelen ter instandhouding van dat gebied kunnen zijn.

39 Voorts heeft het Hof ten aanzien van preventieve maatregelen reeds geoordeeld dat het

bepaalde in artikel 6, lid 2, van de habitatrichtlijn het mogelijk maakt te voldoen aan het hoofddoel, namelijk het behoud en de bescherming van de kwaliteit van het milieu, met inbegrip van de instandhouding van de natuurlijke habitats en de wilde flora en fauna, en een algemene beschermingsverplichting oplegt die erin bestaat verslechteringen en verstoringen te voorkomen die gelet op de doelstellingen van deze richtlijn significante gevolgen zouden kunnen hebben (arrest van 14 januari 2010, Stadt Papenburg, C-226/08, EU:C:2010:10, punt 49 en aldaar aangehaalde rechtspraak).

40 Een preventieve maatregel is dus slechts in overeenstemming met artikel 6, lid 2, van deze richtlijn indien is gegarandeerd dat hij niet leidt tot een verstoring die significante gevolgen kan hebben voor de doelstellingen van de richtlijn, met name voor de daarmee nagestreefde instandhoudingsdoelstellingen (arrest van 14 januari 2016, Grüne Liga Sachsen e.a., C-399/14, EU:C:2016:10, punt 41 en aldaar aangehaalde rechtspraak).

41 Bijgevolg is artikel 6, leden 1 en 2, van de habitatrichtlijn in omstandigheden als in de hoofdgedingen niet van toepassing.

[…]

48 Wat meer in het bijzonder het antwoord betreft dat op de gestelde vraag moet worden gegeven, dient in de eerste plaats erop te worden gewezen dat het Hof in punt 29 van het arrest van 15 mei 2014, Briels e.a. (C-521/12, EU:C:2014:330), heeft geoordeeld dat de beschermingsmaatregelen die in een project worden opgenomen om de schadelijke gevolgen van dit project voor een Natura 2000-gebied te compenseren, bij de door artikel 6, lid 3, opgelegde beoordeling van de gevolgen van dit project niet in aanmerking kunnen worden genomen.

49 Het is juist dat in de hoofdgedingen geen sprake is van dezelfde omstandigheden als in de zaak die heeft geleid tot het arrest van 15 mei 2014, Briels e.a. (C-521/12, EU:C:2014:330), aangezien de daarin voorgenomen maatregelen moeten worden uitgevoerd voordat er aantastingen plaatsvinden, terwijl in die andere zaak de maatregelen zouden worden uitgevoerd na het plaatsvinden van de aantastingen.

50 Evenwel wordt in de rechtspraak van het Hof benadrukt dat de overeenkomstig artikel 6, lid 3, van de habitatrichtlijn verrichte beoordeling geen leemten mag vertonen en volledige, precieze en definitieve constateringen en conclusies moet bevatten die elke redelijke wetenschappelijke twijfel over de gevolgen van de geplande werkzaamheden voor het betrokken beschermde gebied kunnen wegnemen (arrest van 14 januari 2016, Grüne Liga Sachsen e.a., C-399/14, EU:C:2016:10, punt 50 en aldaar aangehaalde rechtspraak).

51 Daarbij is het zo dat de passende beoordeling van de gevolgen van een plan of project voor het betrokken gebied die ingevolge artikel 6, lid 3, moet worden verricht, inhoudt dat, op basis van de beste wetenschappelijke kennis ter zake, alle aspecten van het betrokken plan of project die op zichzelf of in combinatie met andere plannen of projecten de instandhoudingsdoelstellingen van dat gebied in gevaar kunnen brengen, moeten worden geïnventariseerd (arrest van 14 januari 2016, Grüne Liga Sachsen e.a., C-399/14, EU:C:2016:10, punt 49 en aldaar aangehaalde rechtspraak).

52 Verder dient erop te worden gewezen dat de eventuele positieve gevolgen van het later

ontwikkelen van een nieuwe habitat waarmee het verlies aan oppervlakte en kwaliteit van ditzelfde type habitat in een beschermd gebied dient te worden gecompenseerd, in de regel onzeker zijn, en dat deze gevolgen hoe dan ook slechts binnen enkele jaren zichtbaar zullen worden (zie in die zin arrest van 15 mei 2014, Briels e.a., C-521/12, EU:C:2014:330, punt 32).

53 In de tweede plaats is het zo dat in artikel 6, lid 3, van de habitatrichtlijn tevens het voorzorgsbeginsel ligt besloten, en dat met deze bepaling op efficiënte wijze kan worden voorkomen dat de natuurlijke kenmerken van beschermde gebieden worden aangetast als gevolg van plannen of projecten. Met een minder streng toestemmingscriterium dan het daarin genoemde criterium zou de met deze bepaling beoogde verwezenlijking van de doelstelling van bescherming van die gebieden niet even goed kunnen worden gegarandeerd (zie in die zin arrest van 14 januari 2016, Grüne Liga Sachsen e.a., C-399/14, EU:C:2016:10, punt 48 en aldaar aangehaalde rechtspraak).

54 Het voorzorgsbeginsel verlangt van de bevoegde nationale instantie dat zij bij de toepassing van artikel 6, lid 3, van de richtlijn de gevolgen van het project voor het betrokken gebied beoordeelt in het perspectief van de instandhoudingsdoelstellingen van dit gebied, rekening houdend met de in dit project vastgestelde beschermingsmaatregelen waarmee wordt beoogd de eventuele schadelijke gevolgen die rechtstreeks uit dit project voortvloeien, te voorkomen of te verminderen, teneinde ervoor te zorgen dat het betrokken project de natuurlijke kenmerken van dat gebied niet aantast (arrest van 15 mei 2014, Briels e.a., C-521/12, EU:C:2014:330, punt 28).

55 In casu staan de aantastingen van het betrokken Natura 2000-gebied vast, aangezien de verwijzende rechter de omvang ervan heeft kunnen bepalen. Daarnaast zijn de uit de ontwikkeling van natuurkerngebieden voortvloeiende voordelen reeds meegenomen in de beoordeling en bij het aantonen dat er geen betekenisvolle aantasting van dat gebied is, terwijl het resultaat van de ontwikkeling van die natuurkerngebieden onzeker is, daar de ontwikkeling onvoltooid is.

56 De omstandigheden in de hoofdgedingen en die welke aan de orde waren bij het wijzen van het arrest van 15 mei 2014, Briels e.a. (C-521/12, EU:C:2014:330), zijn dan ook vergelijkbaar in de zin dat bij de beoordeling van de gevolgen van het plan of project voor het betrokken gebied, wordt uitgegaan van dezelfde aanname dat toekomstige voordelen de significante effecten op dat gebied zullen mitigeren, terwijl de ontwikkelingsmaatregelen in kwestie niet zijn voltooid.

57 In de derde plaats moet erop worden gewezen, zoals in punt 33 van dit arrest is aangegeven, dat in de bewoordingen van artikel 6 van de habitatrichtlijn geen sprake is van enigerlei „mitigerende maatregelen".

58 Zoals het Hof in dit verband reeds heeft verklaard, bestaat de nuttige werking van de in artikel 6 van de habitatrichtlijn genoemde beschermingsmaatregelen erin te voorkomen dat de bevoegde nationale instantie via zogenoemde mitigerende maatregelen, die in werkelijkheid compenserende maatregelen zijn, de in dit artikel vastgelegde specifieke procedures omzeilt en krachtens lid 3 van dat artikel projecten toestaat die de natuurlijke kenmerken van het betrokken gebied aantasten (arrest van 15 mei 2014, Briels e.a., C-521/12, EU:C:2014:330, punt 33)".

10.12. In het arrest Kolencentrale Moorburg is de vraag aan de orde of een vistrap een beschermingsmaatregel is die betrokken mag worden in een passende beoordeling voor de bouw van de kolencentrale. Het koelsysteem van de kolencentrale zal leiden tot de dood van bepaalde vissoorten waarvoor Natura 2000-gebieden die stroomopwaarts van de kolencentrale liggen, zijn aangewezen. Met de vistrap, die tussen de kolencentrale en de Natura 2000-gebieden is voorzien, wordt beoogd de negatieve gevolgen van het visverlies dat wordt veroorzaakt door de werking van het koelsysteem van de centrale, te verminderen. Het Hof van Justitie overwoog:

"34. De Duitse autoriteiten dienden rekening te houden met de in de bouwplannen vastgestelde beschermingsmaatregelen teneinde zich ervan te vergewissen dat het project voor de bouw van de centrale van Moorburg de natuurlijke kenmerken van de betrokken Natura 2000-gebieden niet aantast. Het is dienaangaande vaste rechtspraak dat het voorzorgsbeginsel van de bevoegde nationale instantie verlangt dat zij bij de toepassing van artikel 6, lid 3, van de habitatrichtlijn met name rekening houdt met de in dit project vastgestelde beschermingsmaatregelen waarmee wordt beoogd de eventuele schadelijke gevolgen die rechtstreeks uit dit project voortvloeien, te voorkomen of te verminderen, teneinde ervoor te zorgen dat het betrokken project de natuurlijke kenmerken van het beschermde gebied niet aantast (arresten van 15 mei 2014, Briels e.a., C521/12, EU:C:2014:330, punt 28, en van 21 juli 2016, Orleans e.a., C387/15 en C388/15, EU:C:2016:583, punt 54).

35 In casu moet worden opgemerkt dat uit het aan het Hof overgelegde dossier blijkt dat, naast andere maatregelen die zijn getroffen om de negatieve gevolgen van de onttrekking van water te verhinderen, zoals de installatie van visafweersystemen, het terugleiden van vissen en de vermindering van de activiteiten van de centrale van Moorburg wanneer het zuurstofgehalte een voor vissen kritiek niveau heeft bereikt, een vistrap is geïnstalleerd ter hoogte van de Geesthachtdam.

36 Deze vistrap zou tot een uitbreiding van de bestanden van migrerende vissen kunnen leiden door die soorten de mogelijkheid te bieden sneller hun voortplantingsgebieden in de midden- en bovenloop van de Elbe te bereiken. De uitbreiding van de bestanden zou de verliezen bij de centrale van Moorburg compenseren zodat er geen significante gevolgen zouden zijn voor de instandhoudingsdoelstellingen van de stroomopwaarts van deze centrale gelegen Natura 2000-gebieden.

37 Uit de effectbeoordeling blijkt echter dat deze geen definitieve bevindingen bevat omtrent de doeltreffendheid van de vistrap, maar enkel preciseert dat pas na verscheidene jaren van toezicht zal worden bevestigd of deze vistrap al dan niet doeltreffend is.

38 Hoewel deze vistrap tot doel had om rechtstreekse significante gevolgen voor de stroomopwaarts van de centrale van Moorburg gelegen Natura 2000-gebieden te verminderen, dient dus te worden vastgesteld dat hij, samen met de andere in punt 35 van het onderhavige arrest vermelde maatregelen, op het ogenblik van de afgifte van de vergunning niet alle redelijke twijfel kon wegnemen dat die centrale de natuurlijke kenmerken van het gebied niet zou aantasten in de zin van artikel 6, lid 3, van de habitatrichtlijn.

39 Aan deze conclusie kan niet worden afgedaan door de argumenten van de Bondsrepubliek Duitsland met betrekking tot het beheer van risico's en de meegedeelde bevindingen voor de

jaren 2011 tot en met 2014.

[…]

43 Voorts volstaat het toezicht in verscheidene fases niet om de nakoming van de verplichting van artikel 6, lid 3, van de habitatrichtlijn te waarborgen."

Rechtspraak van de Afdeling

10.13. Zoals de Afdeling eerder heeft overwogen (uitspraak van 24 december 2014, ECLI:NL:RVS:2014:4672) leidt zij uit het arrest Briels af dat voor het oordeel of een project de natuurlijke kenmerken van een gebied aantast, alle rechtstreekse gevolgen van dat project in het licht van de instandhoudingsdoelstellingen moeten worden beschouwd, waarbij bepalend is of de bepalende kenmerken van het gebied die verband houden met de natuurlijke habitats waarvoor instandhoudingsdoelstellingen zijn gesteld, duurzaam behouden blijven. Dit oordeel betreft derhalve de gevolgen van het project voor het totale bestaande areaal van een habitattype in een Natura 2000-gebied. Voor dit onderzoek dienen de gevolgen per habitattype, per locatie van voorkomen van het habitattype, in kaart gebracht te worden.

De Afdeling heeft in voornoemde uitspraak voorts overwogen dat zij uit het arrest Briels afleidt dat bij de beoordeling of een project leidt tot een aantasting van de natuurlijke kenmerken van een gebied, slechts die beschermingsmaatregelen mogen worden betrokken, waarmee wordt beoogd de schadelijke gevolgen die rechtstreeks uit het project voortvloeien te voorkomen of te verminderen ter plaatse van de locatie van het voorkomen van het habitattype dat negatieve gevolgen van het project zou ondervinden als deze maatregelen niet worden getroffen. Positieve gevolgen van maatregelen voor een areaal van een habitattype waarvoor een project geen negatieve effecten heeft, kunnen niet worden betrokken bij de beoordeling of een project leidt tot een aantasting van de natuurlijke kenmerken van het gebied. De hiervoor bedoelde beschermingsmaatregelen die in een passende beoordeling mogen worden betrokken worden in de rechtspraak van de Afdeling aangeduid als mitigerende maatregelen.

10.14. De Afdeling heeft in haar rechtspraak voorts aangenomen dat in een passende beoordeling van de gevolgen van een plan of project onder voorwaarden rekening gehouden mag worden met instandhoudingsmaatregelen en autonome ontwikkelingen.

10.15. Herstelmaatregelen die geheel los van een specifiek plan of project worden getroffen om de instandhoudingsdoelen te kunnen realiseren, worden in de rechtspraak van de Afdeling geduid als instandhoudingsmaatregelen (vergelijk de uitspraken van de Afdeling van 30 oktober 2013, ECLI:NL:RVS:2013:1694 en 10 december 2014, ECLI:NL:RVS:2014:4431). Deze instandhoudingsmaatregelen kunnen wegens het ontbreken van een directe samenhang met een plan of project niet als mitigerende of compenserende maatregel worden aangemerkt. De Afdeling oordeelde in de uitspraak van 27 januari 2016, ECLI:NL:RVS:2016:170, dat instandhoudingsmaatregelen als feitelijke ontwikkeling in een passende beoordeling kunnen worden betrokken als aan de volgende voorwaarden is voldaan: "met een voldoende mate van zekerheid dient vast te staan dat de maatregelen daadwerkelijk zullen worden uitgevoerd. Verder dienen niet alleen de verwachte positieve effecten, maar ook eventuele negatieve effecten daarvan op de kwalificerende habitattypen, habitatsoorten en vogelsoorten te worden beoordeeld in het licht van de

instandhoudingsdoelstellingen voor de betrokken Natura 2000-gebieden. In dat verband moet inzichtelijk worden gemaakt dat geen afbreuk wordt gedaan aan de effectiviteit van de beoogde instandhoudingsmaatregelen die een bijdrage moeten leveren aan het bereiken van de instandhoudingsdoelstellingen voor habitattypen en soorten in het Natura 2000-gebied. Uit de passende beoordeling moet ten slotte blijken dat het - op termijn - behalen van die instandhoudingsdoelstellingen niet in gevaar wordt gebracht".

10.16. Herstelmaatregelen die specifiek met het oog op de uitvoering van een plan of project en aanvullend op het bestaande beheer worden getroffen, worden door de Afdeling geduid als mitigerende maatregel als de maatregel gericht is op het voorkomen of verminderen van de gevolgen ter plaatse van voorkomens van habitattypen die negatieve gevolgen van het plan of project zouden ondervinden als de maatregel niet zou worden getroffen (bijv. de uitspraak van 24 december 2014, ECLI:NL:RVS:2014:4672). Deze laatste voorwaarde heeft de Afdeling afgeleid uit het arrest Briels waarin het Hof oordeelde dat in een passende beoordeling alleen beschermingsmaatregelen mogen worden betrokken die de rechtstreekse gevolgen van een plan of project voorkomen of verzachten. Verder dient de effectiviteit van de maatregel zeker te zijn (zeker moet zijn dat de positieve gevolgen zich ook daadwerkelijk zullen manifesteren) en dient het treffen van de mitigerende maatregel juridisch verzekerd te zijn. Dat laatste gebeurt in de regel door aan de vergunning voor het project voorschriften te verbinden.

Ook herstelmaatregelen die weliswaar niet gericht zijn op de afname van depositie, maar op het voorkomen van de gevolgen van depositie voor stikstofgevoelige natuurwaarden, zoals het maaien en plaggen waarmee stikstofdepositie uit het gebied wordt weggehaald, worden door de Afdeling als mitigerende maatregel geduid.

10.17. Met de autonome daling van stikstofdepositie die het gevolg is van bestaand beleid kan volgens de Afdeling rekening worden gehouden in een passende beoordeling (vergelijk de uitspraak van 20 april 2016, ECLI:NL:RVS:2016:1060). Uit de uitspraken van 7 december 2011, ECLI:NL:RVS:2011:BU7002 en 31 oktober 2012, ECLI:NL:RVS:2012:BY1743, volgt dat met de autonome daling rekening mag worden gehouden mits voldoende zeker is dat de afname optreedt en inzichtelijk is gemaakt in hoeverre het plan of project het behalen van verbeterdoelstellingen zal vertragen, dan wel van het behalen daarvan in de weg zal staan.

10.18. Zoals uit het voorgaande kan worden afgeleid hebben de in de rechtspraak van de Afdeling geformuleerde voorwaarden, waaronder instandhoudingsmaatregelen, passende maatregelen, mitigerende maatregelen en autonome ontwikkelingen in een passende boordeling kunnen worden betrokken, met name betrekking op de plaats waar de maatregelen zullen worden getroffen of de plaats waar de positieve gevolgen van maatregelen zich voordoen, de zekerheid dat de maatregelen zullen worden getroffen en de zekerheid dat de maatregelen effectief zullen zijn. Herstelmaatregelen die positieve gevolgen hebben voor arealen van habitattypen waarvoor een project in het geval die maatregelen niet zouden worden getroffen, negatieve gevolgen zou hebben, zijn volgens de Afdeling te duiden als beschermingsmaatregelen als bedoeld in punt 28 van het arrest Briels, die in een passende beoordeling mogen worden betrokken. Als deze rechtspraak wordt toegepast op de passende beoordeling voor het PAS, dan zouden de maatregelen en ontwikkelingen meegenomen kunnen worden in de passende beoordeling indien voldaan is aan de volgende voorwaarden:

(a) de bron- en herstelmaatregelen dienen positieve gevolgen te hebben ter plaatse van de arealen van stikstofgevoelige natuurwaarden waarvoor de benutting van de depositieruimte waarin het programma voorziet in het geval de bron- en herstelmaatregelen niet zouden worden getroffen, negatieve gevolgen zou hebben;

(b) het treffen van de bron- en herstelmaatregelen dient in het kader van de uitvoering van het programma voldoende te zijn verzekerd;

(c) de effectiviteit van de bron- en herstelmaatregelen moet voldoende zeker zijn en voor zover het de afname van stikstofdepositie betreft dient de afname gebaseerd te zijn op een realistische prognose van de daling;

(d) met de autonome daling van stikstofdepositie kan rekening worden gehouden mits voldoende zeker is dat de afname optreedt en de afname gebaseerd is op een realistische prognose.

Het PAS in het licht van de rechtspraak van het Hof en de Afdeling

10.19. De Afdeling stelt voorop dat er een belangrijk verschil bestaat tussen het PAS en de situatie die aan de orde was in de zaken Briels en Orleans. In de zaken Briels en Orleans stond immers vast dat de realisering van de projecten zou leiden tot een blijvende aantasting of verlies van bestaande arealen van een habitattype waarvoor het Natura 2000-gebied was aangewezen en het tot ontwikkeling brengen van nieuwe natuur op een andere locatie binnen hetzelfde Natura 2000-gebied. Uit de passende beoordeling van het PAS volgt dat de depositie in 2014 en de depositie die kan ontstaan door benutting van de depositieruimte, rekening houdend met de autonome daling van de stikstofdepositie, de PAS-bronmaatregelen en de bestaande en voorgenomen herstelmaatregelen, de bestaande arealen van habitattypen en leefgebieden van soorten niet zal aantasten.

10.20. Over het betoog van de Werkgroep dat de uitgifte van ontwikkelingsruimte direct na de inwerkingtreding van het PAS kan leiden tot een toename van stikstofdepositie waardoor een verslechtering van habitattypen niet is uitgesloten, wordt het volgende overwogen.

In de gebiedsanalyses is rekening gehouden met de mogelijkheid dat een deel van de ontwikkelingsruimte direct na de inwerkingtreding van het PAS kan worden toegedeeld (zie hierover 6.5). Daarbij is in aanmerking genomen dat een tijdelijke toename van de stikstofdepositie kan plaatsvinden wanneer de uitgifte en benutting sneller verlopen dan de daling van de stikstofdepositie. Wanneer de stikstofdepositie tijdelijk toeneemt zou dat voorafgaand aan of tijdens de uitvoering van de herstelmaatregelen kunnen leiden tot zuurdere en voedselrijkere condities (van bodem en water) en tot een grotere beschikbaarheid van voedingsstoffen en mineralen voor de vegetatie. Volgens de passende beoordeling wordt met de voorgestelde herstelmaatregelen voorkomen dat die tijdelijke situatie daadwerkelijk tot verslechtering van habitattypen leidt. De habitattypen hebben een relatief lange responstijd op veranderingen in het abiotische systeem, terwijl een aantal maatregelen die aan het begin van het tijdvak worden genomen een korte responstijd hebben en dus relatief een snel effect hebben. Dit houdt volgens de passende beoordeling in dat binnen de responstijd van de habitattypen op een eventuele toename van depositie, de noodzakelijke maatregelen worden genomen die ervoor zorgen dat er geen achteruitgang van de kwaliteit of het oppervlakte van habitattypen optreedt. Een tijdelijke toename van

depositie in de eerste helft van het tijdvak van het programma leidt volgens de gebiedsanalyses dan ook niet tot een ecologische verslechtering van de stikstofgevoelige habitattypen en leefgebieden.

Gelet op het gestelde in de gebiedsanalyse, volgt de Afdeling de Werkgroep niet in haar betoog dat verslechtering van habitattypen niet is uitgesloten omdat direct na inwerkingtreding van het PAS ontwikkelingsruimte kan worden uitgegeven.

10.21. De Werkgroep wijst er voorts op dat projecten en andere handelingen die op grond van het PAS-beoordelingskader kunnen worden toegestaan ook na afloop van de looptijd van het PAS van zes jaar nog stikstofdepositie zullen veroorzaken. Zij acht niet uitgesloten dat deze projecten na afloop van de looptijd van het PAS alsnog verslechtering van stikstofgevoelige habitats tot gevolg zullen hebben. Zij stelt dat een programma met een looptijd van zes jaar niet ten grondslag mag worden gelegd aan vergunningen voor onbepaalde tijd, omdat de in het PAS voorziene herstelmaatregelen slechts zijn verzekerd voor zes jaar.

10.21.1. In artikel 19kg, vijfde lid, van de Nbw 1998 is voorgeschreven dat het programma ten minste eenmaal in de zes jaar wordt vastgesteld en voor een tijdvak van zes jaar geldt. Daaruit volgt dat na de eerste periode van zes jaar (2015-2021) een nieuw programma moet worden vastgesteld. Bij de vaststelling van een programma moet artikel 19kh, eerste lid, van de Nbw 1998 in acht worden genomen. Dit betekent onder meer dat, gelet op het bepaalde in het eerste lid, onder a en f, zowel de omvang van de stikstofdepositie aan het begin van het tijdvak moet worden beschreven als de wijze waarop en frequentie waarmee de rapportage plaatsvindt over de voortgang en uitvoering van de getroffen of te treffen in het programma beschreven en genoemde maatregelen en de effecten daarvan op de depositie. Daarnaast moeten op grond van het eerste lid, onder g, de getroffen of te treffen herstelmaatregelen in het programma worden beschreven of genoemd. Dat in artikel 5, derde lid, van de Regeling staat dat in een toestemmingsbesluit dat geldig is voor onbepaalde tijd het bevoegd gezag ontwikkelingsruimte eenmalig toekent voor onbepaalde tijd betekent, anders dan de Werkgroep betoogt, dan ook niet dat na afloop van het tijdvak van zes jaar projecten of andere handelingen die stikstof veroorzaken zullen zijn toegestaan zonder dat in de desbetreffende gebieden herstelmaatregelen reeds zijn of zullen worden getroffen en hun effect al hebben of zullen hebben. Het samenstel van voormelde bepalingen verzekert derhalve dat het huidige programma zal worden opgevolgd door een nieuw programma, waarbij rekening moet worden gehouden met zowel de reeds uitgegeven ontwikkelingsruimte bij de bepaling van de omvang van de stikstofdepositie aan het begin van dat tijdvak als met de reeds getroffen en de nog te treffen herstelmaatregelen. Als blijkt dat de stikstofdepositie hoger is dan verwacht zal de beschikbare depositieruimte in het tweede tijdvak daarop moeten worden aangepast.

Gelet op het voorgaande volgt de Afdeling de Werkgroep niet in haar betoog dat verslechtering van habitattypen niet is uitgesloten omdat vergunningen voor onbepaalde tijd worden verleend, terwijl het PAS voor een periode van 6 jaar wordt vastgesteld.

10.22. Aangezien uit de passende beoordeling van het PAS volgt dat het PAS, mede gelet op de te treffen maatregelen en ontwikkelingen, niet zal leiden tot een verslechtering van de bestaande habitattypen en leefgebieden is in deze zaak, anders dan in de zaken Briels en Orleans, de vraag aan de orde of, en zo ja onder welke voorwaarden in een passende

beoordeling rekening mag worden gehouden met instandhoudingsmaatregelen, passende maatregelen en maatregelen die specifiek voor het programma worden genomen (mitigerende maatregelen) en met autonome ontwikkelingen die de negatieve gevolgen van stikstofdepositie voor bestaande arealen van habitattypen en leefgebieden van soorten zodanig voorkomen of verminderen dat verslechtering van de bestaande arealen van de habitattypen en de leefgebieden wordt voorkomen en herstel, waar nodig, mogelijk blijft. Daarnaast speelt de vraag of voor beschermingsmaatregelen en de andere maatregelen en ontwikkelingen waarmee verslechtering van bestaande arealen van habitattypen en leefgebieden door stikstofdepositie wordt voorkomen geldt, dat deze uitsluitend in een passende beoordeling mogen worden betrokken als ze op dat moment genomen zijn en resultaat hebben gehad.

Duiding van de maatregelen en ontwikkelingen die in de passende beoordeling zijn betrokken in het licht van artikel 6 Habitatrichtlijn

10.23. In de geschiedenis van de totstandkoming van de regeling over het PAS in de Nbw 1998 is aandacht besteed aan de juridische duiding van de PAS-bronmaatregelen en de herstelmaatregelen (Kamerstukken II 2013/14, 33 669, nr. 6, blz. 9). Hierin staat: "Het maatregelenpakket dat onderdeel uitmaakt van dat programma bevat zowel brongerichte maatregelen, die leiden tot een vermindering van de stikstofdepositie op Natura 2000-gebieden, als gebiedsgerichte maatregelen, die de kwaliteit van de Natura 2000-gebieden herstellen en verbeteren. De maatregelen zijn uit een juridisch oogpunt aan te duiden als instandhoudingsmaatregelen en passende maatregelen in de zin van artikel 6, eerste en tweede lid, van de Habitatrichtlijn en als mitigerende maatregelen die bij de passende beoordeling van een project mogen worden betrokken. Het is daarbij feitelijk niet mogelijk om per maatregel aan te geven of deze als instandhoudingsmaatregel, passende maatregel of mitigerende maatregel moet worden beschouwd. Dat is ook niet relevant, aangezien met het pakket aan maatregelen op programmaniveau uitvoering wordt gegeven aan zowel de verplichtingen op grond van artikel 6, eerste en tweede lid, van de Habitatrichtlijn als de verplichtingen in het kader van de concrete toestemmingverlening voor projecten met mogelijk significant negatieve gevolgen overeenkomstig artikel 6, derde lid, van de Habitatrichtlijn".

De Afdeling zal hierna weergeven hoe zij de in de passende beoordeling betrokken ontwikkelingen en maatregelen in het licht van artikel 6 van de Habitatrichtlijn duidt.

10.24. In de passende beoordeling is rekening gehouden met de verwachte daling van de stikstofdepositie door de PAS-bronmaatregelen (zie ook 6.8).

De PAS-bronmaatregelen betreffen stalmaatregelen, maatregelen voor emissiearme bemesting en voer- en managementmaatregelen. De stalmaatregelen bestaan uit emissie-eisen waaraan bij uitbreiding of wijziging van een agrarisch bedrijf moet worden voldaan. Deze maatregel zal gedurende een periode van 20 tot 30 jaar effect hebben. De aanpassing van de mestregelgeving zal na wijziging van de regelgeving direct tot een verlaging van de emissie leiden. De voer- en managementmaatregelen zullen, indien een bedrijf de maatregelen doorvoert, direct effect hebben.

De Afdeling beschouwt de PAS-bronmaatregelen in het licht van artikel 6 van de Habitatrichtlijn als passende maatregel of instandhoudingsmaatregel, als bedoeld in lid 1 en

2. De maatregelen leiden in het algemeen tot een afname van stikstofemissie door agrarische bedrijven, die kan leiden tot een afname van stikstofdepositie op Natura 2000-gebieden. De afname kan een bijdrage leveren aan het behoud of herstel van de stikstofgevoelige natuurwaarden in Natura 2000-gebieden.

10.25. In het PAS is ervan uitgegaan dat de autonome daling van de stikstofemissie zich zal blijven manifesteren in de eerste en vervolgens ook in de tweede en derde PAS-periode. De autonome daling is het gevolg van bestaande en voorgenomen beleidsmaatregelen die geen verband houden met het PAS. Voorbeelden zijn aanscherping van de emissienormen voor het wegverkeer (Euro 5 en Euro 6 normen) en de aanscherping van de emissienormen voor de industrie. In de autonome daling is ook rekening gehouden met een daling van de emissiebijdrage uit het buitenland, door de uitvoering van het Europese luchtkwaliteitsbeleid.

De autonome daling van de stikstofdepositie is een ontwikkeling die zich ook los van het PAS zou voordoen. De autonome daling draagt bij aan de verbetering van de specifieke milieukenmerken die van belang zijn voor het behoud of het herstel van stikstofgevoelige natuurwaarden. De maatregelen die deze daling veroorzaken kunnen, nu deze niet specifiek ter uitvoering van de Habitatrichtlijn zijn genomen, niet worden beschouwd als passende maatregel of instandhoudingsmaatregel. Het betreft een feitelijke ontwikkeling.

10.26. De herstelmaatregelen die in een Natura 2000-gebied moeten worden getroffen zijn uitgewerkt en beschreven in de gebiedsanalyse die voor elk Natura 2000-gebied is opgesteld. De herstelmaatregelen dragen enerzijds bij aan het op termijn realiseren van de instandhoudingsdoelstellingen van de stikstofgevoelige natuurwaarden in Natura 2000-gebieden en anderzijds aan het voorkomen van verslechtering van deze stikstofgevoelige natuurwaarden door de benutting van de depositie- en ontwikkelingsruimte. Voor de vergunningzaken is onder meer de gebiedsanalyse neergelegd in het rapport "PAS-analyse herstelmaatregelen voor de Natura 2000-gebieden 139 Deurnsche Peel & Mariapeel en 140 Groote Peel", gedateerd 19 november 2015 (hierna: de gebiedsanalyse), relevant.

De herstelmaatregelen die in de Groote Peel en de Deurnsche Peel & Mariapeel zullen worden getroffen betreffen onder meer het periodiek verwijderen van de opslag van berken en trosbosbes, het verwijderen van bomen, het begrazen aangevuld met plaggen en maaien. De maatregelen worden grotendeels getroffen aanvullend op het regulier beheer en zijn niet in andere plannen voorzien. Naast deze beheermaatregelen zijn hydrologische maatregelen voorzien, zoals het dempen en afdammen van watergangen en het opzetten van het waterpeil. De maatregelen hebben een hoger en stabieler grondwaterpeil tot doel. Een groot deel van de hydrologische maatregelen zijn of worden genomen als onderdeel van een landinrichtingsplan, de GGOR-plannen en LIFE-projecten. Een aantal maatregelen dat in het kader van de uitvoering van de GGOR-plannen zal worden uitgevoerd is een direct uitvloeisel van het PAS.

Als gevolg van de vernatting van het gebied zullen de overgangen van nat naar droog (en voedselarm naar voedselrijker) verschuiven van de centra van de Peelgebieden naar de randen van Peelgebieden. Deze overgangen zijn leefgebieden van vogelsoorten. Voor deze vogelsoorten zullen randzones worden ingericht zodat zij mee kunnen bewegen met de veranderingen in het Natura 2000-gebied.

10.26.1. De Afdeling stelt vast dat de herstelmaatregelen die in de gebiedsanalyse voor de

gebieden Groote Peel en Deurnsche Peel & Mariapeel zijn betrokken gericht zijn op het voorkomen van verslechtering van de bestaande arealen van stikstofgevoelige habitats, en waar nodig en mogelijk op het herstel daarvan. De herstelmaatregelen die ook los van het PAS worden getroffen kunnen worden geduid als instandhoudingsmaatregel of passende maatregel. De herstelmaatregelen die specifiek in het kader van het PAS worden getroffen houden direct verband met de gevolgen van de benutting van de depositie-/ontwikkelingsruimte voor de bestaande arealen van stikstofgevoelige habitats. Het karakter van deze maatregelen wordt besproken in 10.35.

10.26.2. Het betoog van de Werkgroep dat het inrichten van de randzones voor de vogelsoorten een compenserende maatregel is omdat deze maatregel gericht is op de ontwikkeling van nieuw leefgebied voor soorten, deelt de Afdeling niet. Uit de gebiedsanalyse volgt dat de hydrologische maatregelen die getroffen worden met het oog op het behoud en herstel van de stikstofgevoelige habitattypen herstellend hoogveen en actief hoogveen, zullen leiden tot vernatting van deze arealen. Deze arealen zijn na de vernatting minder geschikt als leefgebied voor vogelsoorten waarvoor het Natura 2000-gebied eveneens is aangewezen. Voor deze soorten worden maatregelen getroffen zodat daarvoor voldoende geschikt leefgebied aanwezig blijft. De Afdeling is van oordeel dat die maatregel niet samenhangt met de benutting van depositie-/ontwikkelingsruimte, maar het gevolg is van het uitvoeren van de herstelmaatregelen voor de stikstofgevoelige habitattypen. Het betreft derhalve geen maatregel ter compensatie van door de toegestane depositie-/ontwikkelingsruimte verloren gegane natuurwaarden, maar een instandhoudings- of passende maatregel.

Instandhoudings- en passende maatregelen en autonome ontwikkelingen in de passende beoordeling

10.27. Zoals uit 10.24 en 10.26.1 volgt, duidt de Afdeling de PAS-bronmaatregelen en de herstelmaatregelen die los van het PAS worden genomen als instandhoudingsmaatregel of passende maatregel. De autonome daling van de stikstofdepositie betreft een feitelijke ontwikkeling. Over de vraag of deze maatregelen en ontwikkelingen in een passende beoordeling mogen worden betrokken, wordt het volgende overwogen.

10.28. De Werkgroep leidt uit de punten 33 tot en met 41 van het arrest Orleans af dat in de passende beoordeling niet mede de effecten van instandhoudingsmaatregelen en passende maatregelen voor de bestaande habitats en leefgebieden van soorten mogen worden betrokken. Verder stelt zij dat in het geval instandhoudingsmaatregelen, passende maatregelen, en autonome ontwikkelingen in een passende beoordeling kunnen worden betrokken, dat alleen kan wanneer de positieve gevolgen van die maatregelen en ontwikkelingen zich hebben voorgedaan op het moment waarop de passende beoordeling wordt gemaakt. De Werkgroep leidt dit af uit punt 32 van het arrest Briels en punt 56 van het arrest Orleans. Zij wijst erop dat in de passende beoordeling van het PAS maatregelen zijn betrokken waarvan het resultaat nog moet worden afgewacht. Ook de depositiedaling waarmee in het PAS rekening wordt gehouden is een ontwikkeling die zich nog niet heeft voorgedaan, maar die zich zal moeten manifesteren tijdens de looptijd van het programma, aldus de Werkgroep.

10.29. De punten 33 tot en met 41 van het arrest Orleans hebben betrekking op de vraag of een plan waarin als voorwaarde is opgenomen dat nieuwe natuurwaarden eerst tot

ontwikkeling zullen worden gebracht voordat de uitbreiding van de haven van Antwerpen die zal leiden tot het verlies van habitattypen mag plaatsvinden, als een instandhoudings- of passende maatregel kan worden geduid. Het Hof van Justitie beantwoordt die vraag ontkennend omdat de maatregel, het ontwikkelen van nieuwe natuur, geen betrekking heeft op de instandhouding van de natuurlijke habitats en evenmin verslechtering of verstoring daarvan voorkomt, nu het plan tot een verlies van 20 hectare van een habitattype leidt.

In punt 32 van het arrest Briels wijst het Hof van Justitie erop dat de eventuele positieve gevolgen van het achteraf tot ontwikkeling brengen van een nieuwe habitat in de regel onzeker zijn en slechts binnen enkele jaren zichtbaar zullen worden, zodat daarmee in het kader van de passende beoordeling geen rekening meegehouden kan worden. In het arrest Orleans verduidelijkt het Hof dat standpunt in punt 56 waaruit volgt dat bij de beoordeling van de gevolgen van een plan of project voor een Natura 2000-gebied niet mag worden uitgegaan van de aanname dat toekomstige voordelen de significante effecten op het gebied zullen mitigeren, terwijl de ontwikkelingsmaatregelen in kwestie niet zijn voltooid.

10.30. De Afdeling overweegt dat het oordeel van het Hof van Justitie in de hiervoor aangehaalde punten uit de arresten Briels en Orleans betrekking heeft op maatregelen die niet in een passende beoordeling mogen worden betrokken, omdat deze niet zijn gericht op het voorkomen van de aantasting van bestaande arealen van habitattypen. Zoals hiervoor is weergegeven strekken de maatregelen waarmee in de passende beoordeling van het PAS rekening is gehouden tot het behoud en voorkomen van verslechtering en significante verstoring van bestaande stikstofgevoelige natuurwaarden. Het PAS en de daarin betrokken maatregelen verschillen naar het oordeel van de Afdeling dan ook wezenlijk van de ingreep en de maatregel die in de arresten Briels en Orleans door het Hof van Justitie werden beoordeeld.

10.31. De aanpak van de stikstofproblematiek in het PAS is gericht op de daling van de stikstofdepositie in de Natura 2000-gebieden door bronmaatregelen die langdurig effect zullen hebben en op het treffen van herstelmaatregelen in de gebieden, waarbij bepaalde herstelmaatregelen eerst kunnen worden getroffen nadat andere zijn genomen en andere herstelmaatregelen regelmatig moeten worden herhaald. Zoals de Afdeling heeft overwogen in 10.4 acht zij een programmatische aanpak zoals het PAS dat een integrale aanpak beoogt van de stikstofproblematiek in de Natura 2000-gebieden, niet op voorhand een ongeschikt instrument om uitvoering te geven aan de verplichtingen die uit artikel 6 van de Habitatrichtlijn voortvloeien. Aan een programma zoals het PAS, dat enerzijds gericht is op het behoud en waar mogelijk herstel van natuurwaarden en anderzijds op het scheppen van depositieruimte voor bestaande en toekomstige activiteiten die in samenhang worden beoordeeld, is inherent dat de gevolgen van het benutten van de depositieruimte worden beoordeeld in samenhang met alle maatregelen en autonome ontwikkelingen die zich tijdens de programmaperiode in het Natura 2000-gebied zullen voordoen. Dat zijn voor een deel ook maatregelen die nog niet zijn getroffen of nog geen resultaat hebben gehad.

Indien, zoals de Werkgroep betoogt, een dergelijke integrale beoordeling van de maatregelen, autonome ontwikkelingen en depositieruimte binnen de systematiek van artikel 6 van de Habitatrichtlijn niet is toegestaan, is een programmatische aanpak van de stikstofproblematiek die gericht is op het behoud en waar mogelijk herstel van natuurwaarden en op het scheppen van depositieruimte voor economische activiteiten niet mogelijk. Dat heeft als verstrekkend gevolg dat het PAS niet kan worden gebruikt voor het

verlenen van toestemming voor stikstofveroorzakende activiteiten.

10.32. De Afdeling leidt uit de arresten Briels en Orleans niet af dat instandhoudingsmaatregelen, passende maatregelen en autonome ontwikkelingen niet in de passende beoordeling voor een programma zoals het PAS, mogen worden betrokken, dan wel daarin pas mogen worden betrokken nadat ze zijn uitgevoerd en resultaat hebben gehad. Dergelijke maatregelen en ontwikkelingen kunnen naar het oordeel van de Afdeling in een passende beoordeling worden betrokken indien voldaan wordt aan de in 10.18 vermelde voorwaarden die betrekking hebben op de plaats waar de maatregelen getroffen worden of de positieve gevolgen van de maatregelen zich voor zullen doen, de zekerheid dat de maatregelen zullen worden getroffen en de zekerheid dat de maatregelen effectief zullen zijn. De passende beoordeling dient hierover definitieve bevindingen te bevatten die gebaseerd zijn op de beste wetenschappelijke kennis ter zake.

10.33. De omstandigheid dat in het PAS is voorzien in een jaarlijkse monitoring van zowel de depositieontwikkeling als de voortgang van de uitvoering en het resultaat van de maatregelen en dat bijsturing in het geval de gevolgen van de maatregelen ongunstiger zijn dan waarvan in de passende beoordeling is uitgegaan, indien nodig, plaatsvindt, kan niet afdoen aan de uit artikel 6, derde lid, van de Habitatrichtlijn voortvloeiende verplichting dat op basis van de passende beoordeling de zekerheid moet zijn verkregen dat de natuurlijke kenmerken van de Natura 2000-gebieden niet zullen worden aangetast (vergelijk de uitspraak van de Afdeling van 24 augustus 2011, ECLI:NL:RVS:2011:BR5684 en het hiervoor genoemde arrest Kolencentrale Moorburg van het Hof van Justitie). De Afdeling acht echter aannemelijk dat monitoring van de werking van het programma, de uitvoering van de maatregelen en de op basis van de beste wetenschappelijke kennis gebaseerde definitieve bevindingen over de positieve gevolgen daarvan in een passende beoordeling die ten grondslag ligt aan een programmatische aanpak, zoals het PAS, een vereiste kan zijn. Zij acht daarvoor van belang dat het PAS gedurende de looptijd van het programma de basis biedt voor het verlenen van toestemming voor stikstofveroorzakende activiteiten en dat de gevolgen van het benutten van de depositieruimte die voor die activiteiten in het programma is gereserveerd is beoordeeld in samenhang met alle maatregelen en autonome ontwikkelingen die zich tijdens de programmaperiode in een Natura 2000-gebied zullen voordoen. Monitoring van de werking van het programma maakt het mogelijk dat gedurende de programmaperiode getoetst kan worden of de definitieve bevindingen over de gevolgen waarvan in de passende beoordeling op basis van de beste wetenschappelijke kennis is uitgegaan juist zijn. In het geval de gevolgen ongunstiger zijn dan waarvan in de passende beoordeling is uitgegaan, vindt bijsturing, indien vereist, plaats.. Bijsturing kan bestaan uit het vervangen of toevoegen van herstelmaatregelen of het bijstellen van de beschikbare ontwikkelingsruimte voor nieuwe stikstofveroorzakende activiteiten.

Prejudiciële vragen 3, 3a en 4

10.34. Aangezien naar het oordeel van de Afdeling uit artikel 6 van de Habitatrichtlijn noch uit het arrest Orleans kan worden afgeleid dat in een passende beoordeling voor een programma dat enerzijds gericht is op het behoud en waar mogelijk herstel van natuurwaarden en anderzijds op het scheppen van depositieruimte voor economische activiteiten geen rekening mag worden gehouden met instandhoudingsmaatregelen, passende maatregelen en autonome ontwikkelingen die zich tijdens de programmaperiode zullen voordoen, maar er ook geen rechtspraak van het Hof van Justitie is waaruit volgt dat

dergelijke maatregelen wel in een passende beoordeling voor zo'n programma kunnen worden betrokken, ziet de Afdeling aanleiding hierover de volgende vragen aan het Hof van Justitie te stellen.

3. Mogen in de passende beoordeling als bedoeld in artikel 6, derde lid, van de Habitatrichtlijn, die voor een programma, zoals het Programma Aanpak Stikstof 2015-2021, is gemaakt, de positieve gevolgen van instandhoudingsmaatregelen en passende maatregelen voor bestaande arealen van habitattypen en leefgebieden worden betrokken, die worden getroffen in verband met de verplichtingen die voortvloeien uit artikel 6, eerste en tweede lid, van de Habitatrichtlijn?

3a. Indien vraag 3 bevestigend wordt beantwoord: kunnen de positieve gevolgen van instandhoudingsmaatregelen en passende maatregelen in een passende beoordeling voor een programma worden betrokken als deze ten tijde van de passende beoordeling nog niet zijn uitgevoerd en het positieve effect daarvan nog niet is verwezenlijkt?

Is daarbij, ervan uitgaande dat de passende beoordeling definitieve bevindingen bevat over de gevolgen van deze maatregelen die gebaseerd zijn op de beste wetenschappelijke kennis ter zake, van belang dat de uitvoering en het resultaat van die maatregelen wordt gemonitord en indien daaruit volgt dat de gevolgen ongunstiger zijn dan waarvan is uitgegaan in de passende beoordeling, bijsturing, indien nodig, plaatsvindt?

4. Mogen de positieve gevolgen van de autonome daling van stikstofdepositie die zich zal gaan manifesteren in de periode waarin het Programma Aanpak Stikstof 2015-2021 geldt, in de passende beoordeling als bedoeld in artikel 6, derde lid, van de Habitatrichtlijn, worden betrokken?

Is daarbij, ervan uitgaande dat de passende beoordeling definitieve bevindingen bevat over deze ontwikkelingen die gebaseerd zijn op de beste wetenschappelijke kennis ter zake, van belang dat de autonome daling van stikstofdepositie wordt gemonitord, en indien daaruit volgt dat de daling ongunstiger is dan waarvan is uitgegaan in de passende beoordeling, bijsturing, indien nodig, plaatsvindt?

Beschermingsmaatregelen

10.35. De herstelmaatregelen die specifiek in het kader van het PAS worden getroffen en direct verband houden met de gevolgen van de benutting van de depositie-/ontwikkelingsruimte voor de bestaande arealen van stikstofgevoelige habitats, zijn niet gericht op de afname van stikstofdepositie, maar zijn gericht op het voorkomen van schadelijke gevolgen door stikstofdepositie voor de stikstofgevoelige natuurwaarden. Met vegetatiemaatregelen, zoals maaien, wordt stikstofdepositie uit het gebied verwijderd en de vernatting van het gebied leidt tot een verbetering van de omstandigheden voor de stikstofgevoelige natuurwaarden. Met deze maatregelen worden de negatieve gevolgen van stikstofdepositie voorkomen ter plaatse van een bestaand areaal van een stikstofgevoelige natuurwaarde. De Afdeling heeft in haar rechtspraak dergelijke herstelmaatregelen die in het kader van de uitvoering van een project worden genomen ter plaatse van bestaande habitattypen of leefgebieden van soorten ten einde de gevolgen van een project voor dat habitattype of leefgebied te voorkomen, geduid als beschermingsmaatregel als bedoeld in punt 28 van het arrest Briels (in de rechtspraak van de Afdeling wordt dit geduid als

mitigerende maatregel).

De Werkgroep duidt dergelijke maatregelen als compenserende maatregelen omdat vaststaat dat de stikstofdepositie in de Natura 2000-gebieden Groote Peel en Deurnsche Peel & Mariapeel, zou leiden tot verslechtering van de stikstofgevoelige natuurwaarden als de herstelmaatregelen niet worden genomen. De herstelmaatregelen zijn volgens de Werkgroep te vergelijken met de maatregelen die aan de orde waren in de arresten Briels en Orleans.

De Afdeling is van oordeel dat de herstelmaatregelen die in de passende beoordeling van het PAS zijn betrokken niet zijn te vergelijken met het ontwikkelen van nieuwe natuur op een andere locatie in het Natura 2000-gebied dan de locatie waar een plan of project negatieve gevolgen heeft, de maatregel die in de arresten Briels en Orleans aan de orde was.

In beide arresten heeft het Hof van Justitie geoordeeld dat in een passende beoordeling maatregelen mogen worden betrokken waarmee wordt beoogd de eventuele schadelijke gevolgen die rechtstreeks uit een project voortvloeien te voorkomen of te verminderen, teneinde ervoor te zorgen dat het betrokken project de natuurlijke kenmerken van dat gebied niet aantast. Aangezien uit beide arresten niet met zekerheid kan worden afgeleid of herstelmaatregelen die specifiek in het kader van het PAS worden getroffen en direct verband houden met de gevolgen van de benutting van de depositie-/ontwikkelingsruimte voor de bestaande arealen van stikstofgevoelige habitats en leefgebieden, maar die niet tot een afname van stikstofdepositie leiden, maar wel voorkomen dat de te hoge stikstofdepositie schadelijke gevolgen kan hebben voor die bestaande arealen, kunnen worden geduid als beschermingsmaatregel die in een passende beoordeling mag worden betrokken, ziet de Afdeling aanleiding ook hierover een vraag aan het Hof van Justitie te stellen.

10.36. In het geval de beschermingsmaatregelen in de passende beoordeling kunnen worden betrokken, dan kan dat volgens de Werkgroep alleen en voor zover de positieve gevolgen van die maatregelen en ontwikkelingen zich hebben voorgedaan op het moment waarop de passende beoordeling wordt gemaakt. De Werkgroep leidt dit af uit de arresten Briels en Orleans. Zij wijst erop dat in de passende beoordeling van het PAS herstelmaatregelen zijn betrokken waarvan het resultaat nog moet worden afgewacht.

10.37. Zoals is overwogen in 10.30 heeft het oordeel van het Hof van Justitie in de arresten Briels en Orleans betrekking op maatregelen die niet in een passende beoordeling mogen worden betrokken, omdat deze niet zijn gericht op het voorkomen van de aantasting van bestaande arealen van habitattypen. Maatregelen die in een passende beoordeling kunnen worden betrokken, omdat ze de negatieve gevolgen van een plan of project voorkomen ter plaatse van een areaal van een habitattype waar deze gevolgen zich zouden voordoen als de maatregel niet wordt genomen, zijn naar het oordeel van de Afdeling maatregelen die in het kader van de uitvoering van een plan of project worden getroffen. In het kader van het PAS betreft het herstelmaatregelen die ter uitvoering van het programma worden getroffen. De Afdeling acht aannemelijk dat dergelijke maatregelen niet reeds getroffen hoeven te zijn en effect hoeven te hebben op het moment waarop een passende beoordeling van een plan of project wordt gemaakt. De passende beoordeling kan juist worden gebruikt om te onderzoeken welke maatregelen nodig zijn om de gevolgen van plannen en projecten voor de natuurwaarden waarvoor de Natura 2000-gebieden zijn aangewezen te voorkomen. Wel dient verzekerd te zijn dat de maatregelen die in de passende beoordeling zijn betrokken

daadwerkelijk zullen worden getroffen. Voorts dient voldoende zeker te zijn dat de maatregel effectief zal zijn. De passende beoordeling dient hierover definitieve bevindingen te bevatten die gebaseerd zijn op de beste wetenschappelijke kennis ter zake.

De Afdeling ziet steun voor dit standpunt in de punten 37 en 38 van het arrest Kolencentrale Moorburg, waarin het Hof van Justitie de aanleg van een vistrap die ten tijde van de passende beoordeling nog niet was aangelegd, duidt als beschermingsmaatregel. Daarnaast kan uit het arrest worden afgeleid dat de passende beoordeling definitieve bevindingen dient te bevatten omtrent de doeltreffendheid van maatregelen die tot doel hebben om de rechtstreekse significante gevolgen van een project te voorkomen of te verminderen.

10.38. Zoals ook overwogen in 10.33 kan de omstandigheid dat in het PAS is voorzien in een jaarlijkse monitoring van zowel de depositieontwikkeling als de voortgang van de uitvoering en het resultaat van de maatregelen en dat bijsturing in het geval de gevolgen van de maatregelen ongunstiger zijn dan waarvan in de passende beoordeling is uitgegaan, indien nodig, plaatsvindt, niet afdoen aan de uit artikel 6, derde lid, van de Habitatrichtlijn voortvloeiende verplichting dat op basis van de passende beoordeling de zekerheid moet zijn verkregen dat de natuurlijke kenmerken van de Natura 2000-gebieden niet zullen worden aangetast (vergelijk de uitspraak van de Afdeling van 24 augustus 2011, ECLI:NL:RVS:2011:BR5684 en het hiervoor genoemde arrest Kolencentrale Moorburg van het Hof van Justitie). De Afdeling acht echter aannemelijk dat monitoring van de werking van het programma, de uitvoering van de maatregelen en de op basis van de beste wetenschappelijke kennis gebaseerde definitieve bevindingen over de positieve gevolgen daarvan in een passende beoordeling die ten grondslag ligt aan een programmatische aanpak, zoals het PAS, een vereiste kan zijn. Zij acht daarvoor van belang dat het PAS gedurende de looptijd van het programma de basis biedt voor het verlenen van toestemming voor stikstofveroorzakende activiteiten en dat de gevolgen van het benutten van de depositieruimte die voor die activiteiten in het programma is gereserveerd is beoordeeld in samenhang met alle maatregelen en autonome ontwikkelingen die zich tijdens de programmaperiode in een Natura 2000-gebied zullen voordoen. Monitoring van de werking van het programma maakt het mogelijk dat gedurende de programmaperiode getoetst kan worden of de definitieve bevindingen over de gevolgen waarvan in de passende beoordeling op basis van de beste wetenschappelijke kennis is uitgegaan juist zijn. In het geval de gevolgen ongunstiger zijn dan waarvan in de passende beoordeling is uitgegaan, vindt bijsturing, indien vereist, plaats. Bijsturing kan bestaan uit het vervangen of toevoegen van herstelmaatregelen of het bijstellen van de beschikbare ontwikkelingsruimte voor nieuwe stikstofveroorzakende activiteiten.

Prejudiciële vragen 5 en 5a

10.39. Aangezien naar het oordeel van de Afdeling uit de rechtspraak van het Hof van Justitie niet kan worden afgeleid dat in een passende beoordeling voor een programma dat gericht is op het behoud en waar mogelijk herstel van natuurwaarden en het scheppen van depositieruimte voor economische activiteiten geen rekening mag worden gehouden met de positieve gevolgen van beschermingsmaatregelen die zich gedurende de programmaperiode zullen gaan voordoen, maar er ook geen rechtspraak van het Hof van Jusititie is waaruit volgt dat dergelijke maatregelen in een passende beoordeling kunnen worden betrokken, ziet de Afdeling aanleiding hierover een vraag aan het Hof van Justitie te stellen.

5. Mogen herstelmaatregelen die in het kader van het Programma Aanpak Stikstof 2015-2021 worden getroffen en waarmee wordt voorkomen dat een bepaalde natuurbelastende factor, zoals stikstofdepositie, schadelijke gevolgen kan hebben voor bestaande arealen van habitattypen of leefgebieden, geduid worden als beschermingsmaatregel als bedoeld in punt 28 van het arrest van het Hof van Justitie van 15 mei 2014, Briels, ECLI:EU:C:2014:330, die in een passende beoordeling als bedoeld in artikel 6, derde lid, van de Habitatrichtlijn mogen worden betrokken?

5a. Indien vraag 5 bevestigend wordt beantwoord: kunnen de positieve gevolgen van beschermingsmaatregelen die in de passende beoordeling mogen worden betrokken, daarin worden betrokken, als deze ten tijde van de passende beoordeling nog niet zijn uitgevoerd en het positieve effect daarvan nog niet is verwezenlijkt?

Is daarbij, er van uitgaande dat de passende beoordeling definitieve bevindingen bevat over de gevolgen van deze maatregelen die gebaseerd zijn op de beste wetenschappelijke kennis ter zake, van belang dat de uitvoering en het resultaat van de maatregelen wordt gemonitord en indien daaruit volgt dat de gevolgen ongunstiger zijn dan waarvan is uitgegaan in de passende beoordeling, bijsturing, indien nodig, plaatsvindt?

Het depositieniveau uit 2014 als uitgangspunt

(slechts relevant voor de nationale procedure)

11. De Werkgroep stelt dat in de passende beoordeling ten onrechte het feitelijke stikstofdepositieniveau in 2014 als uitgangspunt is genomen, zonder dat is bezien of voor die deposities Nbw-vergunningen waren verleend. Volgens de Werkgroep dient de ontwikkeling van de stikstofdepositie in beeld gebracht te worden door de vergunde deposities in 2014 af te zetten tegen het verwachte depositieniveau in 2020 en 2030. De Werkgroep betwijfelt of er dan sprake zal zijn van een afname van stikstofdepositie.

Een tweede bezwaar dat de Werkgroep in dit verband aanvoert ziet op de legaliserende rol die het PAS heeft voor stikstofveroorzakende activiteiten die vóór 1 januari 2015 feitelijk bestonden en thans nog bestaan, maar waarvoor geen Nbw-vergunning is verleend. Een Nbw-vergunning kan op grond van het PAS-regime zonder meer worden verleend voor een project of andere handeling in de omvang die hoort bij de hoogste feitelijke veroorzaakte stikstofdepositie in de periode 2012-2014. Deze hoogste feitelijk veroorzaakte depositie in de periode 2012-2014 is voorts in het PAS-regime het uitgangspunt voor de vraag of een uitbreiding van een bestaand bedrijf waarvoor nog niet eerder een Nbw-vergunning is verleend leidt tot een toename van stikstofdepositie. De Werkgroep vindt deze wijze van legaliseren van projecten en andere handelingen die zonder de vereiste toestemming zijn ontstaan nadat artikel 6 van de Habitatrichtlijn van toepassing werd voor een Natura 2000-gebied, in strijd met artikel 6 van de Habitatrichtlijn. Dat is volgens de Werkgroep onder meer aan de orde in de vergunningen voor de bedrijven aan de [locatie 4] te Someren (zaaknr. 201600620/1/R2), Limburglaan 7 te Someren (zaaknr. 201600622/1/R2) en [locatie 2] te Someren (201600617/1/R2).

11.1. Het college stelt dat in de passende beoordeling van het PAS is onderzocht of de depositie die in 2014 plaatsvond en de depositie die gedurende de PAS-periode van zes jaar kan gaan plaatsvinden na benutting van de depositie- en ontwikkelingsruimte door

stikstofveroorzakende activiteiten tot een aantasting van de natuurlijke kenmerken van de relevante Natura 2000-gebieden zal leiden. Het college acht het niet in strijd met artikel 6 van de Habitatrichtlijn dat het PAS de basis biedt voor de verlening van een vergunning voor bedrijfsactiviteiten die in 2014 feitelijk plaatsvonden en thans nog, maar nog niet eerder zijn vergund, omdat de daarmee gepaard gaande deposities passend zijn beoordeeld.

11.2. De Afdeling overweegt dat uit artikel 19kg, eerste lid, van de Nbw 1998 volgt dat het PAS gericht is op de afname van de feitelijke depositie in de Natura 2000-gebieden. In het programma worden, zo volgt uit artikel 19kh, eerste lid, van de Nbw 1998, voor de betrokken Natura 2000-gebieden in elk geval de omvang van de stikstofdepositie aan het begin van het tijdvak van het programma, de verwachte autonome ontwikkelingen, de bijdrage van de bronmaatregelen, en de doelstellingen ten aanzien van de omvang van de stikstofdepositie beschreven of genoemd. De autonome daling van de stikstofdepositie en de daling als gevolg van de PAS-bronmaatregelen zijn in het PAS voor de verschillende tijdvakken in beeld gebracht ten opzichte van de feitelijke depositie in 2014, het jaar voorafgaande aan het eerste PAS. Het op deze wijze in beeld brengen van de verwachte feitelijke daling van de stikstofdepositie gedurende de verschillende PAS-periodes, acht de Afdeling gelet op het bepaalde in de artikelen 19kg en 19kh van de Nbw 1998 niet onjuist.

11.3. De feitelijke depositie in 2014 bestaat uit deposities veroorzaakt door verschillende sectoren in het binnen- en buitenland. Voor geen van die sectoren is een onderscheid gemaakt naar deposities van activiteiten waarvoor al dan niet een Nbw-vergunning is verleend. De depositiebijdrage van de landbouwsector, de sector waarop het betoog van de Werkgroep ziet, is derhalve afkomstig van activiteiten waarvoor een Nbw-vergunning is verleend, activiteiten waarvoor geen Nbw-vergunning is vereist omdat voor de activiteit al toestemming was verleend voordat artikel 6 van de Habitatrichtlijn van toepassing werd voor een Natura 2000-gebied en activiteiten waarvoor, hoewel vereist, geen Nbw-vergunning is verleend.

In het PAS is onderkend dat in de feitelijke depositie die in 2014 plaatsvond, deposities zijn opgenomen waarvoor nog niet eerder een Nbw-vergunning is verleend. Volgens paragraaf 5.5 van het PAS heeft door verschillende oorzaken bij nieuwvestiging van bedrijven of bij uitbreiding of wijziging van bedrijven die plaatsvonden nadat artikel 6 van de Habitatrichtlijn van toepassing werd voor een Natura 2000-gebied, niet altijd vergunningverlening op basis van een passende beoordeling plaatsgevonden. Een van de oorzaken is dat de passende beoordeling op grond van de Nbw 1998 pas op 1 oktober 2005 verplicht werd en pas op 1 februari 2009 voor alle Natura 2000-gebieden gold.

11.4. Met het PAS is beoogd te voorzien in een beoordelingskader voor het verlenen van toestemming voor bestaande en nieuwe stikstofveroorzakende activiteiten. In de passende beoordeling is onderzocht of de feitelijke depositie in 2014 en de benutting van de depositie-/ontwikkelingsruimte in de eerste PAS-periode zullen leiden tot een aantasting van de natuurlijke kenmerken van de Natura 2000-gebieden. De feitelijke depositie van alle stikstofveroorzakende activiteiten, ongeacht of daarvoor een Nbw-vergunning was verleend, is derhalve als uitgangssituatie in de passende beoordeling van het PAS betrokken.

Deze uitgangssituatie speelt op verschillende wijzen een rol bij de toestemmingverlening voor activiteiten die voor 1 januari 2015 zonder de vereiste Nbw-vergunning in bedrijf waren en dat thans ook nog zijn. Als de hoogste feitelijke depositie in de periode 1 januari 2012 - 31

december 2014 die binnen de kaders van een op 1 januari 2015 geldende milieutoestemming mocht worden veroorzaakt (hierna: de feitelijke hoogste depositie in 2012-2014) de grenswaarde van 1 mol N/ha/jr niet overschrijdt, dan zijn deze activiteiten uitgezonderd van de vergunningplicht. Een meldingsplicht geldt dan evenmin. Als de hoogste feitelijke depositie in 2012-2014 de grenswaarde van 1 mol N/ha/jr overschrijdt, dan geldt de vergunningplicht onverkort. Voor de projecten en andere handelingen die het betreft kan een vergunning worden verleend als de aanvraag ziet op een depositie die niet hoger is dan de hoogste feitelijke depositie in 2012-2014. Ontwikkelingsruimte is in dat geval niet nodig. Deze situatie is aan de orde in de zaak 201600620/1/R2 over de vergunning voor het bedrijf aan de [locatie 4] te Someren. Heeft de aanvraag betrekking op een voor 1 januari 2015 bestaande maar nog niet vergunde situatie inclusief een wijziging daarvan dan wordt voor de beoordeling of een vergunning met of zonder ontwikkelingsruimte kan worden verleend, bezien of de aangevraagde situatie leidt tot een toename van stikstofdepositie ten opzichte van de feitelijk hoogste depositie in 2012-2014. De Nbw-vergunning kan zonder ontwikkelingsruimte worden verleend als de aanvraag niet leidt tot een toename. Leidt de aanvraag tot een toename dan is ontwikkelingsruimte nodig. Dit laatste is aan de orde bij de vergunningen voor de bedrijven aan de Limburglaan 7 te Someren (zaaknr. 201600622/1/R2) en [locatie 2] te Someren (201600617/1/R2).

11.5. De Werkgroep vindt dat de emissie die veroorzaakt mocht worden op grond van een milieutoestemming die is verleend voordat artikel 6 van de Habitatrichtlijn van toepassing werd voor een Natura 2000-gebied, of een latere toestemming indien daarbij minder is vergund, het referentiepunt dient te zijn voor de beoordeling of een aanvraag leidt tot een toename van stikstofdepositie. Daarbij zou aangesloten worden bij wijze van beoordeling die in de jurisprudentie van de Afdeling is uiteengezet voordat het PAS-regime gold.

Uit de rechtspraak van de Afdeling die tot stand is gekomen voor de wijziging van de Nbw 1998 in verband met de programmatische aanpak stikstof, die onder meer is verwoord in de uitspraken van 31 maart 2010, ECLI:NL:RVS:2010:BL9656, en 13 november 2013, ECLI:NL:RVS:2013:1891, volgt dat voor de vraag of de uitbreiding of wijziging van een bestaand agrarisch bedrijf significante gevolgen kan hebben, relevant is of de wijziging of uitbreiding leidt tot een verhoging van de depositie ten opzichte van de situatie zoals die was vergund voordat artikel 6 van de Habitatrichtlijn van toepassing werd voor het Natura 2000-gebied. De vergunde situatie kan worden afgeleid uit de milieutoestemming die gold op de datum waarop artikel 6 van de Habitatrichtlijn van toepassing werd voor het Natura 2000-gebied of uit een latere milieutoestemming indien bij die latere vergunning toestemming was verleend voor een activiteit die minder depositie veroorzaakt. Wanneer uit deze beoordeling volgt dat de aangevraagde situatie, die zowel het bestaande bedrijf als de wijziging of uitbreiding daarvan omvat, niet tot een toename van depositie leidt dan is op grond van objectieve gegevens uitgesloten dat de aangevraagde situatie significante gevolgen heeft. De vergunning kon worden verleend zonder passende beoordeling. Leidt de aangevraagde situatie wel tot een toename van stikstofdepositie dan kan de vergunning alleen worden verleend als op grond van een passende beoordeling de zekerheid is verkregen dat het project de natuurlijke kenmerken van het Natura 2000-gebied niet zal aantasten.

11.6. De Afdeling overweegt dat de ratio van deze beoordelingswijze is dat alle wijzigingen van een bedrijf die hebben geleid tot een toename van stikstofdepositie ten opzichte van de situatie waarvoor toestemming was verleend voordat artikel 6 van de Habitatrichtlijn van toepassing werd voor een Natura 2000-gebied alsnog passend worden beoordeeld.

Aangezien in de passende beoordeling van het PAS de feitelijk veroorzaakte stikstofdepositie in 2014 is betrokken, is gewaarborgd dat de gevolgen van stikstofdepositie veroorzaakt door activiteiten waarvoor op basis van het PAS alsnog toestemming kan worden verleend, passend zijn beoordeeld. De omstandigheid dat deze deposities in het PAS-regime niet worden bestempeld als toenames, vloeit voort uit de gekozen systematiek om 2014 als uitgangssituatie te kiezen. De Afdeling acht dat niet in strijd met artikel 6 van de Habitatrichtlijn omdat de depositie van deze activiteiten alsnog passend is beoordeeld.

Het betoog faalt.

G. VERZOEK OM VOORRANG

12. De maatschappelijke en sociaal-economische gevolgen van de verwijzing van deze zaken naar het Hof van Justitie zijn groot. De uitvoering van het Programma Aanpak Stikstof 2015-2021 en de daarbij behorende wet- en regelgeving is relevant voor zowel de Natura 2000-gebieden die zijn opgenomen in het programma als voor alle activiteiten in Nederland die stikstof uitstoten. De beantwoording van de prejudiciële vragen heeft daardoor grote gevolgen voor de Natura 2000-gebieden en de economie in Nederland.

12.1. Het belang voor de Natura 2000-gebieden betreft de uitvoering van de maatregelen gericht op daling van de stikstofdepositie en de herstelmaatregelen die specifiek in het kader van de uitvoering van het Programma Aanpak Stikstof 2015-2021 worden genomen.

12.2. Het belang voor de economie betreft het toestemmingsregime voor activiteiten die stikstofdepositie veroorzaken op Natura 2000-gebieden. In de periode vanaf de inwerkingtreding van het PAS op 1 juli 2015 tot 31 december 2016 zijn 3103 meldingen gedaan en 4299 vergunningen aangevraagd voor activiteiten die stikstofdepositie veroorzaken. Daarbij kan gedacht worden aan de realisering van woningbouwlocaties, de aanleg van wegen, uitbreiding van industriële activiteiten en ontwikkelingen in de veehouderij. Omdat de depositie ver van de bron neerslaat en de Natura 2000-gebieden verspreid over Nederland liggen, verkeren initiatiefnemers van dergelijke projecten in heel Nederland thans in onzekerheid of een vergunning voor hun project kan worden verleend en of die, indien daartegen beroep wordt ingesteld, onherroepelijk zal worden. De onzekerheid over het onherroepelijk worden van de toestemming bemoeilijkt de financiering van deze projecten en kan tot stagnatie van de uitvoering leiden. De Afdeling acht van belang dat de onzekerheid dat deze economische ontwikkelingen doorgang kunnen vinden zo kort mogelijk duurt. Inmiddels heeft de Afdeling ongeveer 200 zaken in behandeling die vergelijkbaar zijn met de zaken die zijn behandeld in de verwijzingsuitspraken. De behandeling van deze zaken moet worden aangehouden in afwachting van de beantwoording van de prejudiciële vragen. Daarnaast is een nog onbekend aantal vergelijkbare procedures aanhangig gemaakt bij de rechtbanken in de elf arrondissementen in Nederland.

13. Vanwege de grote gevolgen die het programma heeft voor de natuur en de economie in Nederland en het grote aantal vergunningprocedures, meldingen en aangehouden en aan te houden gerechtelijke procedures in verband met de prejudiciële verwijzing, heeft de Afdeling overwogen het Hof van Justitie te verzoeken om toepassing van de versnelde procedure (PPA). De complexiteit en omvang van de verwijzingsuitspraken leent zich daar echter niet voor. De Afdeling verzoekt daarom de president om op grond van artikel 53, derde lid, van

het Reglement voor de procesvoering te beslissen de zaken bij voorrang te behandelen, zo mogelijk voor 1 juli 2018. Zij wijst daarbij op het volgende. Het programma heeft een looptijd van zes jaar. Een deel van de depositieruimte voor nieuwe economische activiteiten mag in de eerste helft van de looptijd van het programma worden toebedeeld en een deel van de depositieruimte is gereserveerd voor toedeling in de tweede helft van de looptijd. De tweede helft van de looptijd van het programma vangt aan op 1 juli 2018. De behandeling van de zaken voor 1 juli 2018 zou ertoe bijdragen dat duidelijkheid bestaat over de toepassing van het programma voor toestemmingverlening voordat met de toedeling van de depositieruimte die gereserveerd is voor de tweede helft van de looptijd van het programma wordt aangevangen.

13.1. Een verzoek om behandeling bij voorrang is tevens gedaan in de met de onderhavige zaak samenhangende verwijzingsuitspraak van heden, ECLI:NL:RVS:2017:1260.

H. KEUZES, GEGEVENS EN AANNAMES IN HET PAS

(slechts relevant voor de nationale procedure)

14. Hoewel uit het voorgaande volgt dat de Afdeling aan het Hof van Justitie een aantal prejudiciële vragen moet stellen en de beantwoording hiervan mede bepalend zal zijn voor het antwoord op de vraag of de bestreden vergunningen op grond van het PAS en de daaraan ten grondslag liggende passende beoordeling mochten worden verleend, behandelt de Afdeling hieronder een aantal beroepsgronden van de Werkgroep, die geen onlosmakelijk verband houden met de strekking van de prejudiciële vragen. Daarvoor ziet de Afdeling aanleiding omdat als het Hof van Justitie het PAS-beoordelingskader en de daarbij behorende passende beoordeling verenigbaar acht met artikel 6 van de Habitatrichtlijn, daarmee nog niet gegeven is dat hieraan geen gebreken kleven die moeten worden hersteld. Daarbij wijst de Afdeling er op dat het voorzorgsbeginsel, dat ligt besloten in artikel 6, derde lid, van de Habitatrichtlijn, met zich brengt dat een passende beoordeling, ook als deze geen betrekking heeft op afzonderlijke projecten, maar op een bij wet voorgeschreven programma waarin een bepaalde hoeveelheid stikstofdepositie is beoordeeld, geen redelijke twijfel mag laten bestaan over de vraag of de natuurlijke kenmerken van Natura 2000-gebieden zullen worden aangetast of niet.

14.1. Bij de bestreden besluiten is vergunning verleend voor de exploitatie en/of uitbreiding van agrarische bedrijven, waarbij voor de beoordeling van de stikstofdepositie die wordt veroorzaakt door die bedrijven op stikstofgevoelige habitats in Natura 2000-gebieden, toepassing is gegeven aan het PAS en de daarbij behorende regelgeving. Het algemeen deel van het PAS vormt samen met de gebiedsanalyses van de desbetreffende Natura 2000-gebieden de passende beoordeling die ten grondslag is gelegd aan de vergunningen. Deze passende beoordeling moet derhalve de zekerheid bieden dat de natuurlijke kenmerken van de gebieden niet worden aangetast. In de hierna volgende overwegingen 14.2 tot en met 27 staat niet de vraag centraal of het PAS en de bijbehorende regelgeving onverbindend zijn wegens strijd met hoger recht, maar of het algemeen deel van het PAS met de gebiedsanalyses als passende beoordeling aan de bestreden besluiten ten grondslag mocht worden gelegd.

14.2. Aan het PAS en de bijbehorende passende beoordeling zijn door de ministers van Economische Zaken en van Infrastructuur en Milieu en de staatssecretaris van Economische

Zaken diverse bestuurlijke keuzes en een grote hoeveelheid onderzoeken met gegevens en aannames ten grondslag gelegd. Zo werkt het PAS bij de maatregelen, die de extra daling van de stikstofdepositie moeten bewerkstelligen, en bij de beschikbare depositieruimte voor activiteiten die stikstofdepositie zullen veroorzaken met verschillende marges om onzekerheden in de toekomst op te kunnen vangen. Dat bij een programma met een looptijd van zes jaar wordt gewerkt met verschillende aannames, buffers en marges en wordt uitgegaan van een bepaalde mate van onzekerheid is naar het oordeel van de Afdeling onvermijdelijk. Bij een kleine marge of buffer zal de zekerheid over het effect van de maatregel groter moeten zijn dan bij een grote marge of buffer. De onzekerheid mag echter niet zo groot zijn dat, gelet op het voorzorgsbeginsel, dat mede ten grondslag ligt aan de Habitatrichtlijn, de vraag of de natuurlijke kenmerken van de gebieden zullen worden aangetast niet meer kan worden beantwoord. Derhalve is vereist dat inzicht bestaat in de keuzes, gegevens en aannames die ten grondslag liggen aan het PAS en bijbehorende passende beoordeling. De Afdeling wijst er daarbij op dat naar haar oordeel de mogelijkheden tot monitoring van de gevolgen van het PAS en de mogelijke bijsturing niet mogen afdoen aan de verplichting om voorafgaand zekerheid te hebben over de vraag of de natuurlijke kenmerken van de gebieden kunnen worden aangetast. Monitoring en bijsturing hebben in zoverre slechts een aanvullende en controlerende functie.

14.3. Het PAS beoogt door de combinatie van bron- en herstelmaatregelen en de uitgifte van depositie- en ontwikkelingsruimte zowel economische ontwikkelingen als een aanpak van de bestaande stikstofproblematiek mogelijk te maken. Voordeel van het PAS is dat de gevolgen van de maatregelen en de totale depositie van de economische ontwikkelingen integraal en in samenhang zijn beoordeeld. Daarnaast maakt het verplichte gebruik van AERIUS, zoals beschreven in 6.12-6.17, deels geautomatiseerde en efficiënte besluitvorming mogelijk. Het PAS, de bijbehorende passende beoordeling en AERIUS brengen echter ook het risico met zich dat de deels geautomatiseerde besluitvorming op grond hiervan niet inzichtelijk en controleerbaar is vanwege een gebrek aan inzicht in de gemaakte keuzes en de gebruikte gegevens en aannames. Indien belanghebbenden rechtsmiddelen willen aanwenden tegen op het PAS gebaseerde besluiten kan daardoor een ongelijkwaardige procespositie van partijen ontstaan. Zij kunnen in geval van besluitvorming op basis van een programma dat vanuit hun perspectief is te beschouwen als een zogenoemde "black box" immers niet controleren op basis waarvan tot een bepaald besluit wordt gekomen en of de zekerheid bestaat dat het project of andere handeling de natuurlijke kenmerken van Natura 2000-gebieden niet zal aantasten.

14.4. Ter voorkoming van deze ongelijkwaardige procespositie rust in dit geval op genoemde ministers en de staatssecretaris de verplichting om de gemaakte keuzes en de gebruikte gegevens en aannames volledig, tijdig en uit eigen beweging openbaar te maken op een passende wijze zodat deze keuzes, gegevens en aannames voor derden toegankelijk zijn. Deze volledige, tijdige en adequate beschikbaarstelling moet het mogelijk maken de gemaakte keuzes en de gebruikte gegevens en aannames te beoordelen of te laten beoordelen en zo nodig gemotiveerd te betwisten, zodat reële rechtsbescherming tegen besluiten die op deze keuzes, gegevens en aannames zijn gebaseerd mogelijk is, waarbij de rechter aan de hand hiervan in staat is de rechtmatigheid van deze besluiten te toetsen.

14.5. In onderstaande overwegingen beoordeelt de Afdeling of de aan de bestreden onderdelen van het PAS en de bijbehorende passende beoordeling ten grondslag gelegde keuzes, gegevens en aannames tijdig toegankelijk waren en voldoende onderbouwing bieden

voor de conclusie dat de verlening van de bestreden vergunningen op basis van het PAS de natuurlijke kenmerken van de gebieden niet zal aantasten. Zo worden genoemde ministers en de staatssecretaris in de gelegenheid gesteld mogelijke gebreken te herstellen.

De Afdeling behandelt eerst de beroepsgronden over de autonome daling van stikstofdepositie en het uitgangspunt van een gemiddelde economische groei van 2,5% als buffer bij deze daling. Vervolgens behandelt de Afdeling de beroepsgronden over de daling van de stikstofdepositie als gevolg van PAS-bronmaatregelen. Dan behandelt de Afdeling de beroepsgronden over het systeem van monitoring en bijsturing in het PAS. Daarna behandelt de Afdeling de beroepsgronden met betrekking tot de depositieruimte voor autonome ontwikkelingen. Dit zijn achtereenvolgens de beroepsgronden over de omvang gerelateerd aan economische groei van 2,5 %, de drempelwaarde, de emissieruimte die ontstaat door 'stoppers' en relatie met extern salderen, en onbenutte emissieruimte in bestaande vergunningen. Dit onderdeel wordt afgesloten met een conclusie.

Keuzes, gegevens en aannames over de depositiedaling

De autonome daling van de stikstofdepositie

15. In het PAS staat dat de stikstofdepositie in Nederland de afgelopen decennia fors is gedaald, mede dankzij nationaal en Europees beleid en dat het de verwachting is dat met het vaststaande beleid (zonder PAS) de stikstofdepositie de komende jaren gestaag verder zal dalen, zelfs wanneer wordt uitgegaan van een economische groei van 2,5% (hoog economisch groeiscenario). Dit komt doordat de stikstofemitterende sectoren, zoals het verkeer en de landbouw, steeds schoner worden. Uit analyses van het Rijksinstituut voor Volksgezondheid en Milieu (hierna: RIVM) en het Planbureau voor de Leefomgeving (hierna: het PBL) volgt dat de verwachte autonome ontwikkeling, rekening houdend met een gemiddelde economische groei van 2,5% van 2013 tot en met 2020 en van 2,2% per jaar van 2021 tot en met 2030 zal dalen van 1830 mol/ha/jr in 2012, naar 1585 mol/ha/jr in 2020 en 1515 mol/ha/jr in 2030. Dit betekent een totale daling van 315 mol/ha/jr.

In het PAS zijn maatregelen voorzien die zullen leiden tot een extra daling van depositie ten opzichte van de autonome daling. De verwachte depositieontwikkeling inclusief de PAS-bronmaatregelen is 1830 mol/ha/jr in 2012, 1535 mol/ha/jr in 2020 en 1445 mol/ha/jr in 2030. Het effect van de PAS-bronmaatregelen is derhalve een daling van 50 respectievelijk 75 mol/ha/jr in 2020 en 2030.

15.1. De Werkgroep wijst erop dat de voorspelde daling van de stikstofdepositie in de komende jaren deels is gebaseerd op de aanname dat de stikstofdepositie door bestaand beleid zal dalen. Deze depositieafname is modelmatig berekend. De berekeningen wijken volgens de Werkgroep al geruime tijd af van de feitelijk gemeten ontwikkeling in de ammoniakconcentratie in de lucht. Tussen 2000 en 2012 is een stikstofemissiedaling berekend, terwijl de gemeten ammoniakconcentratie in de lucht sinds 2000 niet meer daalt. Dat is volgens de Werkgroep een indicatie dat ook de depositie niet meer afneemt. De berekeningen die aan het PAS ten grondslag zijn gelegd zijn daarom volgens de Werkgroep onbetrouwbaar, althans niet wetenschappelijk gefundeerd. De Werkgroep verwijst naar de website van het Compendium van de Leefomgeving waaruit volgt dat de stikstofdepositie tussen 2005 en 2014 nauwelijks is gedaald. Voorts verwijst de Werkgroep naar het rapport "Balans van de Leefomgeving 2016" van het PBL (hierna: PBL-rapport 2016). De Werkgroep

stelt dat wanneer de stikstofreductie door bestaand beleid tegenvalt de stikstofdepositie hoger uitvalt. Zij betwijfelt of tegenvallers kunnen worden opgevangen binnen de buffer die volgens het PAS is ingebouwd door uit te gaan van een scenario van 2,5% economische groei. De aanname over de 2,5% groei is niet wetenschappelijk gefundeerd, maar gebaseerd op modelmatige berekeningen.

15.2. Het college stelt dat de aannames over de autonome daling van de stikstofdepositie zijn gebaseerd op ramingen van het RIVM en het PBL waarin per sector het effect van de bestaande en voorgenomen beleidsmaatregelen is opgenomen. Het college erkent dat de metingen van stikstofconcentraties laten zien dat de daling afvlakt, maar stelt dat hiermee rekening is gehouden in de berekening van de autonome daling. Volgens het college daalt de concentratie stikstofdioxide in de lucht nog wel en wordt de concentratie ammoniak in de lucht in AERIUS Monitor heel terughoudend berekend. Zo wordt in de sector landbouw bij sommige diersoorten een krimp in aantallen verwacht, maar is er in de berekening van uitgegaan dat de aantallen van deze soorten gelijk blijven. Verder stelt het college dat er rekening mee is gehouden dat de autonome daling per locatie verschilt.

Voorts stelt het college dat 2,5% economische groei een bovenraming betreft waardoor een buffer is ingebouwd om eventuele tegenvallers in de autonome depositiedaling op te kunnen vangen.

15.3. De Afdeling stelt vast dat de in het PAS verwachte daling van de stikstofdepositie tussen 2012 en 2030 in hoofdzaak wordt bepaald door de autonome ontwikkeling. Van de totale verwachte afname van 385 mol/ha/jr is tussen 2012 en 2020 het effect van de PAS-bronmaatregelen: 50 mol/ha/jr. In de periode tussen 2020 en 2030 is dat 75 mol/ha/jaar. Het verschil tussen de totale afname en de afname door de PAS-bronmaatregelen betreft de autonome daling. Gelet hierop en in aanmerking genomen dat de autonome daling van belang is voor de aannames over het herstel of de verbetering van stikstofgevoelige natuurwaarden en dat een deel van de autonome daling van de stikstofdepositie beschikbaar wordt gesteld als depositie- en ontwikkelingsruimte, dient een goed inzicht te bestaan in de gegevens en aannames die aan de verwachte daling van de stikstofdepositie ten grondslag liggen, en worden hoge eisen gesteld aan de zekerheid die deze gegevens bieden inzake de gevolgen voor de natuurlijke kenmerken van de Natura 2000-gebieden. De Afdeling zal dit hierna beoordelen en daarbij de vraag betrekken of de gemaakte keuzes en de gebruikte gegevens en aannames volledig, tijdig en uit eigen beweging openbaar zijn gemaakt op een passende wijze zodat deze keuzes, gegevens en aannames voor derden toegankelijk zijn.

15.4. Het college heeft toegelicht dat in het rapport "Grootschalige concentratie- en depositiekaarten" uit 2015 van het RIVM (hierna: het RIVM-rapport 2015) staat beschreven op welk beleid de prognose van de toekomstige stikstofemissies is gebaseerd. In het RIVM-rapport 2015 wordt voor een gedetailleerde beschrijving van het vaststaande beleid dat is meegenomen in de prognose verwezen naar het rapport "Referentieraming energie en emissies: actualisatie 2012. Energie en emissies in de jaren 2012, 2020 en 2030" uit 2012 van het PBL en Energieonderzoek Centrum Nederland (hierna: het PBL-rapport 2012). Het PBL-rapport 2012 bevat een raming van het Nederlandse verbruik van energie en de uitstoot van broeikasgassen en luchtverontreinigende stoffen. Hierbij is rekening gehouden met nationaal en Europees beleid dat invloed heeft op deze uitstoot. In bijlage 1 van het PBL-rapport 2012 is een overzicht opgenomen van dit nationale en Europese beleid.

15.5. Het RIVM levert jaarlijks kaarten met grootschalige concentraties voor Nederland van de luchtverontreinigende stoffen waarvoor Europese luchtkwaliteitsnormen bestaan. Deze kaarten geven een beeld van de luchtkwaliteit in Nederland, zowel van het verleden als voor de toekomst. Het RIVM levert ook kaarten met de grootschalige depositie voor Nederland (GDN-kaarten genoemd) van stikstof. De stikstofdepositiekaarten worden gebruikt bij het PAS, zo staat ook in het RIVM-rapport 2015.

15.6. In de passende beoordeling staat dat in lijn met de dalende trend van de stikstofdepositie in de afgelopen decennia een verdere daling van de gemiddelde stikstofdepositie wordt verwacht in de periode tot en met 2030. Naar het oordeel van de Afdeling mag bij de prognose over de autonome ontwikkeling van de stikstofdepositie in beginsel worden uitgegaan van een bestaande dalende trend, zoals voorspeld in het RIVM-rapport 2015, tenzij er contra-indicaties zijn.

In het PBL-rapport 2016 staat dat de emissie van ammoniak sinds 1990 met bijna 70% is verminderd, maar dat de ammoniakconcentraties in de lucht sinds 2000 niet meer dalen en zelfs een stijgende trend lijken te vertonen. In het PBL-rapport 2016 wordt een aantal mogelijke oorzaken genoemd voor het verschijnsel dat de ammoniakemissies dalen maar de ammoniakconcentraties sinds 2000 niet. Volgens het PBL zal nader onderzoek uitsluitsel moeten geven over de oorzaak van het trendverschil. Dit heeft het college ter zitting ook erkend.

Gelet op het voorgaande zijn er contra-indicaties die erop wijzen dat niet zonder meer kan worden uitgegaan van het doorzetten van een bestaande dalende trend van de stikstofdepositie. Het college heeft weliswaar gesteld dat hiermee rekening is gehouden, maar het PAS of de daaraan ten grondslag gelegde stukken bieden geen inzicht op welke wijze bij de berekening van de autonome daling van de stikstofdepositie rekening is gehouden met het gegeven dat de ammoniakconcentratie in de lucht al langere tijd niet daalt. Dit inzicht is nodig om te kunnen beoordelen of de voorspelde daling van de stikstofdepositie is gebaseerd op een realistische prognose. Mede gelet op het grote aandeel dat de autonome daling heeft in de totale in het PAS verwachte depositiedaling dient naar het oordeel van de Afdeling te worden onderbouwd op welke wijze rekening is gehouden met de stagnatie van de ammoniakconcentratie in de lucht dan wel dient te worden onderbouwd waarom aannemelijk is dat de bestaande dalende trend van de stikstofdepositie doorzet.

Gelet op het voorgaande is de onderbouwing van de verwachte daling van de stikstofdepositie, gelet op de eisen die in 14-14.4 zijn geformuleerd, onvoldoende inzichtelijk. Het betoog slaagt.

Economische groei

16. In paragraaf 4.2.1 van het PAS staat dat het programma is gebaseerd op een scenario waarin gerekend wordt met een economische groei van 2,5% per jaar. Voor dit scenario is gekozen om maximaal ruimte te kunnen bieden aan (nieuwe) economische ontwikkelingen en als extra buffer voor onzekerheden in de autonome ontwikkeling van de stikstofdepositie. Ook met dit scenario als uitgangspunt neemt de depositie in Nederland door vaststaand beleid nog steeds af, aldus het PAS.

16.1. In het RIVM-rapport 2015 staat dat de ontwikkeling van emissies afhankelijk is van

nationaal en Europees beleid, autonome maatschappelijke en economische ontwikkelingen. Voor de aannames over de autonome daling is dan ook van belang wat de verwachte economische groei is en tot welke emissies deze leidt.

Voorts staat in het RIVM-rapport 2015 dat in alle ramingen van het RIVM wordt uitgegaan van de werkelijke economische groei in Nederland in 2009 (3,5 procent), 2010 (+1,7 procent) en 2011 (+1,2 procent) zoals gerapporteerd door het Centraal Planbureau (hierna: het CPB) in 2012. Voor de jaren 2012 en 2013 is de verwachte groei gebaseerd op het Centraal Economisch Plan 2012 van het CPB. Voor 2012 werd een krimp verwacht van 0,75 procent en voor 2013 een groei van 1,25 procent. In de referentieraming zit verder een economische groei van +1,5 procent per jaar in 2014-2015 en +1,9 procent per jaar voor de periode 2016-2020. Deze economische ontwikkeling is tot 2020 vergelijkbaar met de gemiddelde groei van 1,7 procent per jaar zoals die in 2014 in de scenario's is gebruikt. In de referentieraming is verder uitgegaan van een bandbreedte in economische groei van 0,75 procentpunt van 2013 tot 2020. Dit geeft een groei van ongeveer 0,9 procent per jaar voor de onderraming, ongeveer 1,7 procent voor de middenraming en ongeveer 2,5 procent per jaar voor de bovenraming. De gemiddelde economische groei is vertaald naar groeicijfers per sector, waarbij rekening is gehouden met sectorspecifieke ontwikkelingen en sectorspecifiek beleid.

Gelet op de hiervoor weergegeven toelichting in het RIVM-rapport 2015 dat 2,5% economische groei een bovenraming is, ziet de Afdeling in het niet nader onderbouwde betoog van de Werkgroep, geen aanleiding voor het oordeel dat dit uitgangspunt mogelijk een onderschatting van de economische groei betreft.

Over het standpunt van het college dat met 2,5% economische groei een buffer is ingebouwd om eventuele tegenvallers in de autonome daling op te kunnen vangen is, overweegt de Afdeling als volgt. Het PAS vermeldt niet welke daadwerkelijke economische groei wordt aangenomen, als gevolg waarvan niet duidelijk is ten opzichte van welk percentage het uitgangspunt van 2,5% economische groei een buffer vormt. Daarmee is evenmin inzichtelijk hoeveel kiloton aan emissies de beoogde buffer groot is. In zoverre is de onderbouwing van de verwachte economische groei van 2,5% als buffer voor tegenvallers in de autonome daling, gelet op de eisen die in 14-14.4 zijn geformuleerd, onvoldoende inzichtelijk.

Het betoog slaagt.

De depositiedaling door de PAS-bronmaatregelen

17. Zoals beschreven in 6.2 worden in het kader van het PAS extra bronmaatregelen getroffen om de stikstofdepositie te laten dalen en daarmee bij te dragen aan de doelen van het programma. Dit zijn de zogeheten 'generieke bronmaatregelen' of 'PAS-bronmaatregelen'. Een deskundige van het Centraal bureau voor de Statistiek heeft een berekening gemaakt van de totale te verwachten gevolgen van deze bronmaatregelen. De resultaten hiervan zijn neergelegd in het document: "Doorrekening maatregelen gericht op beperking ammoniakemissie" van 21 januari 2014 (hierna: de CBS-berekening). Berekend is dat de ammoniakemissie in 2020 met 13,4 kiloton per jaar zal zijn gedaald als gevolg van de PAS-bronmaatregelen ten opzichte van de jaarlijkse emissie zonder het nemen van deze maatregelen. Het programma houdt rekening met een totale daling van de ammoniakemissie met 6,4 kiloton per jaar in 2021. Ter zitting is toegelicht dat met de marge tussen 13,4 en 6,4

kiloton onzekerheden kunnen worden opgevangen.

De helft van de daling van 6,4 kiloton wordt in het programma als ontwikkelingsruimte beschikbaar gesteld. In de passende beoordeling is geconcludeerd dat de uitgifte van deze ontwikkelingsruimte nergens zal leiden tot een toename van de stikstofdepositie en niet kan leiden tot een aantasting van de natuurlijke kenmerken van de gebieden die in het programma zijn betrokken.

17.1. Er zijn drie soorten PAS-bronmaatregelen in het programma opgenomen. Ten eerste is een algemene maatregel van bestuur vastgesteld, het Besluit Emissiearme Huisvesting, waarmee strengere eisen aan de emissies van stallen worden gesteld. Ten tweede zijn en worden er verplichtingen omtrent de aanwending van dierlijke mest ingevoerd. Ten derde zijn afspraken gemaakt over voer- en managementmaatregelen in de veehouderij.

17.2. De Werkgroep heeft aangevoerd dat deze PAS-bronmaatregelen ten onrechte of ten onrechte geheel zijn betrokken bij de beoordeling van het programma. Zij stelt in dit verband dat onvoldoende is verzekerd dat de maatregelen het gestelde effect zullen sorteren, zodat niet kan worden uitgesloten dat stikstofgevoelige habitats in de Peelgebieden te kampen krijgen met een toename van stikstofdepositie wanneer een deel van de berekende daling van de uitstoot in de vorm van ontwikkelingsruimte wordt uitgegeven.

17.3. De Afdeling zal de PAS-bronmaatregelen bespreken in het licht van de beroepsgronden en hetgeen hiervoor onder 14-14.4 is overwogen. Dit betekent dat de verschillende aannames, buffers en marges die aan de beoordeling van de gevolgen van de PAS-bronmaatregelen ten grondslag zijn gelegd zullen worden besproken in het licht van de vraag of de onzekerheid niet zo groot is dat gelet op het voorzorgsbeginsel de vraag of de natuurlijke kenmerken van de gebieden zullen worden aangetast niet meer kan worden beantwoord. Daarbij wordt de vraag betrokken of de gemaakte keuzes en de gebruikte gegevens en aannames volledig, tijdig en uit eigen beweging openbaar zijn gemaakt op een passende wijze zodat deze keuzes, gegevens en aannames voor derden toegankelijk zijn

De Afdeling zal eerst de onderbouwing van de gevolgen van de afzonderlijke maatregelen beoordelen. Omdat het PAS niet alleen uitgaat van de gevolgen van de individuele bronmaatregelen, maar ook de gevolgen van de drie bronmaatregelen gezamenlijk beschouwt, zal de Afdeling vervolgens de onderbouwing van de gezamenlijke gevolgen beoordelen.

De stalmaatregelen in het Besluit Emissiearme Huisvesting (hierna: het BEH)

18. Op 1 augustus 2015 is het BEH in werking getreden. In dit besluit is geregeld dat degene die een inrichting drijft waarin landbouwhuisdieren worden gehouden geen huisvestingssystemen toepast met een emissiefactor voor ammoniak die hoger is dan de maximale emissiewaarde voor ammoniak die zijn vermeld in bijlage 1 bij het BEH. Deze maximale emissiewaarden zijn grotendeels lager dan de emissiewaarden die voor 1 augustus 2015 golden. Hoewel op grond van het overgangsrecht dat in het BEH is opgenomen veel bestaande stallen nog niet hoeven te voldoen aan de verscherpte eisen, is ten behoeve van het PAS rekening gehouden met een ontwikkeling dat in de komende jaren de ammoniakemissie van veehouderijen zal gaan afnemen doordat stalsystemen aan de eisen gaan voldoen, met name als gevolg van bedrijfsvernieuwingen en -ontwikkelingen.

In de voornoemde berekening van het Centraal bureau voor de Statistiek is aan de hand van bekende gegevens over bestaande stalsystemen en ervan uitgaande dat op termijn in 2045 alle stalsystemen zullen zijn aangepast aan het BEH een prognose van de daling van ammoniakemissie gemaakt. Hierin staat dat de totale ammoniakemissie in 2020 zal zijn afgenomen met 2,1 kiloton per jaar als gevolg van deze maatregel. In het PAS is rekening gehouden met een kleinere afname als gevolg van deze maatregel, namelijk dat in het jaar 2021 de totale ammoniakemissie met 1,4 kiloton per jaar zal zijn afgenomen.

De gegevens zijn verwerkt in de weergave van de stikstofdepositie in AERIUS Monitor, waarbij per Natura 2000-gebied de bijdrage van de stalemissies in de totale stikstofdepositie is bepaald. Volgens deze weergave zal in de Groote Peel in 2020 de stikstofdepositie als gevolg van stalemissies met 22 mol/ha/jr zijn afgenomen en in de Deurnsche Peel & Mariapeel met 24 mol/ha/jr.

18.1. Met betrekking tot de vraag of het treffen van deze maatregel als zodanig voldoende is verzekerd, overweegt de Afdeling dat het BEH in werking is getreden.

18.2. Met betrekking tot de vraag of in voldoende mate is verzekerd dat in 2021 deze maatregel zal leiden tot een daling van 1,4 kiloton ammoniakemissie per jaar en dat in combinatie met het uitgeven van ontwikkelingsruimte voor een deel van deze daling geen toename van stikstofdepositie op voorkomens van stikstofgevoelige habitats in de Peelgebieden kan plaatsvinden, heeft de Werkgroep een aantal argumenten ingebracht op grond waarvan zij stelt dat deze vraag ontkennend moet worden beantwoord.

In de eerste plaats stelt de Werkgroep dat de prognose te rooskleurig is over de landelijke ontwikkelingen, omdat uitgegaan wordt van een snellere economische afschrijving van stallen dan in de praktijk het geval is. Hierdoor zullen de strengere emissie-eisen uit het BEH minder snel effect sorteren dan waarvan in het PAS wordt uitgegaan. Daarnaast gaat volgens de Werkgroep de prognose er ten onrechte vanuit dat de maatregelen voor honderd procent werkzaam zullen zijn. In dit verband wijst zij op het rapport "resultaten Brabantbrede toezichtsaanpak luchtwassers 2011-2012" uit april 2013, waaruit volgt dat 59% van de bedrijven in overtreding was bij het gebruik van luchtwassers die bedoeld zijn voor het terugdringen van ammoniakuitstoot.

18.3. Het college stelt zich op het standpunt dat met deze omstandigheden voldoende rekening is gehouden. Ten eerste is een marge aangehouden van 0,7 kiloton tussen de prognose van de emissiedaling en de daling waarmee in het programma rekening is gehouden. Ten tweede wijst het college op nieuwe regelgeving die het mogelijk maakt om de werking van luchtwassers elektronisch te controleren, zodat grootschalige overtreding niet langer kan plaatsvinden.

18.4. Over de vraag of bij de beoordeling of als gevolg van de verlening van de vergunningen de natuurlijke kenmerken van de Peelgebieden kunnen worden aangetast, rekening mocht worden gehouden met een landelijke daling van 1,4 kiloton ammoniakemissie per jaar als gevolg van het BEH overweegt de Afdeling als volgt.

Ten aanzien van de afschrijving van stallen heeft de Werkgroep met redenen omkleed aangegeven dat de levensduur van bestaande stallen - en de daarmee gepaard gaande

hogere ammoniakemissie dan het BEH voorschrijft - langer is dan waarvan in de berekeningen is uitgegaan. Daartegenover heeft het college slechts gewezen op de marge die is aangehouden tussen de prognose van de daling en de daling die in het programma is opgenomen. Noch uit de stukken, noch uit het ter zitting verhandelde is echter inzichtelijk geworden welke aannames en gegevens zijn gehanteerd om het standpunt dat de aangehouden marge voldoende is om een eventuele tegenvallende prognose op te vangen te onderbouwen.

Over het argument dat is gebleken dat enkele jaren geleden in Noord-Brabant meer dan de helft van de stalsystemen niet werkte op een wijze waar de regelgeving over ammoniakemissie en de daarop gebaseerde berekeningen van deze emissie vanuit gaan, overweegt de Afdeling dat een grootschalig handhavingstekort erop zou kunnen wijzen dat de feitelijke prognoses van stikstofdepositie onjuist zijn. De Afdeling ziet in dit geval echter, gelet op de regelgeving over controle die in het Activiteitenbesluit is opgenomen, geen aanleiding voor het oordeel dat het college om deze reden niet uit had mogen gaan van de prognose van de ammoniakdaling als gevolg van het BEH.

18.5. In de tweede plaats stelt de Werkgroep dat ook indien de prognose landelijk bezien juist is, hiermee niet is uitgesloten dat bij uitgifte van ontwikkelingsruimte de depositie van stikstof op hiervoor gevoelige habitattypen in de Peelgebieden zal toenemen. Het BEH bevat namelijk weliswaar normen voor huisvestingssystemen die lagere emissiewaarden voorschrijven dan voor 1 augustus 2015 golden, maar bevat geen verplichting om op bedrijfsniveau de ammoniakemissie te laten dalen. Dit betekent dat door het verplaatsen van dieren naar bedrijven in de buurt van de Peelgebieden de concentratie van vee in de nabijheid van de Peelgebieden kan toenemen. Hierdoor kan bij gelijkblijvende ammoniakemissie op regioniveau de depositie van stikstof op de Peelgebieden toenemen. In de praktijk groeien de bedrijven in de nabijheid van de Peelgebieden, zodat dit een reëel risico is.

Daarnaast is volgens de Werkgroep een deel van de verwachte daling van de stikstofdepositie in de nabijheid van de Peelgebieden aan het BEH toegeschreven, terwijl in de provincie Noord-Brabant op grond van de "Verordening stikstof en Natura 2000 Noord-Brabant 2013" (hierna: de Verordening stikstof) veehouderijen reeds moeten voldoen aan strengere regelgeving voor stalemissies. Daarom is het aannemelijk dat in deze provincie gemiddeld minder daling van ammoniakemissie zal plaatsvinden als gevolg van de landelijke daling door het BEH dan in andere provincies, zodat ten onrechte een deel van de verwachte daling van de stikstofdepositie aan het BEH is toegeschreven. De uitgifte van ontwikkelingsruimte kan juist lokaal resulteren in een toename van de stikstofdepositie, aldus de Werkgroep.

18.6. Het college stelt zich op het standpunt dat in het PAS rekening is gehouden met de mogelijkheid dat bedrijven nabij de Peelgebieden meer dieren kunnen gaan houden. AERIUS gaat namelijk bij het bepalen van de huidige emissies en bij de prognoses uit van de feitelijk bestaande stalsystemen. Dit betekent dat AERIUS ook registreert dat dieren worden verplaatst als gevolg van bedrijfsuitbreidingen, zodat dit bij het bepalen van de depositieruimte wordt betrokken. Hierin ligt volgens het college dus geen reden om te vrezen dat de stikstofdepositie op een locatie binnen de Peelgebieden zal stijgen.

Verder stelt het college dat rekening is gehouden met de reeds bestaande strengere normen in de provincie Noord-Brabant door rekening te houden met de feitelijk bestaande

stalsystemen in AERIUS.

18.7. De vraag is nu of bij de beoordeling van de vraag of als gevolg van de verlening van de vergunningen de natuurlijke kenmerken van de Peelgebieden kunnen worden aangetast uitgegaan mocht worden van een berekende daling van stikstofdepositie op de Groote Peel van 22 mol/ha/jr en op de Deurnsche Peel & Mariapeel van 24 mol/ha/jr, en of verzekerd is dat bij uitgifte van ontwikkelingsruimte de stikstofdepositie op geen stikstofgevoelig habitat binnen deze gebieden kan toenemen. De Afdeling overweegt daarover als volgt.

Ten aanzien van het argument over de verplaatsing van dieren constateert de Afdeling dat in de stukken die bij het systeem AERIUS beschikbaar zijn gemaakt - waaronder de factsheets - niet staat beschreven hoe het systeem rekening houdt met de mogelijkheid dat regionale verschuivingen optreden door het verplaatsen van dieren naar bedrijven in de buurt van de Peelgebieden. De enkele stelling dat hiermee rekening wordt gehouden doordat met de feitelijke stalemissies wordt gerekend, kan de Afdeling in dit verband niet volgen, omdat niet is gebleken dat dit aspect is verwerkt in AERIUS. AERIUS Monitor geeft weliswaar weer wat de verwachte daling van stikstofdepositie is als gevolg van de totale stalemissies, maar niet is inzichtelijk gemaakt welk deel van deze daling het gevolg is van het BEH en welk deel andere oorzaken heeft.

Ten aanzien van het argument over de Verordening stikstof constateert de Afdeling dat in het "Programma Aanpak Stikstof 2015-2021 zoals gewijzigd na partiële herziening op 15 december 2015" op pagina 15 staat dat de provincie Noord-Brabant er niet voor heeft gekozen de Verordening deel uit te laten maken van het programma. Uit pagina 28 lijkt evenwel te volgen dat de effecten van de Verordening betrokken worden bij de effecten van de bronmaatregelen. Het is de Afdeling daarom niet duidelijk of de Verordening stikstof is betrokken in het programma. Daarnaast is het de Afdeling uit de stukken die met AERIUS beschikbaar zijn gesteld niet duidelijk geworden of en hoe in de provincie Noord-Brabant rekening is gehouden met de omstandigheid dat de bestaande feitelijke stalsystemen ten dele voldoen aan de strengere eisen van de Verordening stikstof.

18.8. Gelet op de overwegingen 18.4 en 18.7 is naar het oordeel van de Afdeling onvoldoende inzichtelijk welke gegevens en aannames ten grondslag liggen aan de stelling dat de marge die is aangehouden tussen de prognose van de landelijke daling van ammoniakemissie als gevolg van het BEH en de daling van ammoniakemissie waarvan in het PAS is uitgegaan voldoende is om onzekerheden in de prognose op te vangen. Daarnaast is onvoldoende inzichtelijk welke gegevens en aannames ten grondslag liggen aan de stelling dat lokale stijgingen van de depositie van ammoniak zijn uitgesloten in het licht van de mogelijkheid dat bestaande veehouderijen ook onder het BEH hun veestapel kunnen uitbreiden en de mogelijkheid dat in Noord-Brabant een lagere dan gemiddelde daling van de ammoniakemissie wordt veroorzaakt door het BEH, vanwege reeds bestaande strengere staleisen.

In zoverre is de onderbouwing van de (gevolgen van de) maatregel, gelet op de eisen die in 14-14.4 zijn geformuleerd onvoldoende inzichtelijk gemaakt. Het betoog slaagt.

Besluit gebruik meststoffen (hierna: het Meststoffenbesluit)

19. Door wijzigingen van het Meststoffenbesluit zijn en worden de normen aangescherpt

voor het aanwenden van dierlijke mest op landbouwgronden. Hiermee wordt beoogd om de ammoniakemissie bij het aanwenden van mest te verminderen. De wijzigingen in het Meststoffenbesluit bestaan uit twee maatregelen. Ten eerste is een verplichting ingevoerd om drijfmest op onbeteeld bouwland met een spruitstuk in - sleufjes in - de bodem te brengen. Ten tweede worden emissiearme technieken verplicht gesteld voor de aanwending van mest op klei en veen. Hierbij was voorzien dat deze maatregel op 1 januari 2017 in werking zou treden.

In voornoemde berekening van het Centraal bureau voor de Statistiek is een prognose gemaakt waarin staat dat door deze maatregel de totale ammoniakemissie zal afnemen met 2,2 kiloton per jaar in 2020. Hierbij is voor de eerste maatregel die op 1 januari 2015 in werking is getreden een emissiereductie van 0,5 kiloton ammoniak per jaar voorzien. Voor de maatregel waarvan de inwerkingtreding op 1 januari 2017 was gepland is een emissiereductie van 1,7 kiloton ammoniak per jaar voorzien. Deze getallen staan ook vermeld in de factsheet "Bepalen depositiebijdrage en groeibehoefte mest" die in AERIUS is gebruikt om op basis van de landelijke prognose een prognose te maken van de stikstofdepositie als gevolg van mestaanwending op de betrokken Natura 2000-gebieden. In 2020 wordt verwacht dat deze met 12 mol/ha/jr zal zijn gedaald op de Groote Peel en met 14 mol/ha/jr op de Deurnsche Peel & Mariapeel.

In het PAS is rekening gehouden met een kleinere afname, namelijk dat in het jaar 2021 de totale ammoniakemissie als gevolg van deze maatregelen met 2,0 kiloton per jaar zal zijn afgenomen.

19.1. Met betrekking tot de vraag of het treffen van deze maatregelen als zodanig voldoende is verzekerd, overweegt de Afdeling dat de eerstgenoemde maatregel op 1 januari 2015 in werking is getreden.

De tweede maatregel in het Meststoffenbesluit was voorzien om per 1 januari 2017 in werking te treden. Dat is evenwel niet gebeurd, omdat de technieken hiervoor nog niet beschikbaar zijn, zoals de Werkgroep ook heeft gesteld. Het college stelt dat de technieken inmiddels zijn uitontwikkeld en invoering van deze maatregel is voorzien per 1 januari 2018. Omdat in het PAS ook is gerekend met een invoering vanaf deze datum levert dat in de ogen van het college geen probleem op.

De Afdeling overweegt dat in het PAS rekening is gehouden met prognoses voor de effecten van deze maatregel terwijl onzekerheid bestaat over de vraag wanneer deze maatregel in werking kan treden. De enkele verwachting dat deze maatregel in werking zal treden vanaf 1 januari 2018 acht de Afdeling onvoldoende om te kunnen stellen dat de werking van de maatregel in zoverre is verzekerd. In zoverre slaagt het betoog.

Voor de beoordeling van de gevolgen van deze maatregel hieronder, zal de Afdeling in het kader van deze uitspraak ervan uitgaan dat de tweede maatregel, over emissiearme technieken op klei en veen, daadwerkelijk zal worden genomen.

19.2. Met betrekking tot de vraag of in voldoende mate is verzekerd dat in 2021 deze maatregel zal leiden tot een daling van 2,0 kiloton ammoniakemissie en dat in combinatie met het uitgeven van ontwikkelingsruimte voor een deel van deze daling geen toename van stikstofdepositie op stikstofgevoelige habitats in de Peelgebieden kan plaatsvinden, voert de

Werkgroep het volgende aan. Deze maatregel verhindert niet dat agrariërs het gebruik van hun gronden en daarmee ook de hoeveelheid toegediende mest wijzigen. Hierdoor kan in de nabijheid van de Peelgebieden de aanwending van mest toenemen, als gevolg waarvan een toename van de stikstofdepositie op hiervoor gevoelige locaties kan worden veroorzaakt. Ter onderbouwing van het betoog wijst zij op kaarten die behoren bij het compendium voor de leefomgeving waaruit blijkt dat er grote verschillen bestaan tussen locaties waar over- en onderbenutting van de toegestane bemesting plaatsvindt, waardoor lokaal - op plaatsen met een onderbenutting van de toegestane bemestingsgraad - een sterke toename van de bemestingsgraad kan optreden.

19.3. Het college stelt zich op het standpunt dat door lokale variaties in de aanwending van mest weliswaar kleine verschillen kunnen ontstaan in de lokale stikstofdepositie, maar dat dat niet betekent dat hierdoor substantiële problemen zullen ontstaan inzake de gevolgen van de ammoniakdepositie. Het college verwijst naar een expert-judgement van 19 februari 2016 van deskundigen van het landbouwinstituut Alterra: "Beoordeling van het hanteren van gemiddelde ammoniakemissies per bij mestaanwending in de PAS". In deze beoordeling is aan de hand van proeven bij de bemesting van grasland geconcludeerd dat extreme variaties voor de stikstofdepositie als gevolg van bemesten en beweiden kunnen leiden tot een variatie in stikstofneerslag uit deze bron binnen 1 kilometer van een Natura 2000-gebied van ongeveer 15%. Omdat de gemiddelde bijdrage van stikstofdepositie binnen Natura 2000-gebieden door veehouderijen 5-10% van het totaal bedraagt, zal deze variatie van 15% leiden tot een totale variatie van stikstofdepositie in het gebied van rond de 1%. Deze variatie heeft volgens het college weinig betekenis.

19.4. Over de vraag of bij de beoordeling of als gevolg van de verlening van de vergunningen de natuurlijke kenmerken van de Peelgebieden kunnen worden aangetast, rekening mocht worden gehouden met een daling van 2,0 kiloton ammoniakemissie als gevolg van de genomen en voorziene maatregelen in het Meststoffenbesluit overweegt de Afdeling als volgt.

De Werkgroep heeft met redenen omkleed gesteld dat gelet op de huidige variatie in benutting van de bemestingsruimte en in combinatie met gegevens over lokale onderbenutting van deze ruimte niet kan worden uitgesloten dat er locaties zijn waar de mate van bemesting (sterk) kan toenemen met een toename van stikstofdepositie op stikstofgevoelige habitats tot gevolg.

Uit de stukken die aan het PAS ten grondslag liggen blijkt echter niet dat met deze omstandigheid rekening is gehouden. Weliswaar heeft het college gewezen op de marge die is aangehouden tussen de prognose van een landelijke daling van 2,2 kiloton in 2020 en de daling van 2,0 kiloton in 2021 waarmee het PAS rekening houdt en die is verwerkt in de hiervoor vermelde prognose van de depositieafname in AERIUS Monitor, maar noch uit de stukken, noch uit het ter zitting verhandelde is inzichtelijk geworden welke aannames en gegevens zijn gehanteerd om het standpunt dat de aangehouden marge voldoende is om een eventuele tegenvallende prognose op te vangen te onderbouwen. Dit geldt zowel voor de landelijke (gemiddelde) prognoses als voor het standpunt dat als gevolg van het uitgeven van een deel van deze daling in de vorm van ontwikkelingsruimte, uitgesloten is dat op een locatie de stikstofdepositie zal toenemen.

Verder heeft het college weliswaar op een expert-judgement van 19 februari 2016 gewezen,

maar dit judgement dateert enerzijds van na het nemen van de bestreden besluiten, en anderzijds is in dit stuk een beoordeling gemaakt over een specifiek soort gebruik van landbouwgrond, namelijk grasland waarop wel of niet sprake is van beweiding. Derhalve is niet gebleken dat rekening is gehouden met de grote lokale verschillen in bemestingsgraad van ook andere landbouwgronden dan weidegronden.

In zoverre is de onderbouwing van de (gevolgen van de) maatregel, gelet op de eisen die in 14-14.4 zijn geformuleerd onvoldoende inzichtelijk gemaakt. Het betoog slaagt.

Voer- en managementmaatregelen

20. Op 18 maart 2014 is door de staatssecretaris van Economische Zaken en zeven organisaties die veehouders vertegenwoordigen (hierna: de sectororganisaties) de "Overeenkomst generieke maatregelen in verband met het Programma Aanpak Stikstof" getekend. Deze overeenkomst heeft als doel een reductie in ammoniakemissie te realiseren en bevat daartoe afspraken over vrijwillig door veehouders te treffen voer- en managementmaatregelen. In de overeenkomst zijn verplichtingen opgenomen over maatregelen die veehouderijen kunnen nemen om de ammoniakemissie te reduceren. Hierbij moet gedacht worden aan maatregelen zoals het verbeteren van mineralenefficiëntie van melkveebedrijven en het bevorderen van de weidegang.

In voornoemde berekening van het Centraal bureau voor de Statistiek is een prognose gemaakt waarin staat dat door deze maatregel de totale ammoniakemissie zal afnemen met 9,5 kiloton per jaar. In het PAS is rekening gehouden met een kleinere afname, namelijk dat in het jaar 2021 de totale ammoniakemissie als gevolg van deze maatregel met 3 kiloton per jaar zal zijn afgenomen. Deze afname is verwerkt in de prognose van AERIUS Monitor.

20.1. Met betrekking tot de vraag of het treffen van deze maatregel als zodanig voldoende is verzekerd, overweegt de Afdeling dat de overeenkomst reeds is gesloten.

20.2. Met betrekking tot de vraag of in voldoende mate is verzekerd dat in 2021 deze maatregel zal leiden tot een daling van 3,0 kiloton ammoniakemissie en dat in combinatie met het uitgeven van ontwikkelingsruimte voor een deel van deze daling geen toename van stikstofdepositie op voorkomens van stikstofgevoelige habitats in de Peelgebieden kan plaatsvinden, voert de Werkgroep aan dat de overeenkomst niet afdwingbaar is, zodat geen zekerheid bestaat dat de maatregel ook het effect zal sorteren dat in de prognose is opgenomen. Zij wijst in dit verband op artikel 12 van de overeenkomst waarin de niet-afdwingbaarheid is opgenomen.

20.3. Het college stelt dat de maatregelen in de overeenkomst weliswaar niet direct afdwingbaar zijn, maar dat het systeem van het PAS een prikkel bevat om deze maatregelen uit te voeren. Als deze maatregelen namelijk niet het voorziene effect sorteren - met behulp van het monitoringsprogramma in het PAS wordt dit geregistreerd - dan gaat dit ten koste van de ontwikkelingsruimte. Daarnaast wijst het college op de marge tussen de prognose van de effecten van deze maatregelen en de daling van de ammoniakemissie waarmee in het PAS rekening is gehouden. Deze acht het college voldoende om onzekerheden in de prognose en eventuele gebreken in de naleving op te vangen.

20.4. Over de vraag of bij de beoordeling of als gevolg van de verlening van de vergunningen

de natuurlijke kenmerken van de Peelgebieden kunnen worden aangetast, rekening mocht worden gehouden met een daling van 3,0 kiloton ammoniakemissie als gevolg van de maatregelen uit de overeenkomst overweegt de Afdeling als volgt.

Enerzijds stelt de Werkgroep terecht dat de overeenkomst niet afdwingbaar is. Niet alleen staat in artikel 12 van de overeenkomst: "Deze overeenkomst is niet in rechte afdwingbaar", maar ook kan de nakoming van een overeenkomst tussen twee partijen in beginsel niet door derden worden afgedwongen en bestaat er geen publiekrechtelijke regeling op grond waarvan nakoming van deze overeenkomst kan worden afgedwongen.

De afdwingbaarheid als zodanig is echter thans niet aan de orde. Het gaat om de vraag of in het PAS rekening mocht worden gehouden met een daling van 3,0 kiloton ammoniakemissie. Hierbij kan weliswaar relevant zijn dat de maatregel niet afdwingbaar is, maar dit is niet doorslaggevend, omdat met het ontbreken van de afdwingbaarheid in de prognose van de daling rekening kan worden gehouden.

Ten aanzien van de vraag of de marge die is aangehouden tussen de prognose van een landelijke daling van 9,5 kiloton ammoniakemissie in 2020 en de daling van 3,0 kiloton in 2021 waarmee het PAS rekening houdt, de vereiste zekerheid geeft, oordeelt de Afdeling dat noch uit de stukken, noch uit het ter zitting verhandelde inzichtelijk is geworden welke aannames en gegevens zijn gehanteerd om het standpunt dat de aangehouden marge voldoende is om een eventuele tegenvallende prognose op te vangen, te onderbouwen. Niet inzichtelijk is wat precies de bijdrage is van deze maatregel aan de daling van stikstofdepositie die AERIUS Monitor weergeeft.

In zoverre is de onderbouwing van de (gevolgen van de) maatregel, gelet op de eisen die in 14-14.4 zijn geformuleerd onvoldoende inzichtelijk gemaakt. Het betoog slaagt.

De PAS-bronmaatregelen bij elkaar genomen.

21. Zoals hiervoor onder 17 staat, heeft het college naast de marges die bij de afzonderlijke PAS-bronmaatregelen zijn aangehouden er ook op gewezen dat de prognose van de daling van ammoniakemissie als gevolg van de PAS-bronmaatregelen tezamen genomen 13,4 kiloton per jaar zal zijn, maar dat in het PAS rekening is gehouden met een totale daling van de ammoniakemissie met 6,4 kiloton per jaar in 2021. Derhalve is een marge van 7,0 kiloton ammoniakemissie aangehouden. Het college stelt dat dit voldoende is om onzekerheden in de prognoses op te vangen.

21.1. De Afdeling overweegt dat noch uit de stukken, noch uit het ter zitting verhandelde inzichtelijk is geworden welke aannames en gegevens zijn gehanteerd om het standpunt dat de aangehouden marge voldoende is om een eventuele tegenvallende prognose op te vangen te onderbouwen. Hierbij wijst zij op de conclusies onder 18.8, 19.4 en 20.4 over de afzonderlijke bronmaatregelen en op de omstandigheid dat het college geen separate onderbouwing heeft gegeven dat desondanks gezamenlijk bekeken de desbetreffende marges niettemin voldoende zijn om eventuele tegenvallende prognoses op te vangen. Dit geldt enerzijds voor de landelijke (gemiddelde) prognoses. Anderzijds geldt dit ook voor het standpunt dat als gevolg van het uitgeven van een deel van de depositiedaling in de vorm van ontwikkelingsruimte, uitgesloten is dat op afzonderlijke locaties de stikstofdepositie zal toenemen. Hierbij is van belang dat niet duidelijk is in hoeverre de effecten van de

afzonderlijke maatregelen in dit verband gezamenlijk beschouwd kunnen worden, gelet op de omstandigheid dat niet vaststaat dat de effecten van de bronmaatregelen voor elk van de Natura 2000-gebieden hetzelfde zijn.

De Afdeling stelt voorts vast dat de eerste PAS-periode loopt van 2015 tot 2021. In het PAS wordt rekening gehouden met gevolgen van de PAS-bronmaatregelen tot en met het jaar 2021 (Programma Aanpak Stikstof 2015-2021 zoals gewijzigd na partiële herziening op 15 december 2015"

p. 19). De CBS-berekening die voor de prognoses is gebruikt geeft echter prognoses van de gevolgen van de PAS-bronmaatregelen tot en met het jaar 2020. Niet duidelijk is of en in hoeverre dit verschil in jaartallen gevolgen heeft voor de conclusies die aan het PAS en de bijbehorende passende beoordeling zijn verbonden en zo ja, of met deze eventuele verschillen rekening is gehouden.

Ook in zoverre is de onderbouwing van de (gevolgen van de) maatregel, gelet op de eisen die in 14-14.4 zijn geformuleerd onvoldoende inzichtelijk. Het betoog slaagt.

Monitoren en bijsturen depositiedaling

22. Met onzekerheden die kleven aan de (effecten van) de autonome daling en de daling als gevolg van de drie PAS-bronmaatregelen wordt in het PAS naast het aanhouden van marges ook rekening gehouden door middel van een systeem van monitoren en bijsturen waarmee gezorgd kan worden dat eventuele tegenvallende resultaten kunnen worden opgevangen. De wettelijke grondslag hiervoor is beschreven onder 6.11.

22.1. De Werkgroep stelt dat het systeem van monitoren en bijsturen de onzekerheden over de autonome daling en de effecten van de bronmaatregelen niet wegneemt. Daarbij wijst zij erop dat de mogelijkheid van ingrijpen een vertraging kent, zodat dit systeem de mogelijkheid niet kan wegnemen dat de natuurlijke kenmerken van de Peelgebieden kunnen worden aangetast als gevolg van het PAS.

22.2. Uit de stukken die aan het PAS ten grondslag liggen en het verhandelde ter zitting blijkt dat een systeem in werking is om jaarlijks de feitelijke stikstofdepositie te bepalen. De resultaten hiervan komen in november van het jaar volgend op dat jaar beschikbaar, dat wil zeggen met een maximale vertraging van 23 maanden. Dit systeem werkt gedeeltelijk met metingen en gedeeltelijk met modelberekeningen. Hiermee is het mogelijk om per hexagoon in de PAS-gebieden te bepalen wat de feitelijke depositie van stikstof was in een bepaald jaar.

Naar het oordeel van de Afdeling doet het systeem van monitoren en bijsturen niet af aan de verplichting om voorafgaand aan de vergunningverlening zekerheid te hebben over de vraag of de natuurlijke kenmerken van de gebieden kunnen worden aangetast. De monitoring en bijsturing hebben slechts een aanvullende en controlerende functie die het mogelijk maakt dat gedurende de programmaperiode getoetst kan worden of de gevolgen waarvan in de passende beoordeling op basis van de beste wetenschappelijke kennis is uitgegaan zich ook daadwerkelijk manifesteren, waardoor bijsturing mogelijk is.

22.3. Het aangevoerde geeft de Afdeling geen aanleiding voor het oordeel dat het monitoren

op zichzelf niet adequaat is om vast te stellen wat de feitelijke depositie is op de afzonderlijke locaties van voor stikstof gevoelige habitats in de Natura 2000-gebieden die in het programma zijn opgenomen. Dat de gegevens hierover met enige vertraging beschikbaar komen, doet hier niet aan af. In zoverre slaagt het betoog niet.

Ten aanzien van de beroepsgrond die ziet op de mogelijkheden tot bijsturing in relatie tot de monitoring oordeelt de Afdeling dat niet inzichtelijk is gemaakt op welke wijze de gegevens uit de monitoring kunnen worden gebruikt voor adequate bijsturing. Aan dat oordeel ligt ten grondslag dat in het PAS is gesteld dat bijsturing kan plaatsvinden onder andere door het aanpassen van maatregelen of het nemen van nieuwe maatregelen gericht op depositiedaling, terwijl de monitoring ziet op de ontwikkeling van de depositie van stikstof en niet ziet op het verkrijgen van inzicht in de bijdrage in de daling van de onderscheiden maatregelen en ook niet anderszins is gebleken dat die relatieve bijdragen van de maatregelen wordt onderzocht. Ook is niet duidelijk gemaakt of en op welke wijze, rekening wordt gehouden met de vertraging in monitoringsgegevens, bijvoorbeeld door gehanteerde marges, teneinde een adequate bijsturing mogelijk te maken.

In zoverre is de onderbouwing van de (gevolgen van de) maatregel, gelet op de eisen die in 14-14.4 zijn geformuleerd onvoldoende inzichtelijk gemaakt. Het betoog slaagt.

Keuzes, gegevens en aannames over de omvang van de depositieruimte

Omvang gerelateerd aan economische groei van 2,5%

23. De Werkgroep betwijfelt of in de depositieruimte voldoende rekening is gehouden met alle autonome ontwikkelingen die daarbinnen moeten worden opgevangen. Zo is volgens de Werkgroep niet inzichtelijk hoeveel ruimte de activiteiten die onder de drempelwaarde vallen zullen innemen. Ook betwijfelt zij of rekening is gehouden met onbenutte ruimte in vergunningen en met de mogelijkheid dat nog vergunningen kunnen worden verleend met externe saldering. De aanname in het PAS dat door rekening te houden met een economische groei van 2,5%, waarop de omvang van de depostieruimte voor autonome ontwikkelingen is gebaseerd, een voldoende buffer bestaat, wordt door de Werkgroep bestreden. De aanname over de 2,5% groei is niet wetenschappelijk gefundeerd, maar gebaseerd op modelmatige berekeningen.

23.1. Het uitgangspunt van 2,5% economische groei betreft volgens het college een bovenraming waardoor in het PAS een buffer is ingebouwd om eventuele tegenvallers op te vangen.

23.2. In paragraaf 4.2.2 van het PAS staat dat de omvang van de depositieruimte voor autonome ontwikkelingen is bepaald rekening houdend met het scenario van een economische groei van 2,5%. Bij de berekening van de depositiebehoefte voor autonome ontwikkelingen is rekening gehouden met vaststaande beleidsmaatregelen en regelgeving. De autonome groei is de groei van activiteiten die reeds plaatsvinden bij de aanvang van het programma en waarvoor geen toestemming vooraf is vereist. Het gaat dan bijvoorbeeld om ontwikkelingen als de toename van de productie bij bedrijven - binnen de voorwaarden van een reeds verleende vergunning op grond van de wet - , de groei van het verkeer en consumentengroei.

23.3. De depositieruimte bestaat uit vier onderdelen: depositieruimte voor autonome ontwikkelingen, depositieruimte voor activiteiten onder de grenswaarde, depositieruimte voor prioritaire projecten (segment 1) en depositieruimte voor overige projecten (segment 2). Dit is geïllustreerd in de figuur in overweging 6.3. De benutting van de depositieruimte voor activiteiten onder de grenswaarde en de prioritaire en overige projecten wordt via een meldingen- en toedelingssysteem geregistreerd. De benutting van de depositieruimte voor autonome ontwikkelingen kan alleen achteraf via monitoring in beeld worden gebracht. De benutting van de depositieruimte voor autonome ontwikkelingen betreft dus een onzekerheid, waarvoor gelet op 14.2 geldt dat bij een kleine marge tussen de berekende behoefte van de depositieruimte en de beschikbaar gestelde depositieruimte, de zekerheid over de omvang van de autonome ontwikkelingen groter moet zijn dan bij een grote marge.

23.4. In de gebiedsanalyse voor de Groote Peel en de Deursche Peel & Mariapeel is de verdeling van de depositieruimte over de verschillende onderdelen weergegeven. In de Groote Peel is van de totale depositieruimte van 87 mol/ha/jr, 5 mol/ha/jr beschikbaarbaar gesteld voor autonome ontwikkelingen. In de Deurnsche Peel & Mariapeel is van de totale depositieruimte van 92 mol/ha/jr, 6 mol/ha/jr beschikbaar gesteld voor de autonome ontwikkelingen.

23.5. Zoals overwogen in 16.1 ziet de Afdeling, mede gelet op het RIVM-rapport 2015 waarin staat dat 2,5% economische groei een relatief hoge groeiverwachting is, in het niet nader onderbouwde betoog van de Werkgroep, geen aanleiding voor het oordeel dat dit uitgangspunt mogelijk een onderschatting van de economische groei betreft. Het standpunt van het college dat het een bovenraming betreft waardoor in het PAS een buffer is ingebouwd om eventuele tegenvallers in de depositiedaling op te kunnen vangen is echter niet onderbouwd met gegevens die inzichtelijk maken hoe groot de buffer is. Evenmin is inzichtelijk gemaakt waarom die buffer groot genoeg is om onzekerheden in de benutting van de depositieruimte voor de autonome daling te kunnen opvangen. Dit klemt te meer nu uit 23.4 volgt dat binnen de depositieruimte slechts een klein deel beschikbaar is voor autonome ontwikkelingen en zoals hierna, onder 24, 25 en 26, zal blijken er ook geen inzicht bestaat in de omvang van de depositiebehoefte van enkele specifieke categorieën van activiteiten die binnen de depositieruimte voor autonome ontwikkelingen moeten worden opgevangen, zoals activiteiten die onder de drempelwaarde vallen, de onbenutte ruimte in bestaande Nbw-vergunningen en toenames door extern salderen.

In zoverre is het standpunt dat met het scenario van 2,5% economische groei een buffer is ingebouwd om tegenvallers in de benutting van de depostieruimte voor autonome ontwikkelingen te kunnen opvangen, gelet op de eisen die in 14-14.4 zijn geformuleerd, onvoldoende inzichtelijk gemaakt. Het betoog slaagt.

Drempelwaarde

24. De Werkgroep betoogt dat er ten onrechte geen deugdelijke en voor derden verifieerbare onderbouwing in het PAS bestaat voor het vergunning- en meldingsvrij maken van alle activiteiten die wat stikstofdepositie betreft onder de drempelwaarde van 0,05 mol N/ha/jr blijven. De door het college gestelde indicatieve berekeningen ter onderbouwing van deze drempelwaarde maken geen deel uit van het PAS. De Werkgroep stelt dat er niet vanuit mag worden gegaan dat de maximale cumulatieve stikstofdepositie van alle activiteiten onder deze drempelwaarde geen aantasting van de natuurlijke kenmerken van de gebieden met

zich zal brengen.

24.1. Het college stelt zich op het standpunt dat de keuze voor een drempelwaarde van 0,05 mol N/ha/jr een bestuurlijke keuze is, waarbij rekening is gehouden met de effecten van de (cumulatieve) depositietoenames onder deze drempelwaarde van 0,05 mol N/ha/jr in het PAS. Op basis van indicatieve berekeningen is beoordeeld dat naar verwachting de maximale cumulatieve bijdrage van alle voorziene uitbreidingen van projecten of andere handelingen onder de drempelwaarde, afgezet tegen de te verwachten effecten van de maatregelen, de natuurlijke kenmerken van de gebieden niet zal aantasten. Uit de indicatieve berekeningen blijkt volgens het college dat de cumulatieve deposities van de projecten of andere handelingen onder deze drempelwaarde bij 95% van alle PAS-gebieden onder de 4 mol N/ha/jr blijven. Bij de overige 5% van de PAS-gebieden overstijgen de cumulatieve deposities de 7 mol N/ha/jr niet. De indicatieve berekeningen zijn gebaseerd op het rapport "Ontwikkeling dieraantallen tot 2020" van LEI/Wageningen UR van november 2014 (hierna: LEI-rapport). Vanwege de complexiteit en de omvang van de gegevensbestanden zijn deze berekeningen mondeling toegelicht op 27 juli 2016, waarbij medewerkers van de Stichting Advisering Bestuursrechtspraak voor Milieu en Ruimte, het ministerie van Economische Zaken en het RIVM aanwezig waren.

24.2. Voor de weergave van artikel 19kh, zevende lid, van de Nbw 1998, artikel 2 van het Besluit grenswaarden en artikel 8 van de Regeling wordt verwezen naar 7.2, 7.3 en 7.4.

24.3. Vast staat dat aan de bestuurlijke keuze om projecten en andere handelingen die een stikstofdepositie veroorzaken die de drempelwaarde van 0,05 mol N/ha/jr niet overschrijden vergunning- en meldingsvrij te maken berekeningen ten grondslag zijn gelegd. Uit deze berekeningen zou volgens het college afgeleid kunnen worden dat deze activiteiten een totale stikstofdepositie tot gevolg zouden hebben die als onderdeel van de autonome ontwikkelingen, mede gelet op de te treffen maatregelen, geen aantasting van de natuurlijke kenmerken van de gebieden tot gevolg kan hebben. Bedoelde berekeningen zijn echter niet uit het genoemde LEI-rapport af te leiden en zijn evenmin op een andere wijze openbaar en toegankelijk voor derden gemaakt. De mondelinge toelichting die heeft plaatsgevonden op 27 juli 2016 maakt dat niet anders, reeds omdat de Werkgroep daarvoor niet was uitgenodigd en uit de bij het deskundigenbericht overgelegde sheets van de powerpointpresentatie van deze toelichting deze berekeningen ook niet zijn af te leiden. Nu in zoverre geen volledige, tijdige en adequate beschikbaarstelling van deze berekeningen heeft plaatsgevonden is het voor de Werkgroep niet mogelijk geweest deze te beoordelen. Voor de Afdeling is het, bij het ontbreken van deze berekeningen, evenmin mogelijk om te beoordelen of in het PAS terecht is gesteld dat de maximale cumulatieve bijdrage van alle voorziene uitbreidingen van activiteiten onder de drempelwaarde van 0,05 mol N/ha/jr, is meegenomen in de passende beoordeling. Dat door middel van monitoring gevolgd kan worden of de totale depositie van deze activiteiten blijft binnen de beschikbare depositieruimte voor autonome ontwikkelingen betekent niet reeds dat er vanuit mag worden gegaan dat deze activiteiten cumulatief zo'n geringe depositie met zich zullen brengen dat de natuurlijke kenmerken van de Natura 2000-gebieden niet kunnen worden aangetast.

In zoverre is de onderbouwing van de drempelwaarde, gelet op de eisen die in 14-14.4 zijn geformuleerd, onvoldoende inzichtelijk gemaakt.

Het betoog slaagt.

Emissieruimte die ontstaat door ‘stoppers’ en relatie met extern salderen

25. De hierna te bespreken beroepsgrond heeft betrekking op zogenoemde externe saldering. Daarbij worden rechten op het houden van dieren, waarmee een bepaalde ammoniakemissie samenhangt, (hierna: ammoniakrechten) van de ene veehouderij overgedragen aan een andere veehouderij.

25.1. Volgens de Werkgroep is in het PAS onvoldoende rekening gehouden met de mogelijkheid die bestaat om ook tijdens de looptijd van het PAS externe saldering toe te passen. De Werkgroep wijst op het volgende. Een deel van de depositieruimte wordt geleverd door veehouders die stoppen met hun bedrijfsvoering, de zogenoemde stoppers. Op grond van het overgangsrecht van artikel 19km, vierde lid, van de Nbw 1998 is het volgens de Werkgroep mogelijk dat extern gesaldeerd wordt met de emissieruimte van deze stoppers. Hierbij zijn ammoniakrechten van stoppers overgedragen aan uitbreiders die voor de uitbreiding een Nbw-vergunning hebben gekregen of nog zullen krijgen. Volgens de Werkgroep is niet duidelijk of er rekening mee is gehouden dat de ruimte die in het kader van externe saldering is gebruikt daarnaast ook als depositieruimte wordt ingezet bij de uitvoering van het PAS en daarmee twee keer wordt gebruikt.

25.2. Het college stelt dat het door de Werkgroep genoemde effect zich kan voordoen, maar stelt daar tegenover dat er een zodanig grote veiligheidsmarge is ingebouwd bij de berekening van de stoppersruimte dat ten gevolge van dit effect niet hoeft te worden gevreesd voor een aantasting van de natuurlijke kenmerken van Natura 2000-gebieden. De depositieafname als gevolg van de stoppende veehouderijen is volgens het college in werkelijkheid veel groter dan de zogenoemde stoppersruimte die als depositieruimte beschikbaar wordt gesteld. Het college wijst erop dat alleen de helft van de stoppende veehouderijen en alleen de veehouderijen die liggen op grotere afstand dan 1 km tot Natura 2000-gebieden bij de berekening van de stoppersruimte is meegenomen. Gelet hierop is de stoppersruimte groot genoeg om het effect van het twee keer gebruiken van externe saldering, voor zover dit zich voordoet, op te vangen. Er zal dan ook niet te veel depositieruimte worden uitgegeven. Voorts wordt de werkelijke hoeveelheid emissierechten die beschikbaar komt door bedrijfsbeëindiging gemonitord. Volgens het college is er overigens alleen sprake van enige dubbeltelling van de emissie als de stoppende veehouderij in het referentiejaar van het PAS vee in de stal had staan.

25.3. In artikel 19km, derde lid, van de Nbw 1998 staat dat voor een project dat of een andere handeling die betrekking heeft op een inrichting als bedoeld in artikel 1.1, derde lid, van de Wet milieubeheer en stikstofdepositie veroorzaakt op een in het programma opgenomen Natura 2000-gebied een vergunning als bedoeld in artikel 19d, eerste lid, niet wordt verleend op grond van het feit dat onmiddellijk in verband met dit project of deze andere handeling een afname van de stikstofdepositie plaatsvindt als gevolg van de beëindiging of beperking van een of meer bepaalde andere handelingen buiten die inrichting.

In het vierde lid staat dat het derde lid niet van toepassing is op een besluit op een aanvraag om een vergunning als bedoeld in artikel 19d, eerste lid, die is ingediend vóór het tijdstip van inwerkingtreding van het derde lid.

In artikel 67a staat dat artikel 19km niet van toepassing is op projecten, plannen en andere

handelingen die stikstofdepositie op voor stikstof gevoelige habitats in een Natura 2000-gebied veroorzaken indien:

- voor het project, plan of de andere handeling voor het tijdstip van inwerkingtreding van het programma een besluit als bedoeld in artikel 19km, eerste lid, in voorbereiding is bij het desbetreffende bestuursorgaan;

- de voor het nemen van het desbetreffende besluit beschikbare gegevens en bescheiden naar het oordeel van het desbetreffende bestuursorgaan voldoende zijn voor de beoordeling van de aanvraag of voor de voorbereiding van het desbetreffende besluit en bovendien, ingeval het besluit betrekking heeft op een project als bedoeld in artikel 19f, eerste lid, een volledige passende beoordeling als bedoeld in dat artikellid is gemaakt, en

- degene die het desbetreffende project zal realiseren, onderscheidenlijk de andere handeling zal verrichten, een tijdige uitvoering heeft verzekerd van de maatregelen die in het kader van de realisering van het project, onderscheidenlijk het verrichten van de andere handeling worden getroffen om te verzekeren dat de natuurlijke kenmerken van Natura 2000-gebied niet zullen worden aangetast als gevolg van het project, onderscheidenlijk om verslechteringen of significant verstorende effecten als gevolg van de andere handeling te voorkomen.

25.4. Het onder 25.2 weergegeven verweer is in het PAS als volgt terug te vinden. Bij de berekening van de in het kader van het PAS uit te geven depositieruimte die ontstaat door vrijkomende emissieruimte van stoppende veehouderijen, de zogenoemde stoppersruimte, zijn niet alle veehouderijen betrokken die naar verwachting zullen stoppen. Voor 2020 en 2030 is de helft van het percentage van het aantal landbouwbedrijven dat in de desbetreffende provincie naar verwachting vanaf 1 juli 2015 tot 2030 stopt en die liggen op grotere afstand dan 1 km tot Natura 2000-gebieden betrokken bij deze berekening. Van de stoppersruimte wordt gedurende de hele PAS-periode jaarlijks een even groot deel uitgegeven als depositieruimte. Er is derhalve geen directe koppeling tussen concrete verzoeken om intrekking van ammoniakemissierechten van stoppende veehouderijen en de omvang van de depositieruimte die als stoppersruimte beschikbaar wordt gesteld.

25.5. Uit artikel 19km, derde lid, van de Nbw 1998 volgt dat externe saldering onder het PAS niet is toegestaan. De Afdeling stelt vast dat zowel op grond van artikel 19km, vierde lid, van de Nbw 1998 als op grond van artikel 67a in de looptijd van het PAS onder voorwaarden nog wel toepassing kan worden gegeven aan externe saldering. Zoals de Afdeling eerder heeft overwogen (zie onder meer de uitspraak van 13 november 2013, ECLI:NL:RVS:2013:1931) is extern salderen mogelijk met stikstofdepositie die waren vergund en weliswaar niet feitelijk aanwezig waren maar dat wel konden zijn tot het moment van de intrekking van de vergunning van de saldogever of het sluiten van de overeenkomst over de overname van de ammoniakemissie ten behoeve van de uitbreiding van het saldo-ontvangende bedrijf. Gelet hierop kan door toepassing van externe saldering met een bedrijf dat in 2014 feitelijk geen stikstofdepositie meer veroorzaakte, een toename plaatsvinden van de stikstofdepositie ten opzichte van de stikstofdepositie die feitelijk plaatsvond in het referentiejaar van het PAS. Naar het oordeel van de Afdeling is niet inzichtelijk gemaakt of, en, zo ja, op welke wijze in het PAS rekening is gehouden met deze mogelijke toename en of deze toename kan worden opgevangen binnen de beschikbare depositieruimte voor autonome ontwikkelingen. In zoverre is de onderbouwing van de gevolgen van de mogelijkheden die het overgangsrecht

biedt voor externe saldering, gelet op de eisen die in 14-14.4 zijn geformuleerd onvoldoende inzichtelijk gemaakt.

25.6. De overgangsregeling voor extern salderen sluit voorts niet uit dat gesaldeerd wordt met een bedrijf dat op 1 juli 2015 nog in bedrijf was. In die gevallen is het in elk geval deels mogelijk, zoals het college ook erkent, dat emissieruimte waarmee extern is of nog zal worden gesaldeerd ook wordt vrijgegeven als depositieruimte. Anders dan het college meent, is de Afdeling niet van oordeel dat geen redelijke twijfel bestaat over de vraag of ten gevolge van deze dubbeltelling de natuurlijke kenmerken van Natura 2000-gebieden kunnen worden aangetast. Daarbij is het volgende van belang.

De percentages stoppende landbouwbedrijven per provincie zijn in juli 2014 aangeleverd door de provincies en weergegeven in de factsheet "Bepalen ontwikkelingsruimte stoppers" van AERIUS. Niet is echter inzichtelijk gemaakt op basis waarvan deze percentages zijn vastgesteld. Voor zover het college stelt dat deze percentages mede zijn gebaseerd op het rapport "Vrijkomende agrarische bebouwing in het landelijk gebied" uit maart 2014 van Alterra overweegt de Afdeling dat genoemde percentages niet uit dit rapport zijn af te leiden. De wijze waarop deze percentages zijn vastgesteld is evenmin op een andere wijze inzichtelijk gemaakt. De enkele stelling van het college ter zitting dat de percentages zijn vergeleken met het eerder genoemde LEI-rapport is daarvoor onvoldoende.

Dat alleen sprake kan zijn van een dubbeltelling van de emissie bij stoppende veehouderijen indien bij de betreffende veehouderij in het referentiejaar vee in de stal stond, betekent niet reeds dat er vanuit mag worden gegaan dat de kans op dubbeltelling zodanig klein is dat de natuurlijke kenmerken van de Natura 2000-gebieden niet kunnen worden aangetast. Voorts is niet inzichtelijk gemaakt op welke wijze hiermee rekening is gehouden in de prognose. Dat door middel van monitoring kan worden gecontroleerd wat de werkelijke hoeveelheid emissieruimte is die beschikbaar komt door stoppers betekent niet dat de natuurlijke kenmerken van de Natura 2000-gebieden niet kunnen worden aangetast. Als uit de monitoring blijkt dat de totale emissieruimte die beschikbaar komt kleiner is dan waarmee na toepassing van de marges rekening is gehouden, betekent dit slechts dat bijsturing nodig kan zijn.

Voorts is, zoals in het deskundigenverslag staat, niet uitgesloten dat door gebruik te maken van stoppersruimte de stikstofdepositie lokaal per saldo toeneemt. Dit zal zich voordoen in het geval in de directe omgeving van een Natura 2000-gebied in een bepaald PAS-jaar geen veehouderijen stoppen, terwijl ter plaatse een veehouderij uitbreidt door gebruik te maken van de aan de depositieruimte toegevoegde stoppersruimte. Niet inzichtelijk is gemaakt op welke wijze met dergelijke toenames van lokale stikstofdeposities rekening is gehouden in de passende beoordeling.

In zoverre is de onderbouwing van de gevolgen van een mogelijke dubbeltelling van emissieruimte in het kader van extern salderen, gelet op de eisen die in 14-14.4 zijn geformuleerd, onvoldoende inzichtelijk gemaakt. Het betoog slaagt.

Onbenutte emissieruimte in bestaande vergunningen

26. De Werkgroep voert aan dat in de bestaande Nbw-vergunningen veel onbenutte emissieruimte zit waarmee de veestapel en daarmee de ammoniakemissie kan worden

uitgebreid. In het PAS is geen rekening gehouden met deze onbenutte ruimte. Niet is onderzocht hoe groot de onderbezetting binnen de bestaande Nbw-vergunningen is, bijvoorbeeld aan de hand van landbouwtellingen en andere telgegevens.

26.1. Het college stelt dat in het PAS rekening is gehouden met onbenutte ruimte binnen bestaande Nbw-vergunningen. Het opvullen van deze ruimte is volgens het college een vorm van economische groei en kan binnen de depositieruimte voor autonome ontwikkelingen worden opgevangen.

26.2. De Afdeling stelt vast dat het gebruik maken van onbenutte emissieruimte in bestaande Nbw-vergunningen niet in strijd is met het verbod in artikel 19d, eerste lid, van de Nbw 1998. De bepalingen uit hoofdstuk III, titel 2, paragraaf 2a van de Nbw 1998, het Besluit grenswaarden en de Regeling zijn hierop niet van toepassing. Dit betekent dat onbenutte emissieruimte in bestaande Nbw-vergunningen kan worden opgevuld, met als gevolg dat de stikstofemissie van het vergunninghoudende bedrijf toeneemt, zonder dat daarvoor een meldingsplicht geldt of een vergunning hoeft te worden verleend.

Voor zover het college stelt dat het opvullen van onbenutte emissieruimte in bestaande Nbw-vergunningen kan worden opgevangen binnen de depositieruimte voor autonome ontwikkelingen omdat deze ontwikkeling valt binnen de 2,5% economische groei overweegt de Afdeling als volgt. Dit percentage betreft een landelijk gemiddelde over alle economische sectoren. Weliswaar staat in het PAS dat bij de bepaling van de depositiebehoefte voor nieuwe economische ontwikkelingen rekening is gehouden met sectorspecifieke ontwikkelingen en sectorspecifiek beleid en dat rekening is gehouden met de verschillende landbouwsectoren, maar de wijze waarop dit is uitgevoerd is niet inzichtelijk gemaakt. Daarom is niet inzichtelijk of binnen de depositieruimte voor autonome ontwikkelingen voldoende ruimte aanwezig is om eventuele hoger uitvallende stikstofdeposities ten gevolge van het opvullen van onbenutte emissieruimte in een bepaald jaar op te vangen. Naar het oordeel van de Afdeling is het wel mogelijk om inzicht te krijgen in de hoeveelheid onbenutte emissieruimte in bestaande Nbw-vergunningen. Daarbij betrekt zij dat gegevens beschikbaar zijn over de dierenaantallen per stalsysteem in het referentiejaar van het PAS. Op basis van deze gegevens in combinatie met de verleende Nbw-vergunningen en de sectorspecifieke ontwikkelingen, waaronder de economische groei, en sectorspecifiek beleid is het in beginsel mogelijk om beter inzicht te verschaffen in het te verwachten verloop van de opvulling van onbenutte emissieruimte in bestaande Nbw-vergunningen binnen de PAS-periode. Dit is vooralsnog onvoldoende inzichtelijk gemaakt.

In zoverre zijn de gevolgen van het gebruik van thans onbenutte emissieruimte in bestaande Nbw-vergunningen, gelet op de eisen die in 14-14.4 zijn geformuleerd, onvoldoende inzichtelijk gemaakt. Het betoog slaagt.

Conclusie onderdeel H

27. Hetgeen de Afdeling heeft overwogen onder H betekent dat er gebreken aan het PAS kleven. De bestreden besluiten zijn op het PAS gebaseerd, zodat deze gebreken moeten worden hersteld door genoemde ministers en de staatssecretaris.

Gelet op 15.6 en 16.1 moeten zij onderbouwen waarom, gelet op de geconstateerde contra-indicaties, van de bestaande dalende trend in de stikstofdepositie in het PAS mag worden

uitgegaan en in hoeverre het uitgangspunt van 2,5% gemiddelde economische groei tot een buffer leidt die mogelijke tegenvallers in de autonome daling van de stikstofdepositie kan opvangen.

Gelet op 18.8, 19.4, 20.4 en 21.1 moeten zij onderbouwen in hoeverre aangehouden marges bij de PAS-bronmaatregelen voldoende zijn om tegenvallende resultaten op te vangen en moet worden onderbouwd of de effecten van deze maatregelen zich gelijkmatig over de verschillende Natura 2000-gebieden verdelen.

Gelet op 22.3 moeten zij in verband met het systeem van monitoring en bijsturing onderbouwen hoe op basis hiervan bijsturing kan plaatsvinden, waarbij ook aangegeven moet worden op welke wijze rekening wordt gehouden met de vertraging in het beschikbaar komen van de monitoringsgegevens.

Gelet op 23.5 moeten zij onderbouwen dat een buffer van 2,5% van gemiddelde economische groei voldoende depositieruimte met zich brengt om mogelijke tegenvallers in het kader van de autonome ontwikkelingen op te kunnen vangen.

Gelet op 24.3 moeten zij de berekeningen die ten grondslag zijn gelegd aan de drempelwaarde beschikbaar en toegankelijk maken.

Gelet op 25.5 en 25.6 moeten zij onderbouwen of en zo ja, op welk wijze rekening is gehouden met gevolgen van de overgangsregeling voor extern salderen.

Gelet op 26.2 moeten zij onderbouwen of in het PAS voldoende rekening is gehouden met onbenutte emissieruimte binnen bestaande Nbw-vergunningen binnen de depositieruimte voor autonome ontwikkelingen.

I. SLOT

28. De behandeling van de beroepen zal worden geschorst tot het Hof van Justitie uitspraak heeft gedaan.

29. De Werkgroep heeft de Afdeling verzocht ambtshalve een voorlopige voorziening te treffen in het geval prejudiciële vragen aan het Hof van Justitie worden gesteld.

29.1. De Afdeling ziet, mede gelet op artikel 4, derde lid, van het Verdrag betreffende de Europese Unie, aanleiding te beoordelen of termen aanwezig zijn een voorlopige voorziening als bedoeld in artikel 8:72, vijfde lid, van de Algemene wet bestuursrecht (hierna: Awb) te treffen.

29.2. De omstandigheid dat prejudiciële vragen worden gesteld betekent niet dat een verzoek tot het treffen van een voorlopige voorziening zonder meer zou moeten worden toegewezen zolang het Hof van Justitie nog geen uitspraak over de prejudiciële vragen heeft gedaan. Wel brengt dit met zich, dat dient te worden bezien of afwijzing van het verzoek tot het treffen van een voorlopige voorziening zal leiden tot onomkeerbare gevolgen in het licht van de doelstellingen van de toepasselijke richtlijn en de vragen die zijn gesteld over de uitleg van een aantal bepalingen van die richtlijn.

29.3. De onderwerpen waarover prejudiciële vragen worden gesteld, het PAS-toestemmingsregime en de maatregelen die in een passende beoordeling mogen worden betrokken, geven bij afweging van de betrokken belangen, geen aanleiding voor het treffen van een voorlopige voorziening. De Afdeling stelt daarbij voorop dat zij een programmatische aanpak zoals het PAS, dat enerzijds gericht is op het behoud en waar mogelijk herstel van natuurwaarden en anderzijds op het scheppen van depositieruimte voor bestaande en toekomstige activiteiten, op voorhand geen ongeschikt instrument acht om uitvoering te geven aan de verplichtingen die voortvloeien uit artikel 6 van de Habitatrichtlijn. Zoals overwogen in 9.12 en 9.17 acht de Afdeling aannemelijk dat artikel 6 van de Habitatrichtlijn ruimte biedt voor het PAS-toestemmingsregime op grond waarvan projecten en andere handelingen die stikstofdepositie veroorzaken die een bepaalde drempel- of grenswaarde niet overschrijden zonder individuele toestemming zijn toegestaan en dat bij de verlening van een vergunning gebruik kan worden gemaakt van de passende beoordeling die voor het PAS is gemaakt. Van belang daarbij is dat aan het PAS een passende beoordeling ten grondslag is gelegd. In die passende beoordeling zijn maatregelen en autonome ontwikkelingen betrokken waarvan de Afdeling, indien deze voldoen aan de voorwaarden vermeld in 10.18, in haar rechtspraak heeft aangenomen dat deze in een passende beoordeling betrokken mogen worden. Bovendien is aan een programma dat enerzijds gericht is op het behoud en waar mogelijk herstel van natuurwaarden en anderzijds op het scheppen van depositieruimte voor bestaande en toekomstige activiteiten die in samenhang worden beoordeeld, inherent dat de gevolgen van het benutten van de depositieruimte worden beoordeeld in samenhang met alle maatregelen en autonome ontwikkelingen die zich tijdens de programmaperiode in het Natura 2000-gebied zullen voordoen.

29.4. Uit onderdeel H van deze uitspraak vloeit voort dat de Afdeling een nadere onderbouwing nodig acht van enkele keuzes, gegevens en aannames die ten grondslag liggen aan het PAS en de daarbij behorende passende beoordeling. Met name dient een nadere onderbouwing te worden gegeven van de keuze en omvang van enkele buffers en marges waarmee in het PAS rekening is gehouden bij de voorspelde daling van de stikstofdepositie. Deze vragen raken in de kern de berekende omvang van de depositieruimte. Daarnaast zijn er vragen over de aannames en berekeningen van de depositiebijdrage van de autonome ontwikkelingen. Deze vragen hebben betrekking op de verdeling van de depositieruimte over de verschillende segmenten, in het bijzonder of binnen de depositieruimte voldoende ruimte is gereserveerd voor autonome ontwikkelingen. De Afdeling acht waarschijnlijk dat deze gebreken zodanig hersteld kunnen worden dat het PAS ongewijzigd of in bijgestelde vorm, dat wil zeggen met minder depositieruimte of met bijstelling van de reservering van de depositieruimte voor autonome ontwikkelingen, doorgang kan vinden.

29.5. De Afdeling ziet in deze gebreken thans geen aanleiding voor het treffen van een voorlopige voorziening. Zij overweegt daartoe als volgt. In paragraaf 4.2.6 van het PAS zijn de uitgangspunten voor de verdeling van de depositieruimte in de eerste en tweede helft van het tijdvak (zes jaar) van het programma beschreven. Daaruit volgt dat 60% van de ontwikkelingsruimte voor segment 2 beschikbaar is voor toedeling in de eerste helft van het tijdvak en 40% voor toedeling in het tweede tijdvak. Voor segment 1 (de prioritaire projecten) geldt een dergelijke verdeling niet, maar de verwachting bestaat dat de ontwikkelingsruimte voor dit segment niet geheel zal zijn benut in de eerste helft van het tijdvak.

Het PAS is op 1 juli 2015 in werking getreden, zodat de eerste helft van het tijdvak eindigt op

1 juli 2018. Wanneer de ontwikkelingsruimte volgens de uitgangspunten voor de verdeling van de depositieruimte zoals opgenomen in 4.2.6 van het PAS wordt gereserveerd, dan is verzekerd dat tot 1 juli 2018 de ontwikkelingsruimte die voor de tweede helft van het tijdvak is gereserveerd niet is uitgegeven. Er is als het ware binnen het systeem een buffer ingebouwd. De Afdeling acht niet aannemelijk dat de stikstofdepositie die kan ontstaan door benutting van de depositieruimte en de toedeling van de ontwikkelingsruimte in de eerste helft van het tijdvak van het PAS, uitgaande van de uitvoering van de bron- en herstelmaatregelen zoals in het PAS voorzien, onomkeerbare gevolgen zal hebben. Zij ziet hierin aanleiding te overwegen dat er in beginsel tot 1 juli 2018 geen aanleiding bestaat tot het treffen van een voorlopige voorziening hangende de behandeling van de verwijzingsuitspraak bij het Hof van Justitie.

29.6. Het voorgaande kan anders zijn als voor de betrokken Natura 2000-gebieden alsnog gebruik wordt gemaakt van de in 4.2.9 van het PAS beschreven uitgangspunten voor het verhogen van de depositieruimte dan wel indien gebruik wordt gemaakt van de in artikel 2.14 van het Besluit natuurbescherming geboden mogelijkheid te besluiten dat het verbod op extern salderen voor gebieden die in het PAS zijn opgenomen buiten toepassing blijft, en dit Natura 2000-gebieden betreft die in de betrokken vergunningzaken aan de orde zijn. Daarnaast is van belang dat met de uitvoering van de herstelmaatregelen die in het PAS zijn voorzien wordt doorgegaan.

29.7. Onder de hiervoor genoemde omstandigheden bestaat bij de Afdeling niet de verwachting dat het niet treffen van een voorlopige voorziening zal leiden tot onomkeerbare gevolgen. De Afdeling ziet daarom, bij afweging van de betrokken belangen, geen aanleiding om een voorlopige voorziening als bedoeld in artikel 8:72, vijfde lid, van de Awb te treffen. Dit laat onverlet dat, zolang de vergunningen niet in rechte onaantastbaar zijn, de vergunninghouder op eigen risico daarvan gebruik maakt.

29.8. Indien het Hof van Justitie voor 1 juli 2018 geen uitspraak heeft gedaan op de prejudiciële vragen, dan zal, indien een verzoek om voorlopige voorziening wordt ingediend, worden bezien of er alsdan aanleiding bestaat tot het treffen van een maatregel. Daarbij zal een eventuele reactie van de ministers op de in onderdeel H geconstateerde gebreken, worden betrokken.

Beslissing

De Afdeling bestuursrechtspraak van de Raad van State:

I. verzoekt het Hof van Justitie van de Europese Unie bij wege van prejudiciële beslissing uitspraak te doen op de volgende vragen:

1. Staat artikel 6, tweede en derde lid, van Richtlijn 92/43/EEG van de Raad van de Europese Gemeenschappen van 21 mei 1992 inzake de instandhouding van de natuurlijke habitats en de wilde flora en fauna (PbEG1992 L206; Habitatrichtlijn) in de weg aan een wettelijke regeling die ertoe strekt dat projecten en andere handelingen die stikstofdepositie veroorzaken die een drempel- of grenswaarde niet overschrijden, zijn uitgezonderd van de vergunningplicht en daardoor zonder individuele toestemming zijn toegestaan, ervan uitgaande dat de gevolgen van alle projecten en andere handelingen tezamen die gebruik kunnen maken van de wettelijke regeling voor de vaststelling van die wettelijke regeling

passend zijn beoordeeld?

2. Staat artikel 6, tweede en derde lid, van Richtlijn 92/43/EEG van de Raad van de Europese Gemeenschappen van 21 mei 1992 inzake de instandhouding van de natuurlijke habitats en de wilde flora en fauna (PbEG1992 L206) eraan in de weg dat een passende beoordeling voor een programma waarin een bepaalde totale hoeveelheid stikstofdepositie is beoordeeld ten grondslag wordt gelegd aan de verlening van een vergunning (individuele toestemming) voor een project of andere handeling, die stikstofdepositie veroorzaakt die binnen de in het kader van het programma beoordeelde depositieruimte past?

3. Mogen in de passende beoordeling als bedoeld in artikel 6, derde lid, van Richtlijn 92/43/EEG de Raad van de Europese Gemeenschappen van 21 mei 1992 inzake de instandhouding van de natuurlijke habitats en de wilde flora en fauna (PbEG1992 L206), die voor een programma, zoals het Programma Aanpak Stikstof 2015-2021, is gemaakt, de positieve gevolgen van instandhoudingsmaatregelen en passende maatregelen voor bestaande arealen van habitattypen en leefgebieden worden betrokken, die worden getroffen in verband met de verplichtingen die voortvloeien uit artikel 6, eerste en tweede lid, van die richtlijn?

3a. Indien vraag 3 bevestigend wordt beantwoord: kunnen de positieve gevolgen van instandhoudingsmaatregelen en passende maatregelen in een passende beoordeling voor een programma worden betrokken als deze ten tijde van de passende beoordeling nog niet zijn uitgevoerd en het positieve effect daarvan nog niet is verwezenlijkt?

Is daarbij, ervan uitgaande dat de passende beoordeling definitieve bevindingen bevat over de gevolgen van deze maatregelen die gebaseerd zijn op de beste wetenschappelijke kennis ter zake, van belang dat de uitvoering en het resultaat van die maatregelen wordt gemonitord en indien daaruit volgt dat de gevolgen ongunstiger zijn dan waarvan is uitgegaan in de passende beoordeling, bijsturing, indien nodig, plaatsvindt?

4. Mogen de positieve gevolgen van de autonome daling van stikstofdepositie die zich zal gaan manifesteren in de periode waarin het Programma Aanpak Stikstof 2015-2021 geldt, in de passende beoordeling als bedoeld in artikel 6, derde lid, van Richtlijn 92/43/EEG van de Raad van de Europese Gemeenschappen van 21 mei 1992 inzake de instandhouding van de natuurlijke habitats en de wilde flora en fauna (PbEG1992 L206), worden betrokken?

Is daarbij, ervan uitgaande dat de passende beoordeling definitieve bevindingen bevat over deze ontwikkelingen die gebaseerd zijn op de beste wetenschappelijke kennis ter zake, van belang dat de autonome daling van stikstofdepositie wordt gemonitord, en indien daaruit volgt dat de daling ongunstiger is dan waarvan is uitgegaan in de passende beoordeling, bijsturing, indien nodig, plaatsvindt?

5. Mogen herstelmaatregelen die in het kader van het Programma Aanpak Stikstof 2015-2021 worden getroffen en waarmee wordt voorkomen dat een bepaalde natuurbelastende factor, zoals stikstofdepositie, schadelijke gevolgen kan hebben voor bestaande arealen van habitattypen of leefgebieden, geduid worden als beschermingsmaatregel als bedoeld in punt 28 van het arrest van het Hof van Justitie van 15 mei 2014, Briels, ECLI:EU:C:2014:330, die in een passende beoordeling als bedoeld in artikel 6, derde lid, van Richtlijn 92/43/EEG van de Raad van de Europese Gemeenschappen van 21 mei 1992 inzake de instandhouding van de

natuurlijke habitats en de wilde flora en fauna (PbEG1992 L206) mogen worden betrokken?

5a. Indien vraag 5 bevestigend wordt beantwoord: kunnen de positieve gevolgen van beschermingsmaatregelen die in de passende beoordeling mogen worden betrokken, daarin worden betrokken, als deze ten tijde van de passende beoordeling nog niet zijn uitgevoerd en het positieve effect daarvan nog niet is verwezenlijkt?

Is daarbij, ervan uitgaande dat de passende beoordeling definitieve bevindingen bevat over de gevolgen van deze maatregelen die gebaseerd is op de beste wetenschappelijke kennis ter zake, van belang dat de uitvoering en het resultaat van de maatregelen wordt gemonitord en indien daaruit volgt dat de gevolgen ongunstiger zijn dan waarvan is uitgegaan in de passende beoordeling, bijsturing, indien nodig, plaatsvindt?

II. schorst de behandeling en houdt iedere verdere beslissing aan.

Aldus vastgesteld door mr. B.J. van Ettekoven, voorzitter, en mr. R. Uylenburg en mr. B.J. Schueler, leden, in tegenwoordigheid van mr. J. Verbeek, griffier.

w.g. Van Ettekoven w.g. Verbeek
voorzitter griffier

Uitgesproken in het openbaar op 17 mei 2017

388-459-653-723

ONAFHANKELIJKE

DESKUNDIGENGROEP OP HOOG NIVEAU

INZAKE KUNSTMATIGE INTELLIGENTIE

OPGERICHT DOOR DE EUROPESE COMMISSIE IN JUNI 2018

ETHISCHE RICHTSNOEREN VOOR BETROUWBARE KI

ETHISCHE RICHTSNOEREN voor BETROUWBARE KI

Deskundigengroep op hoog niveau inzake kunstmatige intelligentie

Dit document is opgesteld door de deskundigengroep op hoog niveau inzake kunstmatige intelligentie (AI HLEG). De in dit document genoemde leden van de AI HLEG ondersteunen het algehele kader voor betrouwbare KI dat in deze richtsnoeren uiteen wordt gezet, maar zijn het niet noodzakelijkerwijs eens met iedere afzonderlijke bewering in het document. .

De controlelijst voor betrouwbare KI die in hoofdstuk III van dit document wordt gepresenteerd, wordt gedurende een testfase door belanghebbenden getest om praktische feedback te verzamelen. Begin 2020 wordt er een herziene versie van de controlelijst aan de Europese Commissie voorgelegd, waarin de tijdens de testfase verzamelde feedback wordt verwerkt.

De AI HLEG is een onafhankelijke deskundigengroep die in juni 2018 door de Europese Commissie is opgericht.

Contactpersoon Nathalie Smuha – Coördinator AI HLEG
E-mailadres CNECT-HLG-AI@ec.europa.eu

Europese Commissie
B-1049 Brussel

Document gepubliceerd op 8 april 2019.

Op 18 december 2018 is er een eerste ontwerp van dit document uitgebracht, waarop middels een openbare raadpleging feedback is gegeven door meer dan vijfhonderd belanghebbenden. Wij willen iedereen die feedback op het eerste ontwerp van het document heeft gegeven, uitdrukkelijk en hartelijk bedanken. Deze feedback is meegenomen bij het opstellen van deze herziene versie.

PDF ISBN 978-92-76-11999-9 doi:10.2759/61918 KK-02-19-841-NL-N
Print ISBN 978-92-76-12802-1 doi:10.2759/924378 KK-02-19-841-NL-C

Meer informatie over de deskundigengroep op hoog niveau inzake kunstmatige intelligentie is online beschikbaar (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

INHOUDSOPGAVE

SAMENVATTING

(1) Het doel van deze richtsnoeren is het bevorderen van betrouwbare KI. Betrouwbare KI bestaat uit **drie componenten**, waaraan gedurende de volledige levenscyclus van het systeem moet worden voldaan: de KI moet a) **wettig** zijn, door te voldoen aan alle toepasselijke wet- en regelgeving, b) **ethisch** zijn, door naleving van ethische beginselen en waarden te waarborgen, en c) **robuust** zijn uit zowel technisch als sociaal oogpunt, aangezien KI-systemen ongewild schade kunnen aanrichten, zelfs al zijn de bedoelingen goed. Iedere component is op zichzelf nodig, maar niet voldoende om betrouwbare KI te bewerkstelligen. In het ideale scenario sluiten alle drie de componenten op elkaar aan en valt de werking ervan gedeeltelijk samen. Als er in de praktijk spanningen ontstaan tussen deze componenten, moet de maatschappij zich inspannen om ze weer op één lijn te brengen.

(2) In deze richtsnoeren wordt een **kader voor het bewerkstelligen van betrouwbare KI** geschetst. In het kader wordt niet expliciet ingegaan op de eerste component van betrouwbare KI (wettige KI)[1]. In plaats daarvan is het bedoeld als richtsnoer voor het bevorderen en waarborgen van ethische en robuuste KI (de tweede en derde component). Deze richtsnoeren, die gericht zijn op alle belanghebbenden, vormen een poging verder te gaan dan een lijst van ethische beginselen, doordat wordt toegelicht hoe dergelijke beginselen in de praktijk in sociaal-technische systemen kunnen worden verwerkt. Deze toelichting wordt in drie lagen met verschillend abstractieniveau gegeven: van het meest abstract in hoofdstuk I tot het meest concreet in hoofdstuk III, dat eindigt met voorbeelden van de mogelijkheden en punten van zorg die KI-systemen met zich meebrengen.

I. Op basis van een op grondrechten gefundeerde benadering worden in hoofdstuk I de **ethische beginselen** en de bijbehorende waarden vastgesteld die bij de ontwikkeling, de installatie en het gebruik van KI-systemen moeten worden gerespecteerd.

Essentiële richtsnoeren uit hoofdstuk I:

- ✓ Ontwikkel, installeer en gebruik KI-systemen op dusdanige wijze dat de volgende ethische beginselen worden nageleefd: *respect voor menselijke autonomie, preventie van schade, rechtvaardigheid en verantwoording.* Erken de potentiële spanningen tussen deze beginselen en pak deze aan.
- ✓ Schenk bijzondere aandacht aan situaties waarbij kwetsbaardere groepen betrokken zijn, zoals kinderen, mensen met een beperking en anderen die historisch gezien kansarm zijn of het risico lopen te worden uitgesloten, en aan situaties die worden gekenmerkt door ongelijkheid wat betreft macht of beschikking over informatie, bijvoorbeeld tussen werkgevers en werknemers of tussen bedrijven en consumenten[2].
- ✓ Erken en wees u ervan bewust dat KI-systemen de mens en de samenleving grote voordelen kunnen opleveren, maar ook bepaalde risico's en negatieve gevolgen met zich mee kunnen brengen, die mogelijk lastig te voorspellen, vast te stellen of te meten zijn (bijv. gevolgen voor de democratie, de rechtsstaat en de verdelende rechtvaardigheid of voor de menselijke geest). Neem waar nodig passende maatregelen, evenredig aan de omvang van het risico, om deze risico's te beperken.

II. Voortbouwend op hoofdstuk I worden er in hoofdstuk II richtsnoeren geboden voor de bewerkstelliging van betrouwbare KI, in de vorm van **zeven vereisten** voor KI-systemen. Voor de verwezenlijking hiervan kunnen zowel technische als niet-technische methoden worden gebruikt.

---

1 Alle normatieve beweringen in dit document zijn bedoeld als richtsnoer voor het verwezenlijken van de tweede en derde component van betrouwbare KI (ethische en robuuste KI). Deze beweringen zijn dus niet bedoeld als juridisch advies of als hulp bij het naleven van de toepasselijke wetgeving, hoewel vele ervan tot op zekere hoogte al wel in de bestaande wetgeving zijn opgenomen. Zie met betrekking daartoe punt 21 en verder.

2 Zie de artikelen 24 tot en met 27 van het Handvest van de grondrechten van de EU (het Handvest), waarin wordt ingegaan op de rechten van kinderen en ouderen, de integratie van mensen met een beperking en de rechten van werknemers. Zie ook artikel 38 over de bescherming van consumenten.

Essentiële richtsnoeren uit hoofdstuk II:

- ✓ Zorg ervoor dat de ontwikkeling, de installatie en het gebruik van KI-systemen voldoen aan de vereisten voor betrouwbare KI: 1) menselijke controle en menselijk toezicht, 2) technische robuustheid en veiligheid, 3) privacy en datagovernance, 4) transparantie, 5) diversiteit, non-discriminatie en rechtvaardigheid, 6) milieu- en maatschappelijk welzijn en 7) verantwoordingsplicht.
- ✓ Bestudeer technische en niet-technische methoden om te waarborgen dat die vereisten worden verwezenlijkt.
- ✓ Stimuleer onderzoek en innovatie om bij te dragen aan het controleren van KI-systemen en het verwezenlijken van de vereisten te bevorderen. Verspreid de resultaten en open vragen onder een breder publiek en leid systematisch een nieuwe generatie deskundigen op het gebied van KI-ethiek op.
- ✓ Verspreid informatie over de mogelijkheden en beperkingen van een KI-systeem en over de manier waarop de vereisten worden verwezenlijkt, op een heldere en proactieve manier onder belanghebbenden om onrealistische verwachtingen te voorkomen. Wees open over het feit dat ze met een KI-systeem te maken hebben.
- ✓ Faciliteer de traceerbaarheid en controleerbaarheid van KI-systemen, met name in kritieke omgevingen of situaties.
- ✓ Betrek belanghebbenden bij de volledige levenscyclus van het KI-systeem. Stimuleer onderwijs en opleiding, zodat alle belanghebbenden zich bewust zijn van betrouwbare KI en opgeleid zijn op dat gebied.
- ✓ Wees bedacht op conflicterende beginselen en vereisten. Zoek, evalueer en documenteer deze afwegingen en de oplossingen voortdurend en maak ze kenbaar.

III. In hoofdstuk III wordt een concrete en niet-uitputtende controlelijst voor betrouwbare KI gegeven, die gericht is op de operationalisering van de vereisten uit hoofdstuk II. De **controlelijst** moet worden afgestemd op de specifieke gebruikssituatie van het KI-systeem[3].

Essentiële richtsnoeren uit hoofdstuk III:

- ✓ Stel een controlelijst voor betrouwbare KI vast bij de ontwikkeling, de installatie of het gebruik van KI-systemen en stem deze af op de specifieke situatie waarin het systeem wordt gebruikt.
- ✓ Bedenk dat een dergelijke controlelijst nooit uitputtend zal zijn. Bij betrouwbare KI gaat het niet om het afvinken van elementen op een lijst, maar om het voortdurend vaststellen en verwezenlijken van vereisten, afwegen van oplossingen en waarborgen van betere resultaten gedurende de hele levenscyclus van het KI-systeem, en om het betrekken van de belanghebbenden bij dit proces.

(3) Het laatste gedeelte van het document is bedoeld om een aantal van de binnen het kader besproken kwesties concreet te maken door voorbeelden te geven van voordelige mogelijkheden die moeten worden nagestreefd, en punten van zorg die KI-systemen met zich meebrengen en die zorgvuldig moeten worden afgewogen.

(4) Hoewel deze richtsnoeren bedoeld zijn als ondersteuning voor KI-toepassingen in het algemeen door een horizontale grondslag te creëren voor de bewerkstelliging van betrouwbare KI, brengen verschillende situaties verschillende uitdagingen met zich mee. Daarom moet worden onderzocht of er, gezien de contextgebondenheid van KI-systemen, naast dit horizontale kader ook een sectorale benadering nodig is.

(5) Het is niet de bedoeling dat deze richtsnoeren in de plaats komen van beleid of regelgeving, huidig of toekomstig en in welke vorm dan ook, noch dat zij de invoering daarvan ontmoedigen. Ze moeten worden

[3] Overeenkomstig de bij punt 2 beschreven reikwijdte van het kader biedt deze controlelijst geen advies over het naleven van de wetgeving (wettige KI), maar uitsluitend richtsnoeren voor het verwezenlijken van de tweede en derde component van betrouwbare KI (ethische en robuuste KI).

gezien als een dynamisch document, dat periodiek moet worden herzien en bijgewerkt om relevant te blijven naarmate de technologie, onze sociale omgeving en onze kennis zich ontwikkelen. Dit document wordt beschouwd als een uitgangspunt voor het debat over "betrouwbare KI voor Europa"[4]. Buiten Europa zijn deze richtsnoeren bedoeld om het onderzoek naar, de bezinning op en het debat over een ethisch kader voor KI-systemen op mondiaal niveau te stimuleren.

[4] Het is de bedoeling dat dit ideaal van toepassing is op KI-systemen die in de EU-lidstaten worden ontwikkeld, geïnstalleerd en gebruikt, alsook op systemen die elders worden ontwikkeld of geproduceerd, maar die in de EU worden geïnstalleerd en gebruikt. Wanneer er in dit document naar "Europa" wordt verwezen, worden daarmee de EU-lidstaten bedoeld. Het streven is echter dat deze richtsnoeren ook buiten de EU relevant zijn. Met betrekking daartoe kan ook worden opgemerkt dat zowel Noorwegen als Zwitserland deel uitmaakt van het Gecoördineerd plan inzake KI, dat in december 2018 door de Commissie en de lidstaten is overeengekomen en bekendgemaakt.

A. INLEIDING

(6) In haar mededelingen van 25 april 2018 en 7 december 2018 heeft de Europese Commissie (de Commissie) haar visie op kunstmatige intelligentie (KI) uiteengezet, waarin "ethische, veilige en geavanceerde KI in Europa" wordt ondersteund[5]. De visie van de Commissie rust op drie pijlers: i) het verhogen van publieke en particuliere investeringen in KI om het gebruik ervan te stimuleren, ii) het anticiperen op sociaaleconomische veranderingen en iii) het waarborgen van een passend ethisch en juridisch kader ter versterking van Europese waarden.

(7) Om de uitvoering van deze visie te ondersteunen heeft de Commissie de deskundigengroep op hoog niveau inzake kunstmatige intelligentie (AI HLEG) opgericht, een onafhankelijke groep belast met het opstellen van twee documenten: 1) ethische richtsnoeren voor KI en 2) aanbevelingen voor beleid en investeringen.

(8) Dit document bevat de ethische richtsnoeren voor KI, die zijn herzien na verdere besprekingen binnen onze groep in het kader van de bij de openbare raadpleging ontvangen feedback op de ontwerpversie die op 18 december 2018 was gepubliceerd. In de richtsnoeren wordt voortgebouwd op de werkzaamheden van de Europese Groep ethiek van de exacte wetenschappen en de nieuwe technologieën[6] en wordt uitgegaan van andere soortgelijke inspanningen[7].

(9) De afgelopen maanden hebben de 52 leden van de deskundigengroep vergaderd, gediscussieerd en contact gehad, met het oog op het Europese motto: in verscheidenheid verenigd. Wij zijn ervan overtuigd dat KI de potentie heeft om de samenleving sterk te veranderen. KI is geen doel, maar veeleer een veelbelovend middel om de mens tot bloei te laten komen en zo het individuele en maatschappelijke welzijn te verbeteren, het algemeen belang te dienen en vooruitgang en innovatie teweeg te brengen. In het bijzonder kunnen KI-systemen bijdragen aan het verwezenlijken van de duurzameontwikkelingsdoelstellingen van de VN, zoals het bevorderen van het genderevenwicht, het aanpakken van de klimaatverandering, het rationaliseren van ons gebruik van natuurlijke hulpbronnen, het verbeteren van onze gezondheid, mobiliteit en productieprocessen en het ondersteunen van de manier waarop we de voortgang bijhouden aan de hand van indicatoren van duurzaamheid en sociale cohesie.

(10) Daarvoor moet bij KI-systemen[8] **de mens centraal staan** en moet ernaar worden gestreefd deze systemen te gebruiken in dienst van de mensheid en het algemeen belang, met als doel de verbetering van het welzijn en de vrijheid van de mens. KI-systemen bieden geweldige mogelijkheden, maar brengen ook bepaalde risico's met zich mee, waar passend en in proportie mee moet worden omgegaan. Er ligt nu een belangrijke kans om richting te geven aan de ontwikkeling van deze systemen. Wij willen ervoor zorgen dat we de sociaaleconomische situaties waarin ze worden ingebed, kunnen vertrouwen, en wij willen dat de producenten van de KI-systemen een concurrentievoordeel verkrijgen door betrouwbare KI in hun producten en diensten te verwerken. Hiervoor is het nodig **de voordelen van KI-systemen te maximaliseren** en tegelijk **de risico's ervan te voorkomen en tot een minimum te beperken**.

(11) Wij zijn ervan overtuigd dat het in een context van snelle technologische verandering essentieel is dat vertrouwen het cement blijft van samenlevingen, gemeenschappen, economieën en duurzame ontwikkeling. Daarom kiezen wij **betrouwbare KI als onze fundamentele ambitie,** want mensen en gemeenschappen zullen

---

5 COM(2018) 237 final en COM(2018) 795 final. Merk op dat de term "in Europa" ("made in Europe") in de hele mededeling van de Commissie wordt gebruikt. Deze richtsnoeren beogen echter om niet alleen de in Europa gemaakte KI-systemen te omvatten, maar ook de systemen die elders worden ontwikkeld en in Europa worden geïnstalleerd of gebruikt. In dit document streven we er steeds naar betrouwbare KI "voor" Europa te stimuleren.

6 De Europese Groep ethiek van de exacte wetenschappen en de nieuwe technologieën (EGE) is een adviesgroep van de Commissie.

7 Zie punt 3.3 van COM(2018) 237.

8 In de woordenlijst aan het einde van dit document wordt een definitie van KI-systemen gegeven zoals gebruikt binnen dit document. Deze definitie wordt verder uitgewerkt in een speciaal document dat door de AI HLEG is opgesteld en dat bij deze richtsnoeren is gevoegd, getiteld "Een definitie van KI: de belangrijkste capaciteiten en wetenschappelijke disciplines".

alleen vertrouwen kunnen hebben in de technologische ontwikkelingen en de toepassingen daarvan als er een helder en volledig kader bestaat om de betrouwbaarheid ervan tot stand te brengen.

(12) Dat is in onze optiek de manier waarop Europa zich kan positioneren als thuisbasis en voortrekker voor geavanceerde en ethische technologie. Wij, als Europese burgers, zullen de vruchten plukken van betrouwbare KI op een manier die strookt met onze basiswaarden van respect voor mensenrechten, democratie en de rechtsstaat.

Betrouwbare KI

(13) Betrouwbaarheid is een randvoorwaarde voor mensen en samenlevingen om KI-systemen te kunnen ontwikkelen, installeren en gebruiken. Als KI-systemen – en de mensen erachter – niet aantoonbaar te vertrouwen zijn, kan dit ongewilde gevolgen hebben en kan het gebruik ervan worden belemmerd, waardoor de potentieel enorme sociale en economische voordelen van KI-systemen niet kunnen worden verwezenlijkt. Om Europa te helpen deze voordelen te verwezenlijken is onze visie het gebruik van ethiek als kernpijler om betrouwbare KI te garanderen en af te stemmen.

(14) Vertrouwen in de ontwikkeling, de installatie en het gebruik van KI-systemen heeft niet alleen te maken met de inherente kenmerken van de technologie, maar ook met de eigenschappen van de sociaal-technische systemen waarvoor KI-systemen worden gebruikt[9]. Net als bij (verlies van) vertrouwen in de luchtvaart, kernenergie of voedselveiligheid zijn het niet slechts componenten van het KI-systeem die al dan niet vertrouwen wekken, maar het systeem in de algehele context. Streven naar betrouwbare KI heeft niet alleen te maken met de betrouwbaarheid van het KI-systeem zelf, maar vereist een holistische en systemische benadering, waarbij de betrouwbaarheid van alle actoren en processen die gedurende de volledige levenscyclus van het systeem deel uitmaken van de sociaal-technische context ervan, wordt meegenomen.

(15) Betrouwbare KI bestaat uit **drie componenten,** waaraan gedurende de volledige levenscyclus van het systeem moet worden voldaan:

1. de KI moet **wettig** zijn, door te zorgen dat aan alle toepasselijke wet- en regelgeving wordt voldaan;
2. de KI moet **ethisch** zijn, door te zorgen dat de ethische beginselen en waarden worden nageleefd; en
3. de KI moet **robuust** zijn, uit zowel technisch als sociaal oogpunt, aangezien KI-systemen ongewild schade kunnen aanrichten, zelfs al zijn de bedoelingen goed.

(16) Elk van deze drie componenten is nodig, maar op zichzelf niet voldoende om betrouwbare KI te bewerkstelligen[10]. In het ideale scenario sluiten ze alle drie op elkaar aan en valt de werking ervan gedeeltelijk samen. In de praktijk kunnen er echter spanningen bestaan tussen deze elementen (soms kunnen bijvoorbeeld het toepassingsgebied en de inhoud van de bestaande wetgeving niet in lijn zijn met de ethische normen). Als samenleving hebben wij de individuele en collectieve verantwoordelijkheid ervoor te zorgen dat alle drie de componenten bijdragen aan de totstandbrenging van betrouwbare KI[11].

(17) Een betrouwbare benadering is essentieel om "verantwoordelijke concurrentie" mogelijk te maken. Hiermee wordt de grondslag geboden op basis waarvan iedereen die met KI-systemen te maken heeft, erop kan vertrouwen dat het ontwerp, de ontwikkeling en het gebruik ervan wettig, ethisch en robuust zijn. Met deze richtsnoeren, die zijn bedoeld om verantwoordelijke en duurzame KI-innovatie in Europa te stimuleren, wordt beoogd van ethiek een kernpijler te maken op basis waarvan een unieke benadering van KI tot stand komt, een benadering die erop gericht is de ontwikkeling van zowel individuele personen als het algemeen belang van de samenleving te bevorderen, versterken en beschermen. Wij zijn ervan overtuigd dat Europa zich

9 Deze systemen omvatten mensen, staatsactoren, ondernemingen, infrastructuur, software, protocollen, normen, governance, bestaande wetten, toezichtsmechanismen, stimuleringsstructuren, controleprocedures, beste verslagleggingspraktijken en meer.

10 Dit sluit niet uit dat aanvullende voorwaarden noodzakelijk kunnen zijn/worden.

11 Dit kan ook betekenen dat de wetgevers of beleidsmakers de geschiktheid van de bestaande wetgeving moeten onderzoeken, indien deze mogelijk niet meer in lijn is met de ethische beginselen.

hierdoor zal kunnen profileren als mondiale voortrekker op het gebied van geavanceerde KI die ons individuele en collectieve vertrouwen verdient. Alleen door betrouwbaarheid te garanderen zullen Europeanen ten volle kunnen profiteren van de voordelen van KI-systemen, in de wetenschap dat er maatregelen zijn getroffen om hen tegen de potentiële gevaren ervan te beschermen.

(18) Zoals het gebruik van KI-systemen niet ophoudt bij de nationale grens, geldt dit ook voor de gevolgen ervan. Daarom zijn er mondiale oplossingen nodig voor de mondiale kansen en uitdagingen die KI met zich meebrengt. Wij moedigen alle belanghebbenden dan ook aan om te werken aan een mondiaal kader voor betrouwbare KI door internationale consensus te creëren en tegelijk onze op grondrechten gebaseerde benadering te bevorderen en te handhaven.

Doelpubliek en toepassingsgebied

(19) Deze richtsnoeren zijn gericht aan alle KI-belanghebbenden die KI ontwerpen, ontwikkelen, installeren, toepassen, gebruiken of erdoor worden beïnvloed, met inbegrip van onder meer bedrijven, organisaties, onderzoekers, overheidsdiensten, overheidsinstanties, instellingen, maatschappelijke organisaties, individuele personen, werknemers en consumenten. Belanghebbenden die zich inzetten voor de totstandbrenging van betrouwbare KI, kunnen vrijwillig kiezen om deze richtlijnen te gebruiken als methode voor de operationalisering van hun inzet, met name door de praktische controlelijst uit hoofdstuk III te gebruiken bij het ontwikkelings- en installatieproces van hun KI-systemen. Deze controlelijst kan ook een aanvulling vormen op – en dus worden verwerkt in – bestaande controleprocessen.

(20) Deze richtsnoeren zijn bedoeld als ondersteuning voor KI-toepassingen in het algemeen waarbij wordt voorzien in een horizontale grondslag voor de totstandbrenging van betrouwbare KI. **Verschillende situaties brengen echter verschillende uitdagingen met zich mee.** KI-aanbevelingssystemen voor muziek leveren niet dezelfde ethische vragen op als KI-systemen waarin kritieke medische behandelingen worden voorgesteld. Het gebruik van KI-systemen binnen een relatie tussen bedrijf en consument, tussen bedrijven, tussen werkgever en werknemer of tussen overheid en burger – of, meer in het algemeen: in verschillende sectoren of situaties – levert eveneens verschillende kansen en uitdagingen op. Gezien de contextgebondenheid van KI-systemen wordt daarom erkend dat de uitvoering van deze richtsnoeren op de betreffende KI-toepassing moet worden afgestemd. Bovendien moet de noodzaak van een aanvullende sectorale benadering als aanvulling op het algemenere horizontale kader dat in dit document wordt voorgesteld, worden onderzocht.

Om meer inzicht te krijgen in de manier waarop deze richtsnoeren op horizontaal niveau kunnen worden uitgevoerd, alsook in de zaken waarvoor een sectorale benadering nodig is, nodigen we alle belanghebbenden uit om de controlelijst voor betrouwbare KI (hoofdstuk III), waarin dit kader wordt geoperationaliseerd, te testen en ons feedback te geven. Op basis van de middels deze testfase verzamelde feedback zullen wij de controlelijst uit deze richtsnoeren per begin 2020 herzien. De testfase gaat in de zomer van 2019 van start en duurt tot het einde van het jaar. Alle geïnteresseerde belanghebbenden kunnen deelnemen door hun belangstelling kenbaar te maken via de Europese KI-alliantie.

B. EEN KADER VOOR BETROUWBARE KI

(21) In deze richtsnoeren wordt een kader uiteengezet voor de totstandbrenging van betrouwbare KI op basis van grondrechten zoals vervat in het Handvest van de grondrechten van de Europese Unie (het Handvest) en in de relevante internationale mensenrechtenwetgeving. Hieronder bespreken we kort de drie componenten van betrouwbare KI.

Wettige KI

(22) KI-systemen zijn niet werkzaam binnen een wetteloze wereld. Een aantal juridisch bindende regels op Europees, nationaal en internationaal niveau is al van toepassing op of relevant voor de ontwikkeling, de

installatie en het gebruik van KI-systemen. Relevante wettelijke bronnen zijn onder meer het primaire EU-recht (de verdragen van de Europese Unie en het Handvest van de grondrechten van de Europese Unie) en het secundaire EU-recht (zoals de algemene verordening gegevensbescherming (AVG), anti-discriminatierichtlijnen, de machinerichtlijn, de richtlijn inzake productaansprakelijkheid, de verordening betreffende het vrije verkeer van niet-persoonsgebonden gegevens, het consumentenrecht en de richtlijnen inzake veiligheid en gezondheid op het werk), maar ook de mensenrechtenverdragen van de VN en de verdragen van de Raad van Europa (zoals het Europees Verdrag tot bescherming van de rechten van de mens) en allerlei wetgeving van de EU-lidstaten. Naast horizontaal toepasselijke regels bestaan er ook verschillende domeinspecifieke regels die van toepassing zijn op specifieke KI-toepassingen (zoals de verordening betreffende medische hulpmiddelen in de gezondheidszorgsector).

(23) De wetgeving bevat zowel positieve als negatieve verplichtingen. Dat betekent dat deze niet alleen moet worden uitgelegd in verband met wat er *niet mag*, maar ook in verband met wat er *moet*. In de wetgeving worden niet alleen bepaalde handelingen verboden, maar ook andere mogelijk gemaakt. Met betrekking daartoe kan worden opgemerkt dat het Handvest artikelen bevat over de "vrijheid van ondernemerschap" en de "vrijheid van kunsten en wetenschappen", alsook artikelen waarin wordt ingegaan op gebieden waarmee wij beter bekend zijn als we de betrouwbaarheid van KI proberen te waarborgen, zoals gegevensbescherming en non-discriminatie.

(24) In deze richtsnoeren wordt niet expliciet ingegaan op de eerste component van betrouwbare KI (wettige KI). Er wordt veeleer beoogd richtsnoeren te bieden voor het bevorderen en waarborgen van de tweede en derde component (ethische en robuuste KI). Hoewel die laatste twee tot op zekere hoogte al in de bestaande wetgeving zijn opgenomen, kan de volledige realisatie ervan verder gaan dan de bestaande wettelijke verplichtingen.

(25) Niets in dit document mag worden uitgelegd of opgevat als juridisch advies of richtsnoeren voor de manier waarop toepasselijke bestaande wettelijke normen en voorschriften kunnen worden nageleefd. Door niets in dit document worden aan derden wettelijke rechten toegekend of wettelijke verplichtingen opgelegd. Wij wijzen er echter op dat het de plicht van iedere natuurlijke of rechtspersoon is om de wetgeving na te leven – ongeacht of die op dit moment al geldt of in de toekomst wordt vastgesteld op basis van de ontwikkeling van KI. Voor het vervolg van deze richtsnoeren wordt verondersteld dat **alle wettelijke rechten en verplichtingen die van toepassing zijn op de processen en activiteiten in verband met het ontwikkelen, installeren en gebruiken van KI, blijven gelden en in acht moeten worden genomen.**

Ethische KI

(26) Voor de verwezenlijking van betrouwbare KI is naleving van de wetgeving slechts een van de drie benodigde componenten. De wetgeving is niet altijd actueel in verband met technologische ontwikkelingen, komt soms niet overeen met ethische normen of kan simpelweg niet geschikt zijn voor bepaalde kwesties. Om betrouwbaar te zijn moeten KI-systemen daarom ook ethisch zijn, door afstemming op ethische normen te waarborgen.

Robuuste KI

(27) Zelfs als een ethisch doel wordt gewaarborgd, moeten individuele personen en de samenleving ervan overtuigd zijn dat KI-systemen niet ongewild schade zullen aanrichten. Deze systemen moeten op een veilige, zekere en betrouwbare manier werken en er moeten voorzorgsmaatregelen worden getroffen om onbedoelde negatieve gevolgen te voorkomen. Het is dus belangrijk om ervoor te zorgen dat KI-systemen robuust zijn. Dit is zowel vanuit technisch oogpunt (zorgen voor de technische robuustheid van het systeem zoals gepast in een bepaalde context, bijvoorbeeld het toepassingsgebied of de fase van de levenscyclus) als vanuit sociaal oogpunt (rekening houdend met de context en omgeving waarin het systeem werkzaam is) nodig. Ethische en robuuste KI zijn dus nauw met elkaar verweven en vullen elkaar aan. De in hoofdstuk I beschreven beginselen en de daaruit in hoofdstuk II afgeleide vereisten hebben betrekking op beide componenten.

Het kader

(28) De richtsnoeren in dit document worden in drie lagen met verschillend abstractieniveau gegeven: van het meest abstract in hoofdstuk I tot het meest concreet in hoofdstuk III:

I) Grondslagen van betrouwbare KI. In hoofdstuk I worden de grondslagen van betrouwbare KI en de benadering op basis van grondrechten[12] uiteengezet. De ethische beginselen die moeten worden gevolgd om ethische en robuuste KI te waarborgen, worden vastgesteld en beschreven.

II) Betrouwbare KI verwezenlijken. In hoofdstuk II worden deze ethische beginselen vertaald naar zeven vereisten die bij KI-systemen moeten worden toegepast en waaraan ze gedurende hun volledige levenscyclus moeten voldoen. Ook worden er zowel technische als niet-technische methoden aangereikt die voor de toepassing ervan kunnen worden gebruikt.

III) Betrouwbare KI controleren. Beroepsbeoefenaars op het gebied van KI verwachten concrete richtsnoeren. Daarom wordt in hoofdstuk III een voorlopige en niet-uitputtende controlelijst voor betrouwbare KI gegeven waarmee de vereisten uit hoofdstuk II kunnen worden geoperationaliseerd. Deze controle moet worden afgestemd op de toepassing van het specifieke systeem.

(29) In het laatste gedeelte van het document worden voordelige kansen en punten van zorg in verband met KI-systemen uiteengezet die in overweging moeten worden genomen en waarover we verdere discussies willen stimuleren.

(30) De structuur van de richtsnoeren wordt hieronder in figuur 1 geïllustreerd.

[12] Grondrechten liggen ten grondslag aan zowel internationale als EU-mensenrechtenwetgeving en onderbouwen de juridisch afdwingbare rechten die door de EU-verdragen en het Handvest van de grondrechten van de EU worden gegarandeerd. Omdat ze juridisch bindend zijn, valt naleving van de grondrechten onder de eerste component van betrouwbare KI: "wettige KI". Grondrechten kunnen echter ook worden begrepen als een weergave van bijzondere morele rechten die alle personen hebben op grond van hun menselijkheid, ongeacht de juridisch bindende status ervan. In die zin zijn ze ook onderdeel van de tweede component van betrouwbare KI: "ethische KI".

Figuur 1: De richtsnoeren als een kader voor betrouwbare KI

I. Hoofdstuk I: Grondslagen van betrouwbare KI

(31) In dit hoofdstuk worden de grondslagen van betrouwbare KI uiteengezet, die gebaseerd zijn op de grondrechten en worden weergegeven in vier ethische beginselen die moeten worden gevolgd om ethische en robuuste KI te waarborgen. Dit hoofdstuk is sterk gebaseerd op het domein van de ethiek.

(32) KI-ethiek is een subdomein van toegepaste ethiek dat gericht is op de ethische kwesties rondom de ontwikkeling, de installatie en het gebruik van KI. Het belangrijkste is om te bepalen hoe KI het leven van mensen kan verbeteren of zorgen kan opleveren, of het nu gaat om de levenskwaliteit of om de menselijke autonomie en vrijheid die nodig zijn voor een democratische samenleving.

(33) Ethische reflectie op KI-technologie kan meerdere doelen hebben. Ten eerste kan hiermee worden gestimuleerd dat er wordt nagedacht over de noodzaak om personen en groepen op het meest basale niveau te beschermen. Ten tweede kunnen hiermee nieuwe soorten innovatie worden gestimuleerd waarmee ethische waarden worden bevorderd, zoals innovaties die bijdragen aan de verwezenlijking van de duurzameontwikkelingsdoelstellingen van de VN[13], die stevig verankerd zijn in de komende EU-agenda voor 2030[14]. Hoewel dit document voornamelijk gaat over het eerste van deze doelen, moet het mogelijke belang van ethiek voor het tweede doel niet worden onderschat. Door betrouwbare KI kunnen mensen sterker tot bloei komen en kan het collectieve welzijn worden verbeterd, doordat er welvaart, waardecreatie en vermogensmaximalisatie worden gegenereerd. Betrouwbare KI kan bijdragen aan de verwezenlijking van een rechtvaardige samenleving door de verbetering van de gezondheid en het welzijn van burgers zo te stimuleren dat gelijkheid bij de verdeling van economische, maatschappelijke en politieke kansen wordt bevorderd.

(34) Daarom is het noodzakelijk dat we begrijpen hoe we de ontwikkeling, de installatie en het gebruik van KI het best kunnen ondersteunen als we willen zorgen dat iedereen goed gedijt in een op KI gebaseerde wereld, en als we een betere toekomst willen creëren, maar ook wereldwijd willen blijven concurreren. Zoals elke krachtige technologie brengt ook het gebruik van KI-systemen in onze samenleving een aantal ethische uitdagingen met zich mee, bijvoorbeeld in verband met het effect ervan op mensen en de samenleving, de capaciteit voor besluitvorming en de veiligheid. Als we steeds vaker de hulp gaan inroepen van of beslissingen gaan uitbesteden aan KI-systemen, moeten we zorgen dat deze systemen rechtvaardig zijn wat hun effect op het leven van mensen betreft, dat ze in overeenstemming zijn met onbetwistbare waarden en in staat in overeenstemming daarmee te functioneren, en dat er geschikte verantwoordingsprocessen zijn die daarvoor zorgen.

(35) Europa moet bepalen welke normatieve visie op een toekomst vol KI het wil realiseren, en vervolgens welke definitie van KI er in Europa moet worden onderzocht, ontwikkeld, geïnstalleerd en gebruikt om deze visie te verwezenlijken. Met dit document willen wij daaraan bijdragen door het begrip "betrouwbare KI" te introduceren, dat in onze optiek de juiste manier vormt om te bouwen aan een toekomst met KI. In een toekomst waarin democratie, de rechtsstaat en grondrechten de basis vormen voor KI-systemen en waarin die systemen de democratische cultuur voortdurend verbeteren en beschermen, zal ook een omgeving kunnen worden gecreëerd waarin innovatie en verantwoordelijke concurrentie kunnen gedijen.

(36) Een domeinspecifieke ethische code – hoe consistent, ontwikkeld en verfijnd toekomstige versies ervan ook mogen zijn – kan nooit dienen als vervanging van ethisch denken, aangezien dit altijd gevoelig moet blijven voor contextuele details die niet in algemene richtsnoeren kunnen worden vastgelegd. Voor betrouwbare KI moeten we niet alleen een reeks voorschriften ontwikkelen, maar ook een ethische cultuur en houding creëren en onderhouden door middel van openbare discussies, onderwijs en praktische vormen van leren.

1. Grondrechten als morele en wettelijke rechten

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_nl

14 https://sustainabledevelopment.un.org/?menu=1300

(37) Wij geloven in een benadering van KI-ethiek op basis van de grondrechten die zijn vastgelegd in de EU-verdragen[15], het Handvest van de grondrechten van de EU (het Handvest) en de internationale mensenrechtenwetgeving[16]. Eerbiediging van grondrechten, binnen een kader van democratie en de rechtsstaat, biedt de meest veelbelovende grondslagen voor het vaststellen van abstracte ethische beginselen en waarden die in het kader van KI kunnen worden geoperationaliseerd.

(38) In de EU-verdragen en het Handvest wordt een reeks grondrechten voorgeschreven. De lidstaten en instellingen van de EU zijn wettelijk verplicht deze te eerbiedigen bij de uitvoering van EU-wetgeving. Deze rechten worden in het Handvest beschreven op basis van waardigheid, vrijheden, gelijkheid en solidariteit, de rechten van burgers en gerechtigheid. Van de gemeenschappelijke grondslag van deze rechten kan men zeggen dat deze is geworteld in respect voor de menselijke waardigheid, waarmee een benadering wordt weergegeven waarbij de mens centraal staat doordat deze een uniek en onvervreemdbaar moreel primaat heeft binnen het maatschappelijk, politieke, economische en sociale domein[17].

(39) Hoewel de in het Handvest beschreven rechten juridisch bindend zijn[18], is het belangrijk om te erkennen dat grondrechten niet in alle gevallen volledige wettelijke bescherming bieden. Voor het Handvest is het bijvoorbeeld belangrijk om te benadrukken dat het toepassingsgebied ervan beperkt is tot bepaalde domeinen van de EU-wetgeving. De internationale mensenrechtenwetgeving en in het bijzonder het Europees Verdrag tot bescherming van de rechten van de mens zijn juridisch bindend voor de EU-lidstaten, ook op gebieden die buiten het toepassingsgebied van de EU-wetgeving vallen. Tegelijkertijd moet er worden benadrukt dat grondrechten ook worden verleend aan personen en (tot op zekere hoogte) aan groepen op grond van hun morele status als mens, los van de rechtskracht ervan. Opgevat als juridisch afdwingbare rechten vallen grondrechten daarom onder de eerste component van betrouwbare KI (wettige KI), die de naleving van de wetgeving waarborgt. Opgevat als de rechten van alle mensen, geworteld in de inherente morele status van mensen, ondersteunen ze ook de tweede component van betrouwbare KI (ethische KI), die te maken heeft met ethische normen die niet per definitie juridisch bindend zijn, maar wel cruciaal voor betrouwbaarheid. Omdat dit document niet is bedoeld als richtsnoer voor de eerste component, hebben verwijzingen naar grondrechten in het kader van deze niet-bindende richtsnoeren betrekking op de tweede component.

2. Van grondrechten naar ethische beginselen

2.1 Grondrechten als een basis voor betrouwbare KI

(40) Van de volledige reeks in de internationale mensenrechtenwetgeving, de EU-verdragen en het Handvest beschreven ondeelbare rechten zijn de onderstaande groepen grondrechten in het bijzonder geschikt om toe te passen op KI-systemen. Veel van deze rechten zijn, in bepaalde omstandigheden, juridisch afdwingbaar in de EU, dus naleving van de voorwaarden ervan is wettelijk verplicht. Zelfs wanneer naleving van de juridisch afdwingbare grondrechten is verwezenlijkt, kan ethische reflectie ons helpen te begrijpen welke rol grondrechten en de onderliggende waarden daarvan spelen bij de ontwikkeling, de installatie en het gebruik van KI. Ook kan deze reflectie verfijndere richtsnoeren opleveren om te bepalen wat we met technologie *zouden moeten* doen in plaats van wat we er (momenteel) mee *kunnen* doen.

(41) ***Respect voor de menselijke waardigheid***. Onder menselijke waardigheid valt het idee dat ieder mens een

[15] De EU is gebaseerd op een constitutionele verbintenis om de grondrechten en ondeelbare rechten van mensen te beschermen, eerbiediging van de rechtsstaat te garanderen, democratische vrijheid te waarborgen en het algemeen belang te bevorderen. Deze rechten zijn terug te vinden in de artikelen 2 en 3 van het Verdrag betreffende de Europese Unie en in het Handvest van de grondrechten van de EU.

[16] In andere juridische instrumenten worden deze verbintenissen weergegeven en verder uitgewerkt, bijvoorbeeld in het Europees Sociaal Handvest van de Raad van Europa en in specifieke wetgeving zoals de algemene verordening gegevensbescherming van de EU.

[17] Er moet worden opgemerkt dat voor toewijding aan KI waarbij de mens centraal staat, en voor de verankering daarvan in grondrechten, collectieve maatschappelijke en constitutionele grondslagen nodig zijn waarbij individuele vrijheid en respect voor de menselijke waardigheid zowel praktisch mogelijk als zinvol zijn en er geen onnodig individualistisch beeld van de mens wordt geschetst.

[18] Het Handvest is, op grond van artikel 51 ervan, van toepassing op de EU-instellingen en EU-lidstaten bij de uitvoering van EU-wetgeving.

"intrinsieke waarde" heeft die nooit door anderen – of door nieuwe technologieën als KI-systemen – mag worden beperkt, in gevaar worden gebracht of onderdrukt.[19] In het kader van KI houdt respect voor de menselijke waardigheid in dat alle mensen worden behandeld met het respect waar zij als morele *subjecten*, niet slechts *objecten* die kunnen worden ontleed, gesorteerd, becijferd, gedreven, geconditioneerd of gemanipuleerd, recht op hebben. KI-systemen moeten dus worden ontwikkeld op een manier waarbij de fysieke en geestelijke integriteit van mensen, hun persoonlijke en culturele gevoel van identiteit en de vervulling van hun basisbehoeften worden gerespecteerd, gediend en beschermd[20].

(42) ***Vrijheid van het individu.*** Mensen moeten vrij zijn om hun eigen levenskeuzen te maken. Dat betekent vrijheid van soevereine inmenging, maar er is ook ingrijpen voor nodig door de overheid en niet-gouvernementele organisaties, om te zorgen dat personen die het risico lopen te worden uitgesloten, gelijke toegang hebben tot de voordelen en mogelijkheden van KI. In het kader van KI is voor vrijheid van het individu beperking nodig van (in)directe illegale dwang, bedreigingen voor de mentale autonomie en geestelijke gezondheid, ongerechtvaardigd toezicht, misleiding en oneerlijke manipulatie. Sterker nog, vrijheid van het individu betekent een verbintenis om mensen nóg meer controle over hun eigen leven te geven, met (onder andere) rechten als bescherming van de vrijheid van ondernemerschap, de vrijheid van kunsten en wetenschappen, de vrijheid van meningsuiting, het recht op een privéleven en privacy en de vrijheid van vergadering en vereniging**.**

(43) ***Respect voor democratie, justitie en de rechtsstaat****.* Alle macht van de overheid in constitutionele democratieën moet wettelijk worden toegekend en beperkt. KI-systemen moeten democratische processen onderhouden en bevorderen en de pluraliteit van de waarden en levenskeuzen van mensen respecteren. Ze mogen democratische processen, menselijk overleg of democratische kiesstelsels niet ondermijnen. Ook moet in KI-systemen een verbintenis worden verwerkt om te zorgen dat ze niet op manieren werken waardoor op de rechtsstaat berustende basisbeginselen en de toepasselijke wet- en regelgeving worden ondermijnd en dat een eerlijke procesgang en gelijkheid voor de wet worden gewaarborgd.

(44) ***Gelijkheid, non-discriminatie en solidariteit*** *– inclusief de rechten van mensen die het risico lopen te worden uitgesloten.* Gelijk respect voor de morele waarde en waardigheid van alle mensen moet worden gewaarborgd. Dat gaat verder dan non-discriminatie, waarbij het wordt getolereerd verschil te maken tussen ongelijke situaties op basis van objectieve rechtvaardigingen. In het kader van KI betekent gelijkheid dat de activiteiten van het systeem geen vertekende resultaten mogen opleveren (de gegevens die worden gebruikt om KI-systemen te trainen, moeten bijvoorbeeld zo inclusief mogelijk zijn, zodat verschillende bevolkingsgroepen worden vertegenwoordigd). Hiervoor is ook passend respect nodig voor potentieel kwetsbare personen en groepen[21], zoals werknemers, vrouwen, mensen met een beperking, etnische minderheden, kinderen, consumenten of anderen die het risico lopen te worden uitgesloten.

(45) ***Rechten van burgers***. Burgers hebben allerlei rechten, waaronder kiesrecht, het recht op behoorlijk bestuur, het recht op inzage in openbare documenten en het recht een verzoekschrift tot de overheid te richten. KI-systemen bieden veel potentieel voor de verbetering van de schaal en efficiëntie van de overheid bij het leveren van publieke goederen en diensten aan de maatschappij. Tegelijk kunnen KI-toepassingen ook negatieve gevolgen hebben voor de rechten van burgers en moeten deze rechten worden beschermd. Wanneer hier de term "rechten van burgers" wordt gebruikt, wordt daarmee niet beoogd de rechten van inwoners van derde landen en van personen die onregelmatig (of illegaal) in de EU verblijven, die ook rechten hebben op grond van internationale wetgeving en dus ook op het gebied van KI, te ontkennen of verwaarlozen.

---

19 C. McCrudden, Human Dignity and Judicial Interpretation of Human Rights, *EJIL*, 19(4), 2008.

20 Zie voor een vergelijkbare interpretatie van "menselijke waardigheid" E. Hilgendorf, Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity, in: D. Grimm, A. Kemmerer, C. Möllers (red.), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, blz. 325 ff.

21 Zie de woordenlijst voor een beschrijving van de term zoals deze in dit document wordt gebruikt.

2.2 Ethische beginselen in het kader van KI-systemen[22]

(46) Veel publieke, particuliere en maatschappelijke organisaties hebben inspiratie voor de creatie van ethische kaders voor KI uit grondrechten gehaald[23]. In de EU heeft de Europese Groep ethiek van de exacte wetenschappen en de nieuwe technologieën (EGE) een reeks van negen basisbeginselen voorgesteld op basis van de fundamentele waarden die zijn vastgelegd in de EU-verdragen en het Handvest van de grondrechten van de EU.[24] Wij bouwen voort op dit werk. Daarbij erkennen we het grootste gedeelte van de tot dusver door verschillende groepen voorgestelde beginselen en verhelderen we de doelen die men met alle beginselen beoogt te stimuleren en ondersteunen. Deze ethische beginselen kunnen inspiratie opleveren voor nieuwe en specifieke regelgevingsinstrumenten, hulp bieden bij de uitleg van grondrechten naarmate onze sociaal-technische omgeving in de loop van de tijd verandert, en richting geven aan de motivering voor de ontwikkeling, het gebruik en de toepassing van KI-systemen – waarbij ze zich dynamisch aanpassen naarmate de samenleving zelf verandert.

(47) KI-systemen moeten het individuele en collectieve welzijn verbeteren. In dit gedeelte worden **vier ethische beginselen** genoemd die zijn geworteld in grondrechten en die moeten worden nageleefd om te zorgen dat KI-systemen op betrouwbare wijze worden ontwikkeld, geïnstalleerd en gebruikt. Ze worden gespecificeerd als **ethische geboden**, en beroepsbeoefenaars op het gebied van KI moeten er altijd naar streven deze na te leven. Zonder een hiërarchie te willen opleggen vermelden we de beginselen hieronder in dezelfde volgorde waarin de grondrechten waarop ze zijn gebaseerd, in het Handvest staan[25].

(48) Het betreft de beginselen van:

(i) respect voor de menselijke autonomie;
(ii) preventie van schade;
(iii) rechtvaardigheid;
(iv) verantwoording.

(49) Veel van deze beginselen zijn al grotendeels weergegeven in de bestaande wettelijke voorschriften die verplicht moeten worden nageleefd, en vallen dus ook binnen de reikwijdte van "wettige KI", de eerste component van betrouwbare KI[26]. Zoals hierboven beschreven gaat naleving van de ethische beginselen, hoewel deze in veel wettelijke verplichtingen worden weerspiegeld, echter verder dan formele naleving van bestaande wetgeving[27].

- <u>Het beginsel van respect voor de menselijke autonomie</u>

(50) De grondrechten waarop de EU is gefundeerd, zijn erop gericht respect voor de vrijheid en autonomie van

---

[22] Deze beginselen zijn ook van toepassing op de ontwikkeling, de installatie en het gebruik van andere technologieën en hebben dus niet specifiek betrekking op KI-systemen. Wij hebben geprobeerd in dit gedeelte de specifieke relevantie ervan in een KI-gerelateerde context uiteen te zetten.

[23] Door op grondrechten te bouwen kan ook de onzekerheid worden beperkt wat regelgeving betreft, aangezien men kan voortbouwen op tientallen jaren van bescherming van grondrechten in de EU. Daardoor ontstaat duidelijkheid, leesbaarheid en voorzienbaarheid.

[24] Recenter heeft de taskforce van AI4People een enquête gehouden over de genoemde EGE-beginselen, alsook over 36 andere ethische beginselen die tot dusver naar voren zijn gebracht, en deze beginselen onder vier overkoepelende beginselen ingedeeld: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines 28(4): blz. 689-707.

[25] Respect voor de menselijke autonomie wordt sterk geassocieerd met het recht op menselijke waardigheid en vrijheid (weergegeven in de artikelen 1 en 6 van het Handvest). De preventie van schade houdt sterk verband met de bescherming van de fysieke of geestelijke integriteit (weergegeven in artikel 3). Rechtvaardigheid is nauw verbonden aan het recht op non-discriminatie, solidariteit en gerechtigheid (weergegeven in artikel 21 en verder). Verantwoording en verantwoordelijkheid zijn nauw verbonden aan de rechten die verband houden met rechtspleging (zoals weergegeven in artikel 47).

[26] Denk bijvoorbeeld aan de AVG of de EU-wetgeving inzake bescherming van consumenten.

[27] Zie voor meer informatie over dit onderwerp bijvoorbeeld L. Floridi, Soft Ethics and the Governance of the Digital, *Philosophy & Technology,* maart 2018, jaargang 31, editie 1, blz. 1–8.

mensen te waarborgen. Mensen die met KI-systemen werken, moeten hun volledige en effectieve zelfbeschikking kunnen behouden en kunnen deelnemen aan het democratische proces. KI-systemen mogen mensen niet onterecht onderwerpen, dwingen, misleiden, manipuleren, conditioneren of drijven. Ze moeten veeleer worden ontworpen om de menselijke cognitieve, sociale en culturele vaardigheden te vergroten, aan te vullen en te versterken. De verdeling van functies tussen mensen en KI-systemen moet gebeuren volgens ontwerpbeginselen waarbij de mens centraal staat, en moet zinvolle mogelijkheden voor menselijke keuze openlaten. Er moet dus worden gezorgd voor menselijk toezicht[28] en menselijke controle op de werkprocessen in KI-systemen. De werksfeer kan door KI-systemen ook fundamenteel veranderen. Deze moet mensen ondersteunen in de werkomgeving en gericht zijn op de creatie van zinvol werk.

- Het beginsel van preventie van schade

(51) KI-systemen mogen geen schade[29] veroorzaken of vergroten of anderszins negatieve gevolgen hebben voor mensen[30]. Dat betekent bescherming van de waardigheid, alsook de geestelijke en fysieke integriteit van mensen. KI-systemen en de omgeving waarin zij werken, moeten veilig en zeker zijn. Ze moeten technisch robuust zijn en er moet voor worden gezorgd dat ze geen ruimte bieden voor kwaadwillig gebruik. Kwetsbare personen moeten meer aandacht krijgen en moeten worden betrokken bij de ontwikkeling en installatie van KI-systemen. Er moet specifiek aandacht worden besteed aan situaties waarin KI-systemen negatieve gevolgen kunnen veroorzaken of vergroten vanwege ongelijkheid wat betreft macht of beschikking over informatie, bijvoorbeeld tussen werkgevers en werknemers, tussen bedrijven en consumenten of tussen overheden en burgers. Preventie van schade betekent ook dat er rekening moet worden gehouden met de natuurlijke omgeving en alle levende wezens.

- Het beginsel van rechtvaardigheid

(52) De ontwikkeling, de installatie en het gebruik van KI-systemen moeten rechtvaardig zijn. Wij erkennen dat er veel verschillende interpretaties van rechtvaardigheid bestaan, maar zijn ervan overtuigd dat rechtvaardigheid zowel een inhoudelijke als een procedurele dimensie heeft. De inhoudelijke dimensie impliceert een toezegging om de gelijke en rechtvaardige verdeling van zowel voordelen als kosten te waarborgen en ervoor te zorgen dat personen en groepen vrij zijn van onrechtvaardige vertekening, discriminatie en stigmatisering. Indien onrechtvaardige vertekening kan worden voorkomen, zouden KI-systemen zelfs de maatschappelijke rechtvaardigheid kunnen vergroten. Gelijke kansen wat betreft toegang tot onderwijs, goederen, diensten en technologie moeten ook worden bevorderd. Daarnaast mag het gebruik van KI-systemen nooit tot gevolg hebben dat de (eind)gebruikers worden misleid of worden beperkt in hun keuzevrijheid. Verder impliceert rechtvaardigheid dat beroepsbeoefenaars op het gebied van KI het beginsel van evenredigheid tussen middelen en doelen moeten eerbiedigen en zorgvuldig moeten afwegen hoe ze tegengestelde belangen en doelstellingen in evenwicht kunnen brengen.[31] De procedurele dimensie van rechtvaardigheid omvat het vermogen om beslissingen die worden genomen door KI-systemen en door de mensen die deze systemen beheren, aan te vechten en er effectief beroep tegen in te stellen.[32] Om dat te kunnen doen moet de entiteit

28 Het concept van menselijk toezicht wordt hieronder bij punt 65 verder uitgewerkt.

29 Schade kan individueel of collectief zijn en kan ontastbare schade aan sociale, culturele en politieke omgevingen omvatten.

30 Hieronder valt ook de leefwijze van personen of sociale groepen, om bijvoorbeeld culturele schade te voorkomen.

31 Dit heeft te maken met het evenredigheidsbeginsel (weergegeven in de stelregel dat men niet "met een kanon op een mug moet schieten"). Maatregelen die worden genomen om een bepaald doel te bereiken (bijv. de gegevensextractiemaatregelen die worden uitgevoerd om de KI-optimalisatiefunctie te realiseren), moeten worden beperkt tot het strikt noodzakelijke. Het houdt ook in dat, wanneer meerdere maatregelen in aanmerking komen voor het bereiken van een doel, de voorkeur moet worden gegeven aan de maatregel die het minst ingaat tegen de grondrechten en ethische normen (KI-ontwikkelaars moeten bijvoorbeeld altijd de voorkeur geven aan gegevens uit de publieke sector boven persoonsgegevens). Er kan ook worden verwezen naar de evenredigheid tussen gebruiker en installateur, met de rechten van bedrijven (zoals intellectuele eigendom en vertrouwelijkheid) enerzijds en de rechten van de gebruiker anderzijds.

32 Onder meer door gebruik te maken van hun recht van vereniging door lid te worden van een vakvereniging in een werkomgeving, overeenkomstig artikel 12 van het Handvest van de grondrechten van de EU.

die verantwoordelijk is voor de beslissing, identificeerbaar zijn en moet het besluitvormingsproces verklaarbaar zijn.

- Het beginsel van verantwoording

(53) Verantwoording is cruciaal voor het scheppen en behouden van het vertrouwen van gebruikers in KI-systemen. Dat betekent dat processen transparant moeten zijn, dat de capaciteiten en het doel van KI-systemen openlijk kenbaar moeten worden gemaakt en dat beslissingen – voor zover mogelijk – verklaarbaar moeten zijn aan degenen die er direct of indirect de gevolgen van ondervinden. Zonder die informatie kan een beslissing niet naar behoren worden aangevochten. Het is niet altijd mogelijk om te verklaren waarom een model een bepaald resultaat of een bepaalde beslissing heeft opgeleverd (en welke combinatie van inputfactoren daaraan heeft bijgedragen). Deze gevallen worden "blackbox"-algoritmen genoemd en vereisen speciale aandacht. In die situaties kunnen andere verantwoordingsmaatregelen (zoals traceerbaarheid, controleerbaarheid en transparante communicatie over de capaciteiten van het systeem) nodig zijn, mits het systeem als geheel de grondrechten eerbiedigt. De mate waarin verantwoording nodig is, hangt sterk af van de context en de ernst van de gevolgen, mocht het resultaat onjuist of anderszins onnauwkeurig zijn.[33]

2.3 Spanningen tussen de beginselen

(54) Er kunnen spanningen ontstaan tussen bovengenoemde beginselen en er is geen vaste oplossing om deze te verhelpen. Overeenkomstig de fundamentele toewijding van de EU aan democratische betrokkenheid, een eerlijke procesgang en open politieke participatie moeten er methoden van verantwoordelijk overleg worden vastgesteld om met dergelijke spanningen om te gaan. In verschillende toepassingsdomeinen kunnen *het beginsel van preventie van schade* en *het beginsel van menselijke autonomie* bijvoorbeeld met elkaar in strijd zijn. Neem bijvoorbeeld het gebruik van KI-systemen voor "voorspellend beleid", dat kan helpen de criminaliteit terug te dringen, maar wel op manieren waarbij toezichtsactiviteiten komen kijken die inbreuk maken op de individuele vrijheid en privacy. Verder moeten de algehele voordelen van KI-systemen aanzienlijk groter zijn dan de te voorziene individuele risico's. Hoewel deze beginselen absoluut richtsnoeren bieden voor oplossingen, blijven ze abstracte ethische voorschriften. Van beroepsbeoefenaars op het gebied van KI kan dus niet worden verwacht dat zij de juiste oplossing vinden op basis van bovenstaande beginselen, maar zij moeten ethische dilemma's en afwegingen benaderen aan de hand van logische, empirische reflectie, niet van intuïtie of een willekeurig oordeel. Er kunnen echter situaties zijn waarin er geen ethisch acceptabele compromissen te vinden zijn. Bepaalde grondrechten en bijbehorende beginselen zijn absoluut en kunnen geen onderdeel uitmaken van een compromis (bijv. menselijke waardigheid).

Essentiële richtsnoeren uit hoofdstuk I:

- ✓ Ontwikkel, installeer en gebruik KI-systemen op dusdanige wijze dat de volgende ethische beginselen worden nageleefd: *respect voor menselijke autonomie, preventie van schade, rechtvaardigheid en verantwoording.* Erken de potentiële spanningen tussen deze beginselen en pak deze aan.
- ✓ Schenk bijzondere aandacht aan situaties waarbij kwetsbaardere groepen betrokken zijn, zoals kinderen, mensen met een beperking en andere groepen die historisch gezien kansarm zijn of het risico lopen te worden uitgesloten, en/of aan situaties die worden gekenmerkt door ongelijkheid wat betreft macht of beschikking over informatie, bijvoorbeeld tussen werkgevers en werknemers of tussen bedrijven en consumenten[34].
- ✓ Erken en wees u ervan bewust dat KI-systemen de potentie hebben om de mens en de samenleving vele grote voordelen op te leveren, maar dat sommige ook negatieve gevolgen met zich mee kunnen brengen,

[33] Als een KI-systeem onnauwkeurige koopaanbevelingen voortbrengt, levert dat bijvoorbeeld weinig ethische zorgen op, in tegenstelling tot KI-systemen die evalueren of iemand die voor een strafbaar feit is veroordeeld, voorwaardelijk moet worden vrijgelaten.

[34] Zie de artikelen 24 tot en met 27 van het Handvest, waarin wordt ingegaan op de rechten van kinderen en ouderen, de integratie van mensen met een beperking en de rechten van werknemers. Zie ook artikel 38 over de bescherming van consumenten.

die mogelijk lastig te voorspellen, vast te stellen of te meten zijn (bijv. gevolgen voor de democratie, de rechtsstaat en de verdelende rechtvaardigheid of voor de menselijke geest). Neem waar nodig passende maatregelen, evenredig aan de omvang van het risico, om deze risico's te beperken.

II. Hoofdstuk II: Betrouwbare KI verwezenlijken

(55) In dit hoofdstuk worden richtsnoeren geboden voor de toepassing en verwezenlijking van betrouwbare KI, in de vorm van een lijst van zeven vereisten waaraan moet worden voldaan, voortbouwend op de in hoofdstuk I beschreven beginselen. Daarnaast worden de momenteel beschikbare technische en niet-technische methoden voor de uitvoering van deze vereisten in de volledige levenscyclus van het KI-systeem beschreven.

1. Vereisten van betrouwbare KI

(56) De in hoofdstuk I beschreven beginselen moeten worden vertaald naar concrete vereisten voor de verwezenlijking van betrouwbare KI. Deze vereisten zijn van toepassing op verschillende belanghebbenden die deel uitmaken van de levenscyclus van KI-systemen: ontwikkelaars, installateurs en eindgebruikers, alsook de samenleving in bredere zin. Met ontwikkelaars doelen wij op degenen die onderzoek doen naar KI-systemen en deze ontwerpen en/of ontwikkelen. Met installateurs doelen we op publieke of particuliere organisaties die gebruikmaken van KI-systemen binnen hun bedrijfsprocessen en om producten en diensten aan anderen aan te bieden. Eindgebruikers zijn degenen die, direct of indirect, met het KI-systeem te maken hebben. Onder de samenleving in bredere zin vallen tot slot alle anderen die direct of indirect door KI-systemen worden beïnvloed.

(57) De verschillende categorieën belanghebbenden spelen verschillende rollen om te zorgen dat aan de vereisten wordt voldaan:

a. ontwikkelaars moeten de vereisten uitvoeren en toepassen op de ontwerp- en ontwikkelingsprocessen;
b. installateurs moeten zorgen dat de systemen die zij gebruiken en de producten en diensten die zij aanbieden, aan de vereisten voldoen;
c. eindgebruikers en de samenleving in bredere zin moeten in kennis worden gesteld van deze vereisten en de mogelijkheid hebben te verzoeken dat eraan wordt voldaan.

(58) De onderstaande lijst van vereisten is niet uitputtend[35] en bevat systemische, individuele en maatschappelijke aspecten:

1 Menselijke controle en menselijk toezicht
Omvat grondrechten, menselijke controle en menselijk toezicht

2 Technische robuustheid en veiligheid
Omvat weerbaarheid tegen aanvallen en beveiliging, een uitwijkplan en algemene veiligheid, nauwkeurigheid, betrouwbaarheid en reproduceerbaarheid

3 Privacy en datagovernance
Omvat respect voor privacy, de kwaliteit en integriteit van gegevens en toegang tot gegevens

4 Transparantie
Omvat traceerbaarheid, verklaarbaarheid en communicatie

5 Diversiteit, non-discriminatie en rechtvaardigheid
Omvat het voorkomen van onrechtvaardige vertekening, toegankelijkheid en universeel ontwerp en

[35] Zonder een hiërarchie te willen opleggen vermelden we de beginselen hieronder in dezelfde volgorde waarin de beginselen en rechten waarmee ze verband houden, in het Handvest staan.

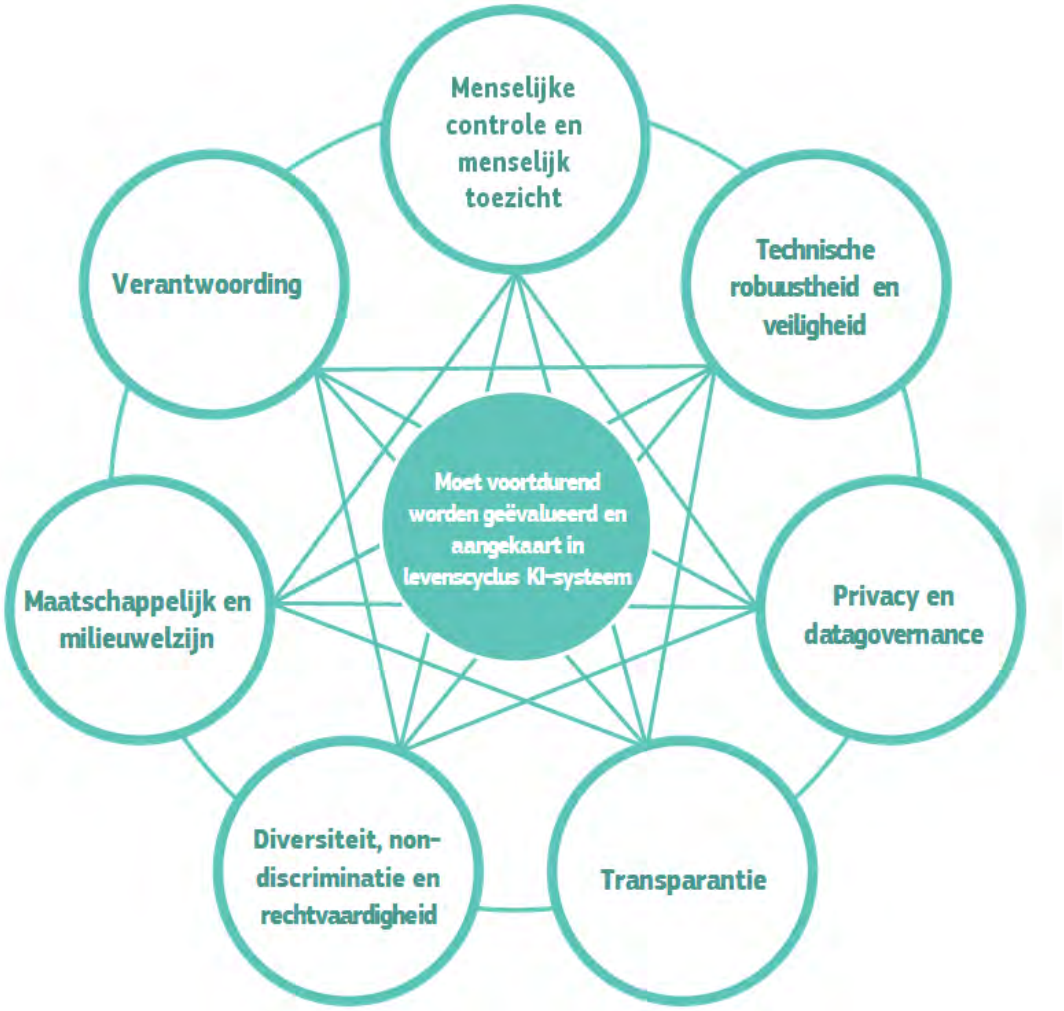
participatie van belanghebbenden

6 **Maatschappelijk en milieuwelzijn**

Omvat duurzaamheid en milieuvriendelijkheid, sociale gevolgen, de samenleving en de democratie

7 **Verantwoording**

Omvat controleerbaarheid, minimalisering en verslaglegging van negatieve gevolgen, afwegingen en beroep.



Figuur 2: Onderlinge verhouding van de zeven vereisten: alle vereisten zijn even belangrijk, ondersteunen elkaar en moeten gedurende de levenscyclus van een KI-systeem worden uitgevoerd en geëvalueerd

(59) Hoewel alle vereisten even belangrijk zijn, moet er rekening worden gehouden met de context en de mogelijke spanningen ertussen wanneer ze in verschillende domeinen en sectoren worden toegepast. De uitvoering van deze vereisten moet plaatsvinden gedurende de volledige levenscyclus van het KI-systeem en hangt af van de specifieke toepassing. De meeste vereisten zijn op alle KI-systemen van toepassing, maar er wordt bijzondere aandacht besteed aan de vereisten die direct of indirect van invloed zijn op personen. Daarom kunnen ze voor sommige toepassingen (bijvoorbeeld in een industriële context) minder belangrijk zijn.

(60) De bovenstaande vereisten bevatten elementen die in sommige gevallen al in de bestaande wetgeving worden weergegeven. Wij herhalen dat het – overeenkomstig de eerste component van betrouwbare KI – de verantwoordelijkheid van de ontwikkelaars en installateurs van KI-systemen is om te zorgen dat ze voldoen aan hun wettelijke verplichtingen wat betreft zowel horizontaal toepasselijke regels als domeinspecifieke

regelgeving.

(61) Bij de volgende punten worden alle vereisten verder uitgewerkt.

1. <u>Menselijke controle en menselijk toezicht</u>

(62) KI-systemen moeten menselijke autonomie en beslissingen ondersteunen, zoals voorgeschreven door het beginsel van *respect voor menselijke autonomie*. Daarvoor is het nodig dat KI-systemen een democratische, florerende en gelijkwaardige samenleving mogelijk maken door de controle van de gebruiker te ondersteunen en grondrechten te bevorderen, en dat ze de mogelijkheid bieden voor menselijk toezicht.

(63) ***Grondrechten***. Zoals veel technologieën kunnen KI-systemen grondrechten evenzeer bevorderen als belemmeren. Mensen kunnen er bijvoorbeeld voordeel van hebben doordat de systemen hen helpen hun persoonsgegevens bij te houden of doordat onderwijs toegankelijker wordt en hun recht op onderwijs dus wordt ondersteund. Gezien de reikwijdte en capaciteit van KI-systemen kunnen deze echter ook negatieve gevolgen hebben voor grondrechten. In situaties waar dergelijke risico's bestaan, moet een effectbeoordeling worden uitgevoerd wat grondrechten betreft. Deze moet voorafgaand aan de ontwikkeling van het systeem worden uitgevoerd en moet een evaluatie omvatten van de vraag of de risico's kunnen worden beperkt of gerechtvaardigd zoals noodzakelijk in een democratische samenleving om de rechten en vrijheden van anderen te respecteren. Daarnaast moeten er mechanismen worden ingesteld om externe feedback te ontvangen ten aanzien van KI-systemen die mogelijk inbreuk maken op grondrechten.

(64) ***Menselijke controle.*** Gebruikers moeten onderbouwde, autonome beslissingen kunnen nemen ten aanzien van KI-systemen. Ze moeten de kennis en hulpmiddelen ontvangen om KI-systemen te kunnen begrijpen en er in bevredigende mate mee te kunnen omgaan, en ze moeten, indien nodig, in staat worden gesteld het systeem redelijkerwijs zelf te controleren of aan te vechten. KI-systemen moeten mensen ondersteunen bij het maken van betere, meer onderbouwde keuzen die overeenkomen met hun doelen. Soms kunnen KI-systemen worden geïnstalleerd om menselijk gedrag te vormen en beïnvloeden via mechanismen die lastig te detecteren zijn, omdat ze gebruikmaken van onbewuste processen, met inbegrip van verschillende vormen van onrechtvaardige manipulatie, misleiding, drijven en conditionering, allemaal bedreigingen voor de individuele autonomie. Het algehele beginsel van gebruikersautonomie moet centraal staan bij de functionaliteit van het systeem. Cruciaal daarbij is het recht niet te worden onderworpen aan een uitsluitend op geautomatiseerde verwerking gebaseerd besluit waaraan voor gebruikers rechtsgevolgen zijn verbonden of dat hen anderszins in aanmerkelijke mate treft[36].

(65) ***Menselijk toezicht***. Menselijk toezicht helpt om te zorgen dat een KI-systeem de menselijke autonomie niet ondermijnt en geen andere negatieve effecten veroorzaakt. Toezicht kan worden verwezenlijkt via governancemechanismen, zoals een benadering met "human-in-the-loop" (HITL), "human-on-the-loop" (HOTL) of "human-in-command" (HIC). HITL verwijst naar de capaciteit voor menselijke interventie in elke besluitcyclus van het systeem, die in veel gevallen mogelijk noch wenselijk is. HOTL verwijst naar de capaciteit voor menselijke interventie gedurende de ontwerpcyclus van het systeem en het monitoren van de werking van het systeem. HIC verwijst naar de capaciteit om de algehele activiteit van het KI-systeem te overzien (inclusief zijn bredere economische, maatschappelijke, juridische en ethische gevolgen) en het vermogen om te kiezen wanneer en hoe het systeem in iedere specifieke situatie wordt gebruikt. Daarbij kan het bijvoorbeeld gaan om de keuze om een KI-systeem in een bepaalde situatie niet te gebruiken, om een bepaalde mate van menselijk oordeel vast te stellen tijdens het gebruik van het systeem of om te garanderen dat een door het systeem genomen beslissing kan worden herroepen. Verder moet ervoor worden gezorgd dat openbare handhavingsinstanties in staat zijn toezicht te houden overeenkomstig hun mandaat. Toezichtsmechanismen kunnen in verschillende mate nodig zijn om andere veiligheids- en

[36] Hier kan worden verwezen naar artikel 22 van de AVG, waar dit recht al is vastgelegd.

controlemaatregelen te ondersteunen, afhankelijk van het toepassingsgebied van het KI-systeem en het potentiële risico. Indien alle andere omstandigheden gelijk blijven, moet een KI-systeem uitgebreider worden getest en is er strengere governance nodig naarmate er minder menselijk toezicht mogelijk is op het systeem.

2. <u>Technische robuustheid en veiligheid</u>

(66) Een cruciale component van de verwezenlijking van betrouwbare KI is technische robuustheid, die nauw verbonden is met het *beginsel van preventie van schade*. Voor technische robuustheid moeten KI-systemen worden ontwikkeld met een preventieve benadering van risico's en op dusdanige wijze dat ze zich betrouwbaar en zoals bedoeld gedragen, terwijl onbedoelde en onverwachte schade zo veel mogelijk wordt beperkt en onacceptabele schade wordt voorkomen. Dit moet ook gelden voor potentiële veranderingen in de omgeving waarin ze werken of de aanwezigheid van andere actoren (menselijk en kunstmatig) die mogelijk op conflicterende wijze met het systeem in aanraking komen. Daarnaast moet de fysieke en geestelijke integriteit van mensen worden gewaarborgd.

(67) ***Weerbaarheid tegen aanvallen en beveiliging***. Zoals alle softwaresystemen moeten KI-systemen worden beschermd tegen kwetsbaarheden waardoor ze door kwaadwillenden kunnen worden misbruikt (bijv. gehackt). Aanvallen kunnen gericht zijn op de gegevens (datavergiftiging), het model (lek in het model) of de onderliggende infrastructuur, zowel software als hardware. Als een KI-systeem wordt aangevallen, bijvoorbeeld bij vijandige aanvallen, kunnen zowel de gegevens als het gedrag van het systeem worden veranderd, waardoor het systeem andere beslissingen neemt of volledig wordt afgesloten. Systemen en gegevens kunnen ook worden aangetast door kwade opzet of door blootstelling aan onverwachte situaties. Ontoereikende beveiligingsprocessen kunnen ook leiden tot onjuiste beslissingen of zelfs fysieke schade. KI-systemen worden pas veilig geacht[37] wanneer er rekening is gehouden met mogelijke onbedoelde toepassingen van KI (bijv. toepassingen voor dubbel gebruik) en potentieel misbruik van een KI-systeem door kwaadwillende actoren en er stappen zijn ondernomen om deze dingen te voorkomen en te beperken.[38]

(68) ***Uitwijkplan en algemene veiligheid***. KI-systemen moeten over waarborgen beschikken die in geval van problemen een uitwijkplan in werking stellen. Dat kan betekenen dat KI-systemen wisselen van een statistische naar een op regels gebaseerde procedure of dat ze om een menselijke beheerder vragen voor ze hun werk voortzetten.[39] Er moet worden gezorgd dat het systeem doet wat het moet doen zonder schade te berokkenen aan levende wezens of het milieu. Daaronder valt ook de minimalisering van onbedoelde gevolgen en fouten. Daarnaast moeten er processen worden ingesteld om potentiële risico's in verband met het gebruik van KI-systemen in verschillende toepassingsdomeinen te verduidelijken en beoordelen. Het benodigde niveau van veiligheidsmaatregelen hangt af van de omvang van het risico dat een KI-systeem met zich meebrengt, die op zijn beurt weer afhangt van de capaciteiten van het systeem. Indien er kan worden voorzien dat het ontwikkelingsproces of het systeem zelf bijzonder hoge risico's zal opleveren, is het cruciaal dat er proactief veiligheidsmaatregelen worden ontwikkeld en getest.

(69) ***Nauwkeurigheid.*** Nauwkeurigheid heeft betrekking op het vermogen van een KI-systeem om correcte afwegingen te maken, bijvoorbeeld door informatie in de juiste categorieën in te delen, of op het vermogen ervan om correcte voorspellingen of aanbevelingen te doen of beslissingen te nemen op basis van gegevens of modellen. Door een expliciet en gedegen ontwikkelings- en evaluatieproces kunnen onbedoelde risico's vanwege onjuiste voorspellingen worden beperkt en gecorrigeerd. Indien onnauwkeurige voorspellingen zo nu

---

37 Zie bijv. de overwegingen onder punt 2.7 van het Gecoördineerd plan inzake kunstmatige intelligentie van de Europese Unie.

38 Voor de beveiliging van KI-systemen kan het hard nodig zijn om binnen onderzoek en ontwikkeling een heilzame cyclus te kunnen ontwikkelen tussen het doorgronden van aanvallen, het ontwikkelen van passende bescherming en de verbetering van evaluatiemethoden. Voor de verwezenlijking daarvan moet convergentie tussen de KI-gemeenschap en de beveiligingsgemeenschap worden gestimuleerd. Daarnaast is het de verantwoordelijkheid van alle betrokken actoren om gemeenschappelijke, grensoverschrijdende veiligheids- en beveiligingsnormen te bepalen en een omgeving van wederzijds vertrouwen te creëren waarin internationale samenwerking wordt bevorderd. Zie voor mogelijke maatregelen Malicious Use of AI" (Avin S., Brundage M. et. al., 2018).

39 Scenario's waarbij menselijke interventie niet direct mogelijk zou zijn, moeten ook in overweging worden genomen.

en dan niet kunnen worden voorkomen, is het belangrijk dat het systeem kan aangeven hoe groot de kans op dergelijke fouten is. Een hoge mate van nauwkeurigheid is in het bijzonder belangrijk in situaties waarin het KI-systeem rechtstreeks gevolgen heeft voor de levens van mensen.

(70) ***Betrouwbaarheid en reproduceerbaarheid.*** Het is essentieel dat de resultaten van KI-systemen reproduceerbaar en betrouwbaar zijn. Een betrouwbaar KI-systeem is een systeem dat goed werkt met allerlei soorten input en in allerlei situaties. Dit is nodig om een KI-systeem te controleren en om onbedoelde schade te voorkomen. Reproduceerbaarheid beschrijft of een KI-experiment hetzelfde gedrag vertoont wanneer het onder gelijke omstandigheden wordt herhaald. Hierdoor kunnen wetenschappers en beleidsmakers nauwkeurig beschrijven wat KI-systemen doen. Replicatiebestanden[40] kunnen het proces van het testen en reproduceren van gedrag vergemakkelijken.

3. Privacy en datagovernance

(71) Nauw verbonden met het *beginsel van preventie van schade* is privacy, een grondrecht waarop KI-systemen in het bijzonder van invloed zijn. Voor de preventie van privacyschade is ook geschikte datagovernance noodzakelijk die betrekking heeft op de kwaliteit en integriteit van de gebruikte gegevens, de relevantie ervan binnen het domein waarop de KI-systemen worden geïnstalleerd, de toegangsprotocollen en de capaciteit om gegevens op dusdanige wijze te verwerken dat de privacy wordt beschermd.

(72) ***Privacy en gegevensbescherming***. KI-systemen moeten privacy en gegevensbescherming garanderen gedurende de volledige levenscyclus van het systeem[41]. Dit omvat de informatie die oorspronkelijk door de gebruiker is aangeleverd, alsook de informatie die in de loop van zijn/haar interactie met het systeem over de gebruiker is gegenereerd (bijv. resultaten die het KI-systeem heeft gegenereerd voor specifieke gebruikers of de manier waarop gebruikers op bepaalde aanbevelingen hebben gereageerd). Uit de digitale documentatie van menselijk gedrag kunnen KI-systemen niet alleen de voorkeuren van mensen afleiden, maar ook hun seksuele geaardheid, leeftijd, geslacht en religieuze of politieke standpunten. Om te zorgen dat mensen het proces van gegevensverzameling kunnen vertrouwen, moet ervoor worden gezorgd dat de over hen verzamelde gegevens niet worden gebruikt om hen onwettig of onrechtvaardig te discrimineren.

(73) ***Kwaliteit en integriteit van gegevens***. De kwaliteit van de gebruikte gegevenssets is cruciaal voor de prestaties van KI-systemen. Wanneer gegevens worden verzameld, kunnen ze sociaal geconstrueerde vertekening, onnauwkeurigheden, fouten en vergissingen bevatten. Deze moeten worden verholpen vóórdat het systeem met een bepaalde gegevensset wordt getraind. Daarnaast moet de integriteit van de gegevens worden gewaarborgd. Als er kwaadwillige gegevens in een KI-systeem worden ingevoerd, kan het gedrag ervan veranderen, met name bij zelflerende systemen. De gebruikte processen en gegevenssets moeten bij elke stap, zoals het plannen, trainen, testen en installeren, worden getest en gedocumenteerd. Dit moet ook gelden voor KI-systemen die niet intern zijn ontwikkeld, maar elders zijn verkregen.

(74) ***Toegang tot gegevens****.* In iedere organisatie die gegevens van personen verwerkt (ongeacht of iemand een gebruiker van het systeem is of niet), moeten gegevensprotocollen worden ingesteld om de toegang tot gegevens te beheren. In deze protocollen moet worden beschreven wie onder welke omstandigheden gegevens kan inzien. Uitsluitend gekwalificeerd personeel met de bevoegdheid en noodzaak om de gegevens van personen in te zien, mag hiertoe de mogelijkheid krijgen.

4. Transparantie

(75) Deze vereiste houdt nauw verband met het *beginsel van verantwoording* en omvat de transparantie van

[40] Hierbij gaat het om bestanden die elke stap van het ontwikkelingsproces van het KI-systeem herhalen, van het onderzoek en de initiële gegevensverzameling tot de resultaten.

[41] Hier kan worden verwezen naar bestaande privacywetgeving, zoals de AVG of de komende e-privacyverordening.

elementen die relevant zijn voor een KI-systeem: de gegevens, het systeem en de bedrijfsmodellen.

(76) ***Traceerbaarheid***. De gegevenssets en de processen waaruit de beslissing van het KI-systeem voortkomt, met inbegrip van die van de verzameling en indeling van gegevens, alsook de gebruikte algoritmen, moeten zo goed mogelijk worden gedocumenteerd om ze traceerbaar te maken en de transparantie te vergroten. Dit geldt ook voor de door het KI-systeem gemaakte beslissingen. Hierdoor wordt het mogelijk vast te stellen waarom een KI-beslissing onjuist was, waardoor vervolgens toekomstige fouten kunnen worden voorkomen. Traceerbaarheid maakt dus controleerbaarheid en verklaarbaarheid mogelijk.

(77) ***Verklaarbaarheid.*** Verklaarbaarheid heeft te maken met het vermogen om zowel de technische processen van een KI-systeem als de daaraan gerelateerde menselijke beslissingen (bijv. de toepassingsgebieden van een KI-systeem) te verklaren. Voor technische verklaarbaarheid is het nodig dat de door een KI-systeem gemaakte beslissingen door mensen kunnen worden begrepen en getraceerd. Daarnaast moeten er mogelijk afwegingen worden gemaakt tussen het verbeteren van de verklaarbaarheid van een systeem (waardoor de nauwkeurigheid zou kunnen afnemen) of het vergroten van de nauwkeurigheid ervan (ten koste van de verklaarbaarheid). Wanneer een KI-systeem significante gevolgen heeft voor het leven van mensen, moet het altijd mogelijk zijn om te vragen om een geschikte verklaring van het besluitvormingsproces van het KI-systeem. Een dergelijke verklaring moet tijdig worden gegeven en zijn afgestemd op de mate van deskundigheid van de betrokken belanghebbende (bijv. leek, regelgever of onderzoeker). Verder moeten er verklaringen beschikbaar zijn van de mate waarin een KI-systeem het organisatorische besluitvormingsproces, de ontwerpkeuzen van het systeem en de motivering voor de installatie ervan beïnvloedt en vormt (zodat de transparantie van het bedrijfsmodel wordt gewaarborgd).

(78) ***Communicatie.*** KI-systemen mogen zich tegenover gebruikers niet als mensen voordoen. Mensen hebben het recht ervan in kennis te worden gesteld dat ze met een KI-systeem te maken hebben. Dat houdt in dat KI-systemen als zodanig herkenbaar moeten zijn. Daarnaast moet, waar dit nodig is om naleving van de grondrechten te waarborgen, de optie worden geboden om te besluiten menselijke interactie aan te gaan in plaats van deze interactie. Bovendien moeten de capaciteiten en beperkingen van het KI-systeem op een voor de betreffende situatie gepaste manier aan beroepsbeoefenaars op het gebied van KI of aan eindgebruikers worden gemeld. Daaronder kan communicatie vallen over de mate van nauwkeurigheid van het systeem, alsook over de beperkingen ervan.

5. Diversiteit, non-discriminatie en rechtvaardigheid

(79) Om betrouwbare KI te verwezenlijken moeten we inclusie en diversiteit mogelijk maken gedurende de gehele levenscyclus van het KI-systeem. Daarbij moet niet alleen gedurende het hele proces met alle betrokken belanghebbenden rekening worden gehouden, maar moet ook worden gezorgd voor gelijke toegang via inclusieve ontwerpprocessen, alsook voor gelijke behandeling. Deze vereiste is nauw verbonden met *het beginsel van rechtvaardigheid*.

(80) ***Voorkomen van onrechtvaardige vertekening***. In door KI-systemen gebruikte gegevenssets (zowel voor training als voor werkzaamheden) kan sprake zijn van onbedoelde historische vertekening, onvolledigheid of slechte governancemodellen. Als dergelijke vertekening wordt behouden, kunnen er onbedoelde (in)directe vooroordelen en discriminatie[42] tegen bepaalde groepen of mensen ontstaan, waardoor vooroordelen en marginalisering mogelijk worden versterkt. Er kan ook schade ontstaan als gevolg van de onbedoelde uitbuiting van vertekening (door consumenten) of door oneerlijke mededinging, zoals de homogenisering van

[42] Zie voor een definitie van directe en indirecte discriminatie bijvoorbeeld artikel 2 van Richtlijn 2000/78/EG van de Raad van 27 november 2000 tot instelling van een algemeen kader voor gelijke behandeling in arbeid en beroep. Zie ook artikel 21 van het Handvest van de grondrechten van de EU.

prijzen door middel van samenspanning of een niet-transparante markt.[43] Aanwijsbare en discriminerende vertekening moet in de verzamelingsfase waar mogelijk worden verwijderd. Er kan ook sprake zijn van vertekening in de manier waarop KI-systemen worden ontwikkeld (bijv. het programmeren van algoritmen). Deze vertekening kan worden tegengegaan door toezichtsprocessen in te stellen om het doel, de beperkingen, de vereisten en de beslissingen van het systeem op een heldere en transparante manier te analyseren en te behandelen. Daarnaast kan de diversiteit van meningen worden gewaarborgd door personeel met diverse achtergronden en uit verschillende culturen en disciplines in dienst te nemen, dus dit moet worden aangemoedigd.

(81) ***Toegankelijkheid en universeel ontwerp***. Met name in domeinen tussen bedrijf en consument moeten bij systemen de gebruikers centraal staan en moeten ze zo worden ontworpen dat mensen KI-producten of -diensten kunnen gebruiken ongeacht hun leeftijd, geslacht, vermogens of eigenschappen. Het is in het bijzonder van belang dat deze technologie toegankelijk is voor mensen met een beperking, die in alle groepen van de samenleving voorkomen. In KI-systemen moet geen gebruik worden gemaakt van één vaste aanpak voor alle situaties, maar er moet rekening worden gehouden met beginselen van universeel ontwerp[44] die zijn gericht op het breedst mogelijke scala aan gebruikers, volgens relevante toegankelijkheidsnormen.[45] Daardoor worden voor alle mensen gelijke toegang tot en actieve participatie aan bestaande en nieuwe door de computer ondersteunde menselijke activiteiten en ten aanzien van ondersteunende technologieën mogelijk.[46]

(82) ***Participatie van belanghebbenden***. Om betrouwbare KI-systemen te ontwikkelen is het raadzaam om belanghebbenden te raadplegen die direct of indirect met het systeem te maken hebben gedurende de levenscyclus ervan. Het is nuttig om, ook na de installatie, regelmatig om feedback te vragen en langetermijnmechanismen voor de participatie van belanghebbenden in te stellen, bijvoorbeeld door de informatie voor en de raadpleging en participatie van werknemers te garanderen gedurende het volledige toepassingsproces van KI-systemen bij organisaties.

6. Maatschappelijk en milieuwelzijn

(83) Overeenkomstig de *beginselen van rechtvaardigheid* en *preventie van schade* moeten de samenleving in bredere zin, andere wezens met gevoel en het milieu gedurende de KI-levenscyclus ook als belanghebbenden worden beschouwd. Duurzaamheid en ecologische verantwoordelijkheid van KI-systemen moeten worden aangemoedigd en onderzoek naar KI-oplossingen die ingaan op mondiale punten van zorg, zoals de duurzameontwikkelingsdoelstellingen, moet worden gestimuleerd. In het ideale scenario wordt KI gebruikt ten voordele van alle mensen, inclusief toekomstige generaties.

(84) ***Duurzame en milieuvriendelijke KI***. KI-systemen hebben de potentie om een aantal van de meest urgente maatschappelijke zorgen te verhelpen, maar er moet wel worden gezorgd dat dit zo milieuvriendelijk mogelijk gebeurt. Het ontwikkelings-, installatie- en gebruiksproces van het systeem, alsook de volledige toeleveringsketen, moet met betrekking daartoe worden gecontroleerd, bijvoorbeeld door via een kritisch onderzoek naar het gebruik van hulpbronnen en de energieconsumptie tijdens de training te kiezen voor minder schadelijke opties. Maatregelen om de milieuvriendelijkheid van de volledige toeleveringsketen van het KI-systeem te waarborgen, moeten worden aangemoedigd.

---

[43] Cf. paper van het Bureau van de Europese Unie voor de grondrechten:
"BigData: Discrimination in data-supported decision making (2018)" http://fra.europa.eu/en/publication/2018/big-data-discrimination.

[44] Op grond van artikel 42 van de richtlijn inzake het plaatsen van overheidsopdrachten moet bij technische specificaties rekening worden gehouden met toegankelijkheid en ontwerp voor iedereen.

[45] Bijvoorbeeld EN 301 549.

[46] Deze vereiste houdt verband met het Verdrag van de Verenigde Naties inzake de rechten van personen met een handicap.

(85) ***Sociale gevolgen***. Als we op alle terreinen van ons leven aan sociale KI-systemen[47] worden blootgesteld (of het nou gaat om onderwijs, werk, zorg of vermaak), kan onze opvatting van sociale controle veranderen of kan dit gevolgen hebben voor onze sociale relaties en hechting. Hoewel KI-systemen kunnen worden gebruikt om sociale vaardigheden te vergroten,[48] kunnen ze evenzeer bijdragen aan de verslechtering ervan. Dat zou ook gevolgen kunnen hebben voor het fysieke en geestelijke welzijn van mensen. De effecten van deze systemen moeten daarom zorgvuldig worden gemonitord en afgewogen.

(86) ***Samenleving en democratie.*** Naast de beoordeling van de gevolgen van de ontwikkeling, de installatie en het gebruik van een KI-systeem voor individuele personen, moeten deze gevolgen ook vanuit maatschappelijk oogpunt worden beoordeeld, waarbij rekening moet worden gehouden met het effect op instellingen, de democratie en de samenleving als geheel. Het gebruik van KI-systemen moet zorgvuldig worden afgewogen, met name in situaties die verband houden met het democratische proces, met inbegrip van niet alleen politieke besluitvorming, maar ook verkiezingen.

7. Verantwoording

(87) De vereiste van verantwoording vormt een aanvulling op de bovenstaande vereisten en is nauw verbonden aan het *beginsel van rechtvaardigheid*. Op grond van deze vereiste moeten mechanismen worden ingesteld om de verantwoordelijkheid en verantwoording voor KI-systemen en de resultaten daarvan te garanderen, zowel voor als na de toepassing.

(88) ***Controleerbaarheid.*** Controleerbaarheid houdt in dat het mogelijk wordt gemaakt de algoritmen, gegevens en ontwerpprocessen te controleren. Dat betekent niet noodzakelijkerwijs dat informatie over bedrijfsmodellen en intellectuele eigendom in verband met het KI-systeem altijd openbaar beschikbaar moet zijn. Evaluatie door interne en externe controleurs en de beschikbaarheid van dergelijke evaluatieverslagen kunnen bijdragen aan de betrouwbaarheid van de technologie. Bij toepassingen die van invloed zijn op grondrechten, met inbegrip van veiligheidskritieke toepassingen, moeten KI-systemen onafhankelijk kunnen worden gecontroleerd.

(89) ***Minimalisering en verslaglegging van negatieve gevolgen***. Het vermogen om verslag te doen van handelingen of beslissingen die bijdragen aan een bepaald resultaat van het systeem, alsook het vermogen om op de gevolgen van een dergelijk resultaat te reageren, moet worden gewaarborgd. De vaststelling, beoordeling, verslaglegging en minimalisering van de potentiële negatieve effecten van KI-systemen is in het bijzonder cruciaal voor degenen die er (in)direct de gevolgen van ondervinden. Er moet gedegen bescherming beschikbaar zijn voor klokkenluiders, ngo's, vakverenigingen of andere entiteiten wanneer zij melding maken van legitieme zorgen over een op KI gebaseerd systeem. Het gebruik van effectbeoordelingen (bijv. red teaming of vormen van algoritmische effectbeoordeling), zowel voor als tijdens de ontwikkeling, de installatie en het gebruik van KI-systemen, kan nuttig zijn voor het minimaliseren van negatieve gevolgen. Deze beoordelingen moeten in verhouding staan tot het risico dat de KI-systemen met zich meebrengen.

(90) ***Afwegingen.*** Bij de uitvoering van de bovenstaande vereisten kunnen er spanningen tussen ontstaan, die tot onvermijdelijke afwegingen kunnen leiden. Dergelijke afwegingen moeten op een rationele en methodologische manier worden benaderd binnen de stand van de technologie. Dat houdt in dat er moet worden vastgesteld welke relevante belangen en waarden door het KI-systeem in het gedrang komen en dat,

---

[47] Hierbij gaat het om KI-systemen die met mensen communiceren en contact hebben door sociaal gedrag te simuleren in interactie met een menselijke robot (belichaamde KI) of als avatars in een virtuele werkelijkheid. Daardoor hebben die systemen de potentie om onze sociaal-culturele praktijken en de structuur van ons sociale leven te veranderen.

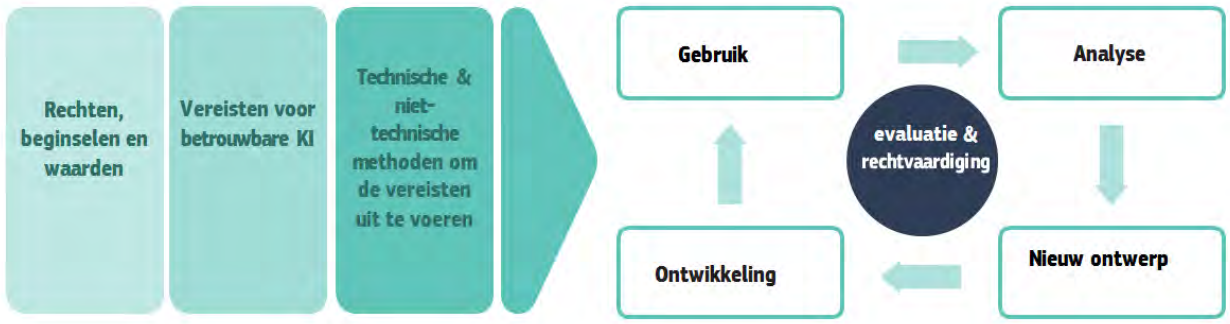
[48] Zie bijvoorbeeld het door de EU gefinancierde project waarbij software op basis van KI wordt ontwikkeld om robots in staat te stellen tot effectievere interactie met autistische kinderen tijdens door mensen geleide sessies en de kinderen zo te helpen hun sociale en communicatieve vaardigheden te verbeteren:
http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968

in geval van een conflict, de afwegingen ertussen uitdrukkelijk moeten worden erkend en geëvalueerd wat betreft het risico voor ethische beginselen, met inbegrip van grondrechten. In situaties waarin geen ethisch acceptabele compromissen kunnen worden gevonden, mogen de ontwikkeling, de installatie en het gebruik van het KI-systeem niet in die vorm worden voortgezet. Alle beslissingen over te maken afwegingen moeten worden onderbouwd en goed worden gedocumenteerd. De beslisser moet aansprakelijk zijn voor de manier waarop de passende afweging wordt gemaakt, en moet de gepastheid van de genomen beslissing voortdurend controleren, zodat waar nodig de noodzakelijke aanpassingen aan het systeem kunnen worden gemaakt. [49]

(91) ***Beroep***. Wanneer zich een negatief effect voordoet, moeten er toegankelijke mechanismen bestaan waarmee geschikte mogelijkheden om beroep in te stellen worden gewaarborgd[50]. De wetenschap dat het mogelijk is om beroep in te stellen wanneer er iets verkeerd gaat, is essentieel voor het vertrouwen. Er moet bijzondere aandacht worden besteed aan kwetsbare personen of groepen.

2. Technische en niet-technische methoden voor de verwezenlijking van betrouwbare KI

(92) Voor de uitvoering van de bovenstaande vereisten kunnen zowel technische als niet-technische methoden worden gebruikt. Hieronder vallen alle fasen van de levenscyclus van een KI-systeem. De methoden die worden gebruikt om de vereisten uit te voeren, moeten voortdurend worden geëvalueerd. Ook moet er voortdurend verslag worden gedaan van de aanpassingen aan de uitvoeringsprocessen en moeten deze worden gerechtvaardigd.[51] Omdat KI-systemen voortdurend veranderen en binnen een dynamische omgeving werkzaam zijn, is de verwezenlijking van betrouwbare KI een doorlopend proces, dat hieronder in figuur 3 wordt weergegeven.



Figuur 3: Verwezenlijking van betrouwbare KI gedurende de gehele levenscyclus van het systeem

(93) De volgende methoden kunnen worden beschouwd als aanvulling op of alternatief voor elkaar, aangezien voor verschillende vereisten – en verschillende gevoeligheden – mogelijk verschillende uitvoeringsmethoden nodig zijn. Dit overzicht beoogt niet volledig, uitputtend of verplicht te zijn. Het doel ervan is veeleer om een lijst te geven van voorgestelde methoden die kunnen helpen bij de verwezenlijking van betrouwbare KI.

1. Technische methoden

(94) In dit gedeelte worden de technische methoden beschreven voor de waarborging van betrouwbare KI die in de

---

[49] Dit kan met behulp van verschillende governancemodellen worden verwezenlijkt. De aanwezigheid van een interne en/of externe ethische (en sectorspecifieke) deskundige of raad kan bijvoorbeeld nuttig zijn om potentiële conflictgebieden uit te lichten en manieren voor te stellen waarop dat conflict het best kan worden opgelost. Zinvolle raadpleging van en gesprekken met belanghebbenden, met inbegrip van degenen die het risico lopen negatieve gevolgen te ondervinden van een KI-systeem, is ook nuttig. Europese universiteiten moeten een voortrekkersrol spelen bij de opleiding van de benodigde ethiekdeskundigen.

[50] Zie ook het advies van het Bureau van de Europese Unie voor de grondrechten: "Improving access to remedy in the area of business and human rights at the EU level" (2017), https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] Hieronder valt bijvoorbeeld rechtvaardiging van de bij het ontwerp, de ontwikkeling en de installatie van het systeem gemaakte keuzen om de bovenstaande vereisten erin te verwerken.

ontwerp-, ontwikkelings- en gebruiksfase van een KI-systeem kunnen worden verwerkt. De onderstaande methoden variëren in mate van volwassenheid[52].

- *Architecturen voor betrouwbare KI*

(95) Vereisten voor betrouwbare KI moeten worden vertaald naar procedures en/of beperkingen van procedures, die in de architectuur van het KI-systeem moeten worden verankerd. Dit kan worden bereikt door middel van een "witte lijst" met een reeks regels (gedragingen of toestanden) die het systeem altijd moet volgen, een "zwarte lijst" met beperkingen van gedragingen of toestanden die het systeem nooit mag overtreden, en combinaties daarvan of complexere, aantoonbare garanties ten aanzien van het gedrag van het systeem. De monitoring van de naleving van deze beperkingen door het systeem gedurende de werkzaamheden kan via een afzonderlijk proces worden verwezenlijkt.

(96) KI-systemen met leercapaciteiten die hun gedrag dynamisch kunnen aanpassen, kunnen worden gezien als non-deterministische systemen die mogelijk onverwacht gedrag vertonen. Naar deze systemen wordt vaak gekeken vanuit het theoretische oogpunt van een cyclus van "waarnemen-plannen-handelen". Om deze architectuur aan te passen voor de verwezenlijking van betrouwbare KI moeten de vereisten worden geïntegreerd bij alle drie de stappen van de cyclus: i) bij de stap "waarnemen" moet het systeem zo worden ontwikkeld dat het alle omgevingselementen herkent die nodig zijn om naleving van de vereisten te waarborgen; ii) bij de stap "plannen" moet het systeem uitsluitend plannen overwegen die aansluiten bij de vereisten; iii) bij de stap "handelen" moeten de handelingen van het systeem beperkt blijven tot gedrag waarmee de vereisten worden verwezenlijkt.

(97) De architectuur zoals hierboven geschetst, is algemeen en biedt voor de meeste KI-systemen slechts een imperfecte beschrijving. Toch biedt zij aanknopingspunten voor beperkingen en beleidspunten die in specifieke modules moeten worden weergegeven, zodat er een algeheel systeem ontstaat dat betrouwbaar is en ook zo wordt gezien.

- *Ethiek en rechtsstaat door ontwerp (X door ontwerp)*

(98) Methoden om waarden door ontwerp te waarborgen verbinden de abstracte beginselen waaraan het systeem moet voldoen en de specifieke uitvoeringsbeslissingen nauwkeurig en expliciet. Het idee dat naleving van normen kan worden opgenomen in het ontwerp van het KI-systeem is voor deze methode essentieel. Het is de verantwoordelijkheid van bedrijven om de gevolgen van hun KI-systemen en de normen waaraan hun KI-systeem moet voldoen om negatieve gevolgen te voorkomen, vanaf het allereerste begin vast te stellen. Verschillende "door-ontwerp"-concepten worden al veelvuldig gebruikt, zoals *privacy door ontwerp* en *beveiliging door ontwerp.* Zoals hierboven aangegeven, moeten de processen, gegevens en resultaten van KI veilig zijn, wil het systeem vertrouwen winnen. Ook moet het systeem robuust worden ontworpen om kwaadwillige gegevens en aanvallen te weerstaan. Er moet een mechanisme voor veiligheidsuitschakeling worden opgenomen, en het systeem moet zijn werkzaamheden kunnen vervolgen na gedwongen uitschakeling (bijv. een aanval).

- *Verklaringsmethoden*

(99) Een systeem kan pas betrouwbaar zijn als we kunnen begrijpen waarom het zich op een bepaalde manier heeft gedragen en waarom het een bepaalde interpretatie heeft gegeven. Er is een heel onderzoeksgebied, verklaarbare KI (VKI), gewijd aan deze kwestie om de onderliggende mechanismen van het systeem beter te begrijpen en oplossingen te vinden. Momenteel is dit nog een open uitdaging voor KI-systemen op basis van neurale netwerken. Trainingsprocessen met neurale netwerken kunnen leiden tot netwerkparameters die zijn

[52] Hoewel sommige van deze methoden momenteel al beschikbaar zijn, is voor andere nog verder onderzoek nodig. De gebieden waarop verder onderzoek nodig is, worden ook gebruikt als input voor het tweede product van de AI HLEG: de aanbevelingen voor beleid en investeringen.

ingesteld op numerieke waarden die lastig te verenigen zijn met resultaten. Bovendien kunnen kleine veranderingen in de gegevenswaarden soms leiden tot drastische veranderingen in de interpretatie, waardoor het systeem bijvoorbeeld een schoolbus met een struisvogel verwart. Van deze kwetsbaarheid kan ook gebruik worden gemaakt tijdens aanvallen op het systeem. Methoden met VKI-onderzoek zijn cruciaal, niet alleen om het gedrag van het systeem te verklaren tegenover gebruikers, maar ook om betrouwbare technologie te installeren.

- *Testen en valideren*

(100) Vanwege het non-deterministische en contextgebonden karakter van KI-systemen is traditioneel testen niet voldoende. Fouten van de door het systeem gebruikte concepten en representaties komen mogelijk alleen voor wanneer een programma op gegevens wordt toegepast die realistisch genoeg zijn. Om de verwerking van gegevens te controleren en valideren moet het onderliggende model daarom tijdens zowel de training als de installatie zorgvuldig worden gemonitord op stabiliteit, robuustheid en werking binnen heldere en voorspelbare grenzen. Er moet worden gezorgd dat het resultaat van het planningsproces consistent is met de input en dat de beslissingen op dusdanige wijze worden genomen dat het onderliggende proces kan worden gevalideerd.

(101) Het testen en valideren van het systeem moet zo vroeg mogelijk gebeuren om te zorgen dat het systeem zich gedurende de volledige levenscyclus en met name na de installatie gedraagt zoals bedoeld. Alle componenten van een KI-systeem moeten erin worden opgenomen, met inbegrip van gegevens, vooraf getrainde modellen, omgevingen en het gedrag van het systeem als geheel. Het systeem moet worden ontworpen en uitgevoerd door een zo divers mogelijke groep mensen. Er moeten meerdere maatstaven worden ontwikkeld voor de categorieën die vanuit verschillende oogpunten worden getest. Kwaadwillig testen door vertrouwde en diverse "red teams" die opzettelijk proberen het systeem te "breken" om kwetsbaarheden te vinden, alsook "bug bounties" die buitenstaanders stimuleren om fouten en zwakheden in het systeem op te sporen en op verantwoordelijke wijze te melden, kan worden overwogen. Tot slot moet er worden gezorgd dat de resultaten of handelingen consistent zijn met de resultaten van de voorgaande processen door ze te vergelijken met het eerder vastgestelde beleid, zodat dit niet wordt geschonden.

- *Kwaliteit van de dienstindicatoren*

(102) Er kan worden beschreven wat de geschikte kwaliteit van de dienstindicatoren is, om te zorgen dat men in de basis begrijpt of ze zijn getest en ontwikkeld met beveiligings- en veiligheidsoverwegingen in het achterhoofd. Bij deze indicatoren kunnen maatregelen worden opgenomen om het testen en trainen van algoritmen te evalueren, alsook maatstaven van traditionele software voor: functionaliteit; prestaties; bruikbaarheid; betrouwbaarheid; veiligheid; en onderhoudbaarheid.

2. Niet-technische methoden

(103) In dit gedeelte wordt een scala aan niet-technische methoden beschreven die een waardevolle rol kunnen spelen bij het waarborgen en in stand houden van betrouwbare KI. Ook deze methoden moeten **voortdurend** worden geëvalueerd.

- *Regelgeving*

(104) Zoals hierboven genoemd bestaat er momenteel al regelgeving ter ondersteuning van betrouwbare KI – denk bijvoorbeeld aan wetgeving inzake productveiligheid en kaders voor aansprakelijkheid. Voor zover er in onze optiek regelgeving moet worden herzien, aangepast of ingevoerd – om betrouwbare KI te waarborgen of mogelijk te maken – geven we dit aan in ons tweede product, dat bestaat uit aanbevelingen voor KI-beleid en -investeringen.

- *Gedragscodes*

(105) Organisaties en belanghebbenden kunnen zich aansluiten bij deze richtsnoeren door hun handvest van maatschappelijke verantwoordelijkheid, hun kernprestatie-indicatoren (KPI's), hun gedragscodes of hun internationale beleidsdocumenten aan te passen door toe te voegen dat ze streven naar betrouwbare KI. Meer in het algemeen kunnen organisaties die aan een KI-systeem werken, hun bedoelingen documenteren en deze onderschrijven met normen voor bepaalde wenselijke waarden, zoals grondrechten, transparantie en het voorkomen van schade.

- *Normalisatie*

(106) Normen, bijvoorbeeld voor ontwerp-, productie- en bedrijfspraktijken, kunnen voor KI-gebruikers, consumenten, organisaties, onderzoeksinstellingen en overheden als een kwaliteitsbeheersysteem fungeren door de mogelijkheid te bieden tot het herkennen en stimuleren van ethisch gedrag via hun aankoopbeslissingen. Naast traditionele normen bestaan er ook medereguleringsbenaderingen: accreditatiesystemen, professionele ethische codes of normen voor ontwerp conform de grondrechten. Huidige voorbeelden zijn ISO-normen of de normen uit de IEEE P7000-reeks, maar in de toekomst zou een label voor "betrouwbare KI" nuttig kunnen zijn om, aan de hand van specifieke technische normen, te bevestigen dat het systeem bijvoorbeeld voldoet aan de normen voor veiligheid, technische robuustheid en verklaarbaarheid.

- *Certificering*

(107) Omdat niet kan worden verwacht dat iedereen de werking en gevolgen van KI-systemen volledig kan begrijpen, kunnen er organisaties in overweging worden genomen die aan het grotere publiek kunnen bevestigen dat een KI-systeem transparant, verantwoordelijk en rechtvaardig is[53]. Voor deze certificeringen zouden normen worden gebruikt die voor verschillende toepassingsdomeinen en KI-technieken zijn ontwikkeld en goed zijn afgestemd op de sectorale en maatschappelijke normen van de betreffende context. Certificering kan echter nooit in de plaats komen van verantwoordelijkheid. Daarom moet deze worden aangevuld met verantwoordingskaders, met inbegrip van disclaimers en mechanismen voor beoordeling en sanering[54].

- *Verantwoording via governancekaders*

(108) Organisaties moeten zowel interne als externe governancekaders creëren om verantwoording te waarborgen voor de ethische dimensies van beslissingen in verband met de ontwikkeling, de installatie en het gebruik van KI. Daarbij kan het bijvoorbeeld gaan om de aanstelling van iemand die gaat over ethische kwesties in verband met KI, of van een interne/externe ethische commissie of raad. Een van de mogelijke taken van een dergelijke persoon, commissie of raad bestaat uit het houden van toezicht en het geven van advies. Zoals hierboven beschreven kunnen certificeringsspecificaties hier ook een rol in spelen. Er moet worden gezorgd voor communicatiekanalen met de bedrijfstak en/of groepen voor publiek toezicht, zodat beste praktijken kunnen worden uitgewisseld, dilemma's kunnen worden besproken of opkomende ethisch zorgelijke kwesties kunnen worden gemeld. Dergelijke mechanismen kunnen dienen ter aanvulling op, maar niet ter vervanging van wettelijk toezicht (bijv. in de vorm van de aanstelling van een functionaris voor gegevensbescherming of vergelijkbare maatregelen die wettelijk verplicht zijn op grond van het gegevensbeschermingsrecht).

- *Onderwijs en bewustzijn ter bevordering van een ethische denkwijze*

(109) Voor betrouwbare KI is het belangrijk dat alle belanghebbenden deelnemen en goed geïnformeerd zijn. Communicatie, onderwijs en opleiding zijn daarbij van groot belang om te zorgen dat kennis van de potentiële gevolgen van KI-systemen wijdverbreid is en om mensen bewust te maken van het feit dat ze de

[53] Zoals voorgesteld door bijvoorbeeld het initiatief Ethically Aligned Design (Ethisch afgestemd ontwerp) van het IEEE: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Zie voor meer informatie over de beperkingen van certificering: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

maatschappelijke ontwikkeling mede vorm kunnen geven. Het gaat daarbij om alle belanghebbenden, zoals degenen die betrokken zijn bij het maken van de producten (de ontwerpers en ontwikkelaars), de gebruikers (bedrijven of individuele personen) en andere beïnvloede groepen (die misschien geen KI-systeem aanschaffen of gebruiken, maar voor wie beslissingen worden gemaakt door een KI-systeem, alsook de samenleving als geheel). Basiskennis op het gebied van KI moet in de gehele samenleving worden bevorderd. Voor het onderwijzen van het grote publiek is het noodzakelijk dat ethici van gedegen vaardigheden en opleiding worden voorzien op dit gebied.

- *Participatie van belanghebbenden en sociale dialoog*

(110) KI heeft veel voordelen en Europa moet zorgen dat deze voor iedereen beschikbaar zijn. Daarvoor is een open debat en de betrokkenheid van sociale partners, belanghebbenden en het grote publiek nodig. Veel organisaties maken al gebruik van panels van belanghebbenden voor de bespreking van het gebruik van KI-systemen en gegevensanalyse. In deze panels zetelen verschillende soorten leden, zoals juridische deskundigen, technische deskundigen, ethici, vertegenwoordigers van consumenten en werknemers. Door participatie en de dialoog op het gebied van het gebruik en de gevolgen van KI-systemen te stimuleren wordt de evaluatie van resultaten en methoden ondersteund. Dit kan in het bijzonder nuttig zijn in complexe gevallen.

- *Diversiteit en inclusieve ontwerpteams*

(111) Diversiteit en inclusie spelen een cruciale rol bij het ontwikkelen van KI-systemen voor gebruik in de echte wereld. Naarmate KI-systemen meer taken zelfstandig uitvoeren, is het essentieel dat de teams die deze systemen ontwerpen, ontwikkelen, testen, onderhouden, installeren en/of aanschaffen, dezelfde diversiteit vertonen als de gebruikers en de samenleving in het algemeen. Op die manier wordt bijgedragen aan de objectiviteit en wordt er rekening gehouden met verschillende perspectieven, behoeften en doelen. In het ideale scenario zijn de teams niet alleen divers wat geslacht, cultuur en leeftijd betreft, maar ook als het gaat om professionele achtergrond en vaardigheden.

Essentiële richtsnoeren uit hoofdstuk II:

- ✓ Zorg dat het KI-systeem gedurende de gehele levenscyclus voldoet aan de vereisten voor betrouwbare KI: 1) menselijke controle en menselijk toezicht, 2) technische robuustheid en veiligheid, 3) privacy en datagovernance, 4) transparantie, 5) diversiteit, non-discriminatie en rechtvaardigheid, 6) milieu- en maatschappelijk welzijn en 7) verantwoordingsplicht.
- ✓ Bestudeer technische en niet-technische methoden om te waarborgen dat die vereisten worden verwezenlijkt.
- ✓ Stimuleer onderzoek en innovatie om aan het controleren van KI-systemen bij te dragen en het verwezenlijken van de vereisten te bevorderen. Verspreid de resultaten en open vragen onder een breder publiek en leid systematisch een nieuwe generatie deskundigen op het gebied van KI-ethiek op.
- ✓ Verspreid informatie over de mogelijkheden en beperkingen van een KI-systeem en over de manier waarop de vereisten worden verwezenlijkt, op een heldere en proactieve manier onder belanghebbenden om onrealistische verwachtingen te voorkomen. Wees open over het feit dat ze met een KI-systeem te maken hebben.
- ✓ Faciliteer de traceerbaarheid en controleerbaarheid van KI-systemen, met name in kritieke omgevingen en situaties.
- ✓ Betrek belanghebbenden bij de volledige levenscyclus van het KI-systeem. Stimuleer onderwijs en opleiding, zodat alle belanghebbenden zich bewust zijn van betrouwbare KI en opgeleid zijn op dat gebied.
- ✓ Wees bedacht op conflicterende beginselen en vereisten. Zoek, evalueer en documenteer deze afwegingen en de oplossingen voortdurend en maak ze kenbaar.

III. Hoofdstuk III: Betrouwbare KI controleren

(112) Op basis van de kernvereisten uit hoofdstuk II wordt in dit hoofdstuk een niet-uitputtende **controlelijst voor betrouwbare KI** (testversie) gegeven voor de **operationalisering van betrouwbare KI**. Deze lijst is in het bijzonder van toepassing op KI-systemen met rechtstreekse interactie met gebruikers en is met name bedoeld voor ontwikkelaars en installateurs van KI-systemen (ongeacht of zij het systeem zelf hebben ontwikkeld of het van derden hebben verkregen). In deze controlelijst wordt niet ingegaan op de operationalisering van de eerste component van betrouwbare KI (wettige KI). Het volgen van deze controlelijst vormt geen bewijs dat de wetgeving wordt nageleefd. De lijst is ook niet bedoeld als richtsnoer voor de naleving van de toepasselijke wetgeving. Omdat KI-systemen sterk toepassingsgebonden zijn, moet de controlelijst worden afgestemd op de specifieke gebruikssituatie en -context waarbinnen een systeem werkzaam is. Verder wordt in dit hoofdstuk een algemene aanbeveling gedaan over de manier waarop de controlelijst voor betrouwbare KI kan worden toegepast via een governancestructuur waarin zowel het operationele als het bestuurlijke niveau is opgenomen.

(113) De controlelijst en governancestructuur worden ontwikkeld in nauwe samenwerking met belanghebbenden uit de publieke en particuliere sector. Het proces wordt uitgevoerd als een testproces, zodat er uitgebreide feedback mogelijk is vanuit twee parallelle processen:

a. een kwalitatief proces dat de representativiteit waarborgt, waarbij een klein aantal geselecteerde bedrijven, organisaties en instellingen (uit verschillende sectoren en van verschillende grootte) zich aanmeldt om de controlelijst en de governancestructuur in de praktijk te testen en diepgaande feedback te geven;

b. een kwantitatief proces, waarbij alle geïnteresseerde belanghebbenden zich kunnen aanmelden om de controlelijst te testen en feedback te geven via een openbare raadpleging.

(114) Na de testfase integreren wij de resultaten van het feedbackproces in de controlelijst en stellen we begin 2020 een herziene versie op. Het doel is om een kader te creëren dat horizontaal kan worden gebruikt voor alle toepassingen en dat dus een grondslag biedt voor de verwezenlijking van betrouwbare KI in alle domeinen. Wanneer een dergelijk kader eenmaal is vastgesteld, zou er een sectoraal of toepassingsspecifiek kader kunnen worden ontwikkeld.

Governance

(115) Het kan nuttig zijn voor bedrijven, organisaties en instellingen om na te denken over de manier waarop de controlelijst voor betrouwbare KI binnen hun organisatie kan worden toegepast. Dit kan door het controleproces in de bestaande governancemechanismen op te nemen of door nieuwe processen in te voeren. Die keuze hangt af van de interne structuur, de omvang en de beschikbare middelen van de organisatie.

(116) Uit onderzoek[55] is gebleken dat aandacht van het hoogste management essentieel is om verandering teweeg te brengen. Ook is aangetoond dat de acceptatie en relevantie van het invoeren van nieuwe processen (al dan niet technologisch) kunnen worden bevorderd door alle belanghebbenden binnen een bedrijf, organisatie of instelling erbij te betrekken[56]. Daarom raden wij aan om een proces in te voeren waarbij zowel het operationele niveau als het hoogste bestuurlijke niveau betrokken zijn.

---

55 https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation

56 Zie bijvoorbeeld A. Bryson, E. Barth en H. Dale-Olsen, The Effects of Organisational change on worker well-being and the moderating role of trade unions, *ILRReview*, 66(4), juli 2013; Jirjahn, U. en Smith, S.C. (2006). "What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations", 45(4), blz. 650–680; Michie, J. en Sheehan, M. (2003). "Labour market deregulation, 'flexibility' and innovation", *Cambridge Journal of Economics*, 27(1), blz. 123–143.

Niveau	Relevante taken (afhankelijk van de organisatie)
Management en raad	Het hoogste management bespreekt en evalueert de ontwikkeling, installatie of aankoop van KI en dient als escalatieraad voor de evaluatie van alle KI-innovaties en -toepassingen wanneer er punten van zorg naar voren komen. Het betrekt degenen op wie de eventuele invoering van KI-systemen van invloed zou zijn (bijv. werknemers), alsook hun vertegenwoordigers, bij het hele proces via informatie-, raadplegings- en participatieprocedures.
Afdeling naleving/Juridische afdeling/Afdeling maatschappelijke verantwoordelijkheid	De afdeling verantwoordelijkheid monitort het gebruik van de controlelijst en de noodzakelijke ontwikkeling ervan voor aanpassing aan de veranderingen in technologie of regelgeving. De afdeling werkt de normen of het interne beleid inzake KI-systemen bij en zorgt ervoor dat bij het gebruik van dergelijke systemen het huidige wettelijke en regelgevingskader en de waarden van de organisatie worden gevolgd.
Product- en dienstenontwikkeling of gelijkwaardig	De afdeling product- en dienstenontwikkeling gebruikt de controlelijst om op KI gebaseerde producten en diensten te evalueren en houdt alle resultaten bij. Deze resultaten worden op managementniveau besproken. Daar worden de nieuwe of herziene op KI gebaseerde toepassingen uiteindelijk ook goedgekeurd.
Kwaliteitsborging	De afdeling kwaliteitsborging (of gelijkwaardig) waarborgt en controleert de resultaten van de controlelijst en neemt maatregelen om een kwestie naar een hoger niveau te tillen indien het resultaat niet bevredigend is of er onvoorziene resultaten worden ontdekt.
HR	De HR-afdeling zorgt voor de juiste combinatie van vaardigheden en diversiteit binnen de profielen van de ontwikkelaars van KI-systemen. Ook zorgt de afdeling ervoor dat het juiste niveau van opleiding op het gebied van betrouwbare KI wordt verzorgd binnen de organisatie.
Inkoop	De afdeling inkoop zorgt ervoor dat in het inkoopproces voor op KI gebaseerde producten en diensten een controle op betrouwbare KI is opgenomen.
Dagelijkse werkzaamheden	Ontwikkelaars en projectmanagers nemen de controlelijst op in hun dagelijkse werkzaamheden en documenteren de resultaten en uitkomsten van de controle.

De controlelijst voor betrouwbare KI gebruiken

(117) Bij het gebruik van de controlelijst in de praktijk raden wij aan niet alleen aandacht te besteden aan de punten van zorg, maar ook aan de vragen die niet (gemakkelijk) kunnen worden beantwoord. Eén potentieel probleem is het gebrek aan diversiteit wat betreft de vaardigheden en competenties binnen het team dat het KI-systeem ontwikkelt en test. Daarom kan het nodig zijn andere belanghebbenden van binnen of buiten de organisatie erbij te betrekken. Het is sterk aan te raden om alle resultaten bij te houden in zowel technische als bestuurlijke termen, zodat de aanpak van het probleem op alle niveaus van de governancestructuur kan worden begrepen.

(118) De controlelijst is bedoeld als leidraad voor beroepsbeoefenaars op het gebied van KI bij het ontwikkelen, installeren en gebruiken van betrouwbare KI. De controle moet op evenredige wijze worden afgestemd op de specifieke gebruikssituatie. Gedurende de testfase kunnen er specifieke gevoelige gebieden naar voren komen. Bij de volgende stap wordt de noodzaak van verdere specificatie in dergelijke gevallen

geëvalueerd. Deze controlelijst biedt geen concrete antwoorden op de ontstane vragen, maar stimuleert wel reflectie op de stappen die kunnen bijdragen aan de verwezenlijking van de betrouwbaarheid van KI-systemen, en op de potentiële stappen die met betrekking daartoe moeten worden genomen.

Verband met bestaande wetgeving en processen

(119) Het is ook belangrijk dat degenen die betrokken zijn bij de ontwikkeling, de installatie en het gebruik van KI, erkennen dat er verschillende bestaande wetten zijn op grond waarvan bepaalde processen verplicht zijn en bepaalde resultaten verboden, en die (gedeeltelijk) samen kunnen vallen met enkele van de maatregelen op de controlelijst. De wetgeving inzake gegevensbescherming bevat bijvoorbeeld een reeks wettelijke voorschriften die moeten worden nageleefd door degenen die betrokken zijn bij de verzameling en verwerking van persoonsgegevens. Voor betrouwbare KI is het echter ook nodig dat gegevens ethisch worden verwerkt. Met interne procedures en intern beleid gericht op het waarborgen van de naleving van wetgeving inzake gegevensbescherming kan dit mogelijk worden gemaakt en kunnen zo bestaande wettelijke processen worden aangevuld. Het volgen van deze controlelijst vormt echter *geen* bewijs dat de wetgeving wordt nageleefd. De lijst is ook niet bedoeld als richtsnoer voor de naleving van toepasselijke wetgeving. Het doel van de controlelijst is veeleer het aanleveren van een reeks specifieke vragen voor ontvangers die proberen te zorgen dat hun benadering van de ontwikkeling en installatie van KI gericht is op (de verwezenlijking van) betrouwbare KI.

(120) Veel beroepsbeoefenaars op het gebied van KI beschikken ook al over bestaande controlehulpmiddelen en softwareontwikkelingsprocessen om naleving, ook van niet-wettelijke normen, te waarborgen. De onderstaande controle hoeft niet noodzakelijkerwijs los te worden uitgevoerd, maar kan in dergelijke bestaande praktijken worden verwerkt.

CONTROLELIJST VOOR BETROUWBARE KI (TESTVERSIE)

1. Menselijke controle en menselijk toezicht

Grondrechten:

- ✓ Hebt u, in de gebruikssituaties waarin er mogelijk sprake kon zijn van negatieve gevolgen voor de grondrechten, een effectbeoordeling op het gebied van grondrechten uitgevoerd? Hebt u mogelijke afwegingen tussen de verschillende beginselen en rechten vastgesteld en gedocumenteerd?
- ✓ Is er sprake van een wisselwerking tussen het KI-systeem en de beslissingen van menselijke eindgebruikers (bijv. aanbevolen handelingen of beslissingen, aanbod van opties)?
  - Is er in die gevallen een risico dat het KI-systeem de menselijke autonomie aantast door het beslissingsproces van de eindgebruiker op een onbedoelde manier te beïnvloeden?
  - Hebt u overwogen of het KI-systeem aan de gebruikers moet aangeven dat inhoud of een beslissing, advies of uitkomst het resultaat is van een algoritmische beslissing?
  - Indien het KI-systeem een chatbot of gesprekssysteem bevat: zijn de menselijke eindgebruikers in kennis gesteld van het feit dat ze te maken hebben met een niet-menselijke actor?

Menselijke controle:

- ✓ Indien het KI-systeem in de werk- en arbeidsprocessen is opgenomen: hebt u nagedacht over de taakverdeling tussen het KI-systeem en de menselijke werknemers om te zorgen voor zinvolle

interactie en een gepaste mate van menselijk toezicht en menselijke controle?

- Versterkt of vergroot het KI-systeem menselijke capaciteiten?
- Hebt u voorzorgsmaatregelen genomen om overmatig vertrouwen in of op het KI-systeem bij werkprocessen te voorkomen?

Menselijk toezicht:

✓ Hebt u nagedacht over de gepaste mate van menselijke controle voor dit specifieke KI-systeem en deze specifieke gebruikssituatie?

- Kunt u, indien van toepassing, de mate van menselijke controle of betrokkenheid beschrijven? Wie is de "human in control" en wat zijn de momenten of hulpmiddelen voor menselijke interventie?
- Hebt u mechanismen en maatregelen ingesteld om dergelijke potentiële menselijke controle of menselijk toezicht te waarborgen of om te zorgen dat beslissingen onder de algehele verantwoordelijkheid van mensen worden genomen?
- Hebt u maatregelen genomen om controle mogelijk te maken en problemen in verband met het beheren van KI-autonomie te verhelpen?

✓ Indien het KI-systeem of de gebruikssituatie zelflerend of autonoom is: hebt u specifiekere controle- en toezichtsmechanismen ingesteld?

- Wat voor detectie- en responsmechanismen hebt u ingesteld om te controleren of er iets mis kan gaan?
- Hebt u gezorgd voor een "stopknop" of procedure waarmee een activiteit indien nodig veilig kan worden afgebroken? Wordt met deze procedure het proces volledig afgebroken, wordt het gedeeltelijk afgebroken of wordt de controle overgedragen aan een mens?

2. <u>Technische robuustheid en veiligheid</u>

Weerbaarheid tegen aanvallen en beveiliging:

✓ Hebt u potentiële soorten aanvallen onderzocht waarvoor het KI-systeem kwetsbaar zou kunnen zijn?

- Hebt u in het bijzonder nagedacht over verschillende soorten kwetsbaarheid, zoals gegevensverontreiniging, fysieke infrastructuur en cyberaanvallen?

✓ Hebt u maatregelen of systemen ingesteld om de integriteit van het KI-systeem en de weerbaarheid tegen potentiële aanvallen te waarborgen?

✓ Hebt u onderzocht hoe uw systeem zich in onverwachte situaties en omgevingen gedraagt?

✓ Hebt u overwogen of en in hoeverre uw systeem geschikt zou kunnen zijn voor dubbel gebruik? Zo ja, hebt u geschikte preventieve maatregelen genomen tegen deze situatie (met inbegrip van bijvoorbeeld het niet publiceren van het onderzoek of het niet installeren van het systeem)?

Uitwijkplan en algemene veiligheid:

- ✓ Hebt u gezorgd dat uw systeem over een toereikend uitwijkplan beschikt, mocht het te maken krijgen met kwaadwillige aanvallen of andere onverwachte situaties (bijv. technische omschakelingsprocedures of vragen om een menselijke beheerder alvorens verder te gaan)?
- ✓ Hebt u nagedacht over de omvang van het risico dat het KI-systeem in deze specifieke gebruikssituatie met zich meebrengt?
  - Hebt u een proces ingesteld om de risico's en de veiligheid te meten en te controleren?
  - Hebt u de noodzakelijke informatie aangeleverd in geval van een risico voor de menselijke fysieke integriteit?
  - Hebt u een verzekering overwogen voor eventuele schade door het KI-systeem?
  - Hebt u de potentiële veiligheidsrisico's vastgesteld van (andere) voorzienbare toepassingen van de technologie, met inbegrip van onbedoeld of kwaadwillig misbruik ervan? Is er een plan om deze risico's te beperken of te beheren?
- ✓ Hebt u onderzocht of er een goede kans is dat het KI-systeem schade veroorzaakt voor gebruikers of derden? Zo ja, hebt u de waarschijnlijkheid, de potentiële schade, het getroffen publiek en de ernst onderzocht?
  - Indien er een risico bestaat dat het KI-systeem schade veroorzaakt: hebt u nagedacht over regels voor aansprakelijkheid en bescherming van consumenten? Hoe hebt u daar rekening mee gehouden?
  - Hebt u nagedacht over de potentiële gevolgen of het veiligheidsrisico voor het milieu of voor dieren?
  - Hebt u in uw risicoanalyse meegenomen of beveiligings- of netwerkproblemen (zoals gevaren voor de cyberbeveiliging) veiligheidsrisico's of schade met zich meebrengen vanwege onbedoeld gedrag van het KI-systeem?
- ✓ Hebt u de vermoedelijke gevolgen ingeschat van een fout van uw KI-systeem met als gevolg de levering van verkeerde resultaten waardoor uw systeem niet meer beschikbaar is of maatschappelijk onaanvaardbare resultaten oplevert (bijv. discriminerende praktijken)?
  - Hebt u drempels en governance vastgesteld voor de bovenstaande scenario's, zodat er alternatieve of uitwijkplannen in werking worden gesteld?
  - Hebt u uitwijkplannen vastgesteld en getest?

Nauwkeurigheid

- ✓ Hebt u onderzocht welke mate en definitie van nauwkeurigheid nodig zouden zijn in de context van het KI-systeem en in de gebruikssituatie?
  - Hebt u onderzocht hoe de nauwkeurigheid wordt gemeten en gegarandeerd?
  - Hebt u maatregelen ingesteld om te zorgen dat de gebruikte gegevens volledig en actueel zijn?
  - Hebt u maatregelen ingesteld om te controleren of er aanvullende gegevens nodig zijn,

bijvoorbeeld om de nauwkeurigheid te vergroten of vertekening weg te nemen?

- ✓ Hebt u de schade onderzocht die zou worden veroorzaakt als het KI-systeem onnauwkeurige voorspellingen doet?
- ✓ Hebt u voor manieren gezorgd om te meten of uw systeem een onaanvaardbaar aantal onnauwkeurige voorspellingen doet?
- ✓ Indien er onnauwkeurige voorspellingen worden gedaan: hebt u een reeks stappen vastgesteld om het probleem te verhelpen?

Betrouwbaarheid en reproduceerbaarheid:

- ✓ Hebt u een strategie vastgesteld om te monitoren en testen of het KI-systeem in overeenstemming is met de doelen, doelstellingen en beoogde toepassingen?
  - Hebt u getest of er rekening moet worden gehouden met specifieke situaties of omstandigheden om de reproduceerbaarheid te waarborgen?
  - Hebt u controleprocessen of -methoden vastgesteld om verschillende aspecten van betrouwbaarheid en reproduceerbaarheid te meten en te waarborgen?
  - Hebt u processen vastgesteld om te beschrijven wanneer een KI-systeem bij bepaalde soorten instellingen een fout maakt?
  - Hebt u deze processen voor het testen en controleren van de betrouwbaarheid van KI-systemen duidelijk gedocumenteerd en geoperationaliseerd?

    Hebt u gezorgd voor mechanismen of communicatie om (eind)gebruikers te garanderen dat het KI-systeem betrouwbaar is?

3. <u>Privacy en datagovernance</u>

Respect voor privacy en gegevensbescherming:

- ✓ Hebt u, afhankelijk van de gebruikssituatie, mechanismen vastgesteld waardoor anderen problemen in verband met privacy of gegevensbescherming kunnen melden die betrekking hebben op de processen van het KI-systeem in verband met gegevensverzameling (voor zowel training als werking) en gegevensverwerking?
- ✓ Hebt u het type en de reikwijdte van de gegevens in uw gegevenssets onderzocht (bijv. of ze persoonsgegevens bevatten)?
- ✓ Hebt u nagedacht over manieren om het KI-systeem te ontwikkelen of het model te trainen zonder of met zo weinig mogelijk gebruik van potentieel gevoelige gegevens of persoonsgegevens?
- ✓ Hebt u, afhankelijk van de gebruikssituatie, mechanismen ingebouwd voor het opmerken van en de controle over persoonsgegevens (zoals geldige toestemming en de mogelijkheid om deze in te trekken, indien van toepassing)?
- ✓ Hebt u maatregelen genomen om de privacy te vergroten, bijvoorbeeld via encryptie, anonimisering en aggregatie?

- ✓ Indien er een functionaris voor gegevensbescherming is: hebt u deze persoon in een vroeg stadium bij het proces betrokken?

Kwaliteit en integriteit van gegevens:

- ✓ Hebt u uw systeem afgestemd op de potentieel relevante normen (bijv. ISO, IEEE) of veelgebruikte protocollen voor uw dagelijkse gegevensbeheer en governance?
- ✓ Hebt u toezichtsmechanismen vastgesteld voor de verzameling, de opslag, de verwerking en het gebruik van gegevens?
- ✓ Hebt u onderzocht in hoeverre u controle heeft over de kwaliteit van de gebruikte externe gegevensbronnen?
- ✓ Hebt u processen vastgesteld om de kwaliteit en integriteit van uw gegevens te waarborgen? Hebt u andere processen overwogen? Hoe controleert u of uw gegevenssets niet zijn aangetast of gehackt?

Toegang tot gegevens:

- ✓ Welke protocollen, processen en procedures zijn er gevolgd voor het beheer en de waarborging van goede datagovernance?
  - Hebt u onderzocht wie er toegang heeft tot gebruikersgegevens en onder welke omstandigheden?
  - Hebt u gecontroleerd of deze personen gekwalificeerd zijn om de gegevens in te zien, of het noodzakelijk is dat zij de gegevens inzien en of zij over de benodigde vaardigheden beschikken om de details van het gegevensbeschermingsbeleid te begrijpen?
  - Hebt u gezorgd voor een toezichtsmechanisme om bij te houden wanneer, waar, hoe, door wie en met welk doel gegevens worden ingezien?

4. <u>Transparantie</u>

Traceerbaarheid:

- ✓ Hebt u maatregelen ingesteld waarmee de traceerbaarheid kan worden gewaarborgd? Daarbij kan het gaan om de documentatie van:
  - ➢ de voor het ontwerp en de ontwikkeling van het algoritmische systeem gebruikte methoden:
    - o in geval van een KI-systeem op basis van regels moet de programmeermethode of de manier waarop het model is gebouwd, worden gedocumenteerd;
    - o in geval van een KI-systeem op basis van leren, moet de trainingsmethode van het algoritme, met inbegrip van welke input er is verzameld en geselecteerd en de manier waarop dit is gebeurd, worden gedocumenteerd.
  - ➢ de voor het testen en valideren van het algoritmische systeem gebruikte methoden:
    - o in geval van een KI-systeem op basis van regels, moeten de voor het testen en valideren gebruikte scenario's of situaties worden gedocumenteerd;
    - o in geval van een model op basis van leren, moet de voor het testen en valideren

gebruikte informatie worden gedocumenteerd.

- de resultaten van het algoritmische systeem:
  - de resultaten of de door het algoritme genomen beslissingen, alsook potentiële andere beslissingen die in andere situaties zouden ontstaan (bijv. voor andere subgroepen of gebruikers) moeten worden gedocumenteerd.

Verklaarbaarheid:

- ✓ Hebt u onderzocht in hoeverre de door het KI-systeem genomen beslissingen en dus het resultaat kunnen worden begrepen?
- ✓ Hebt u gezorgd dat er voor alle gebruikers die een verklaring wensen, een verklaring begrijpelijk kan worden gemaakt van de redenen dat een systeem een bepaalde keuze heeft gemaakt die tot een bepaalde uitkomst heeft geleid?
- ✓ Hebt u onderzocht in hoeverre de beslissingen van het systeem van invloed zijn op het besluitvormingsproces binnen de organisatie?
- ✓ Hebt u onderzocht waarom dit specifieke systeem op dit specifieke gebied is ingezet?
- ✓ Hebt u het bedrijfsmodel van dit systeem onderzocht (bijv. hoe het waarde creëert voor de organisatie)?
- ✓ Hebt u het KI-systeem vanaf het begin ontworpen met interpreteerbaarheid in het achterhoofd?
  - Hebt u onderzocht welk model het eenvoudigst en het meest interpreteerbaar zou zijn voor de betreffende toepassing, en hebt u geprobeerd dit te gebruiken?
  - Hebt u gecontroleerd of u uw trainings- en testgegevens kunt analyseren? Kunt u deze in de loop van de tijd veranderen en bijwerken?
  - Hebt u onderzocht of u, na de training en ontwikkeling van het model, mogelijkheden hebt om de interpreteerbaarheid ervan te beoordelen en of u toegang hebt tot de interne werkstroom van het model?

Communicatie:

- ✓ Hebt u – via een disclaimer of op een andere manier – aan de (eind)gebruikers kenbaar gemaakt dat ze niet met een andere mens, maar met een KI-systeem te maken hebben? Hebt u uw KI-systeem als zodanig aangemerkt?
- ✓ Hebt u mechanismen ingesteld om gebruikers in kennis te stellen van de redenen en criteria achter de resultaten van het KI-systeem?
  - Is dit duidelijk en begrijpelijk aan de beoogde gebruikers kenbaar gemaakt?
  - Hebt u processen ingesteld waardoor rekening wordt gehouden met de feedback van gebruikers, en gebruikt u deze om het systeem aan te passen?
  - Hebt u ook gecommuniceerd over mogelijke of ervaren risico's, zoals vertekening?
  - Hebt u, afhankelijk van de gebruikssituatie, ook nagedacht over de communicatie en transparantie richting andere doelgroepen, derden of het grote publiek?
- ✓ Hebt u duidelijk gemaakt wat het doel van het KI-systeem is en wie of wat er profijt kan hebben van

het product of de dienst?

- Zijn de gebruiksscenario's van het product beschreven en duidelijk kenbaar gemaakt, en zijn daarbij ook alternatieve communicatievormen overwogen om te zorgen dat de boodschap begrijpelijk en geschikt is voor de gebruiker aan wie deze is gericht?

- Hebt u, afhankelijk van de gebruikssituatie, nagedacht over de menselijke psychologie en potentiële beperkingen, zoals het risico op verwarring, voorkeur voor bevestiging of cognitieve vermoeidheid?

✓ Hebt u de kenmerken, beperkingen en potentiële tekortkomingen van het KI-systeem duidelijk kenbaar gemaakt:

- in geval van ontwikkeling: aan degene die het installeert in een product of dienst?

- in geval van installatie: aan de eindgebruiker of consument?

5. <u>Diversiteit, non-discriminatie en rechtvaardigheid</u>

Voorkomen van onrechtvaardige vertekening:

✓ Hebt u gezorgd voor een strategie of een reeks procedures om te voorkomen dat er onrechtvaardige vertekening wordt gecreëerd of versterkt in het KI-systeem, met betrekking tot zowel het gebruik van inputgegevens als het ontwerp van het algoritme?

- Hebt u de mogelijke beperkingen die voortkomen uit de samenstelling van de gebruikte gegevenssets, onderzocht en erkend?

- Hebt u nagedacht over de diversiteit en de representativiteit van gebruikers in de gegevens? Hebt u getest op specifieke populaties of problematische gebruikssituaties?

- Hebt u onderzoek gedaan naar technische hulpmiddelen om uw begrip van de gegevens, het model en de prestaties te verbeteren, en hebt u deze gebruikt?

- Hebt u processen ingesteld voor het testen en monitoren van potentiële vertekening gedurende de ontwikkelings-, installatie- en gebruiksfase van het systeem?

✓ Hebt u, afhankelijk van de gebruikssituatie, gezorgd voor een mechanisme waardoor anderen problemen in verband met vertekening, discriminatie of slechte prestaties van het KI-systeem kunnen melden?

- Hebt u nagedacht over duidelijke stappen en manieren om te communiceren over hoe en aan wie dergelijke problemen kunnen worden gemeld?

- Hebt u niet alleen nagedacht over de (eind)gebruikers, maar ook over anderen die mogelijk indirect de gevolgen van het KI-systeem kunnen ondervinden?

✓ Hebt u onderzocht of er variabiliteit kan bestaan in de beslissingen die onder gelijke omstandigheden mogelijk zijn?

- Zo ja, hebt u nagedacht over de mogelijke oorzaken daarvan?

- In geval van variabiliteit: hebt u een meet- of controlemechanisme vastgesteld voor de potentiële

gevolgen van die variabiliteit voor de grondrechten?

✓ Hebt u gezorgd voor een geschikte werkdefinitie van "rechtvaardigheid" die u toepast bij het ontwerpen van KI-systemen?

- Wordt uw definitie veel gebruikt? Hebt u andere definities overwogen voordat u deze koos?
- Hebt u gezorgd voor een kwantitatieve analyse of maatstaf om de toegepaste definitie van rechtvaardigheid te meten en te testen?
- Hebt u mechanismen ingesteld om de rechtvaardigheid in uw KI-systemen te waarborgen? Hebt u andere mogelijke mechanismen overwogen?

Toegankelijkheid en universeel ontwerp:

✓ Hebt u gezorgd dat het KI-systeem geschikt is voor een breed scala aan individuele voorkeuren en vermogens?

- Hebt u onderzocht of het KI-systeem kan worden gebruikt door mensen met bijzondere behoeften of een beperking of door mensen die het risico lopen te worden uitgesloten? Hoe is dit in het ontwerp van het systeem verwerkt en hoe wordt het gecontroleerd?
- Hebt u gezorgd dat de informatie over het KI-systeem ook toegankelijk is voor gebruikers van ondersteunende technologieën?
- Hebt u deze gemeenschap betrokken bij de ontwikkelingsfase van het KI-systeem?

✓ Hebt u rekening gehouden met de gevolgen van uw KI-systeem voor de potentiële gebruikers?

- Is het team dat betrokken is bij de bouw van het KI-systeem, representatief voor uw beoogde gebruikers? Is het representatief voor de bevolking als geheel, ook rekening houdend met andere groepen die er mogelijk indirect de gevolgen van ondervinden?
- Hebt u onderzocht of er personen of groepen zijn die onevenredig kunnen worden getroffen door negatieve gevolgen?
- Hebt u feedback gekregen van andere teams of groepen die verschillende achtergronden en ervaringen vertegenwoordigen?

Participatie van belanghebbenden:

✓ Hebt u nagedacht over een mechanisme om de participatie van verschillende belanghebbenden onderdeel te maken van de ontwikkeling en het gebruik van het KI-systeem?

✓ Hebt u de invoering van het KI-systeem in uw organisatie voorbereid door de getroffen werknemers en hun vertegenwoordigers vooraf in kennis te stellen en bij het proces te betrekken?

6. <u>Maatschappelijk en milieuwelzijn</u>

Duurzame en milieuvriendelijke KI:

✓ Hebt u mechanismen ingesteld om de gevolgen van de ontwikkeling, de installatie en het gebruik van het KI-systeem voor het milieu te meten (bijv. de door het datacentrum gebruikte energie, het type

energie dat de datacentra gebruiken enz.)?

- ✓ Hebt u gezorgd voor maatregelen om de gevolgen van de levenscyclus van uw KI-systeem voor het milieu te beperken?

Sociale gevolgen:

- ✓ Indien er sprake is van rechtstreekse interactie tussen het KI-systeem en mensen:
  - Hebt u onderzocht of het KI-systeem mensen stimuleert om verbondenheid en empathie jegens het systeem te ontwikkelen?
  - Hebt u gezorgd dat het KI-systeem duidelijk aangeeft dat zijn sociale interactie gesimuleerd is en dat het niet tot "begrip" en "gevoelens" in staat is?
- ✓ Hebt u gezorgd voor een goed begrip van de sociale gevolgen van het KI-systeem? Hebt u bijvoorbeeld onderzocht of er een risico bestaat op het verlies van banen of vaardigheden onder de arbeidskrachten? Welke stappen zijn er genomen om dergelijke risico's tegen te gaan?

Samenleving en democratie:

- ✓ Hebt u de bredere maatschappelijke gevolgen van het gebruik van het KI-systeem voorbij de individuele (eind)gebruiker onderzocht, zoals potentieel indirect getroffen belanghebbenden?

7. Verantwoording

Controleerbaarheid:

- ✓ Hebt u mechanismen ingesteld die de controleerbaarheid van het systeem door interne en/of externe onafhankelijke actoren mogelijk maken, bijvoorbeeld in de vorm van waarborging van de traceerbaarheid en het bijhouden van de processen en resultaten van het KI-systeem?

Minimalisering en verslaglegging van negatieve gevolgen:

- ✓ Hebt u een risico- of effectbeoordeling van het KI-systeem uitgevoerd waarin rekening wordt gehouden met de verschillende belanghebbenden die direct of indirect de gevolgen ervan ondervinden?
- ✓ Hebt u onderwijs- en opleidingskaders gecreëerd voor de ontwikkeling van verantwoordingspraktijken?
  - Welke werknemers of delen van het team zijn erbij betrokken? Gaat het verder dan de ontwikkelingsfase?
  - Wordt binnen deze opleiding ook het potentiële wettelijke kader onderwezen dat van toepassing is op het KI-systeem?
  - Hebt u overwogen een "onderzoeksraad voor ethische KI" of een vergelijkbaar mechanisme in te stellen voor de bespreking van de algehele verantwoording en ethische praktijken, met inbegrip van eventuele grijze gebieden?
- ✓ Is er, naast de interne initiatieven of kaders voor het toezicht op ethiek en verantwoording, ook sprake van een vorm van externe begeleiding of zijn er ook controleprocessen ingesteld?
- ✓ Zijn er processen ingesteld waarmee derden (bijv. leveranciers, consumenten,

distributeurs/verkopers) of werknemers potentiële kwetsbaarheden, risico's of vertekening in het KI-systeem/de KI-toepassing kunnen melden?

Documentatie van afwegingen:

- ✓ Hebt u een mechanisme ingesteld om relevante belangen en waarden die door het KI-systeem in het gedrang komen, alsook potentiële afwegingen daartussen, vast te stellen?
- ✓ Welk proces gebruikt u voor het maken van dergelijke afwegingen? Hebt u gezorgd dat de beslissing bij de afweging is gedocumenteerd?

Vermogen om beroep in te stellen:

- ✓ Hebt u een geschikte reeks mechanismen ingesteld om het mogelijk te maken om in geval van schade of negatieve gevolgen beroep in te stellen?
- ✓ Hebt u mechanismen ingesteld om (eind)gebruikers/derden te voorzien van informatie over de mogelijkheden om beroep in te stellen?

Wij nodigen alle belanghebbenden uit deze controlelijst in de praktijk te testen en feedback te geven op de toepasbaarheid, de volledigheid en de relevantie ervan voor de specifieke KI-toepassing of het specifieke KI-domein, alsook op de overlap of complementariteit met bestaande nalevings- of controleprocessen. Op basis van deze feedback wordt begin 2020 een herziene versie van de controlelijst voor betrouwbare KI aan de Commissie voorgelegd.

Essentiële richtsnoeren uit hoofdstuk III:

- ✓ Stel een **controlelijst** voor betrouwbare KI vast bij de ontwikkeling, de installatie of het gebruik van KI en stem deze af op de specifieke situatie waarin het systeem wordt gebruikt.
- ✓ Bedenk dat een dergelijke controlelijst **nooit uitputtend** zal zijn. Bij betrouwbare KI gaat het niet om het afvinken van elementen op een lijst, maar om het voortdurend vaststellen van vereisten, afwegen van oplossingen en waarborgen van betere resultaten gedurende de hele levenscyclus van het KI-systeem, en om het betrekken van de belanghebbenden bij dit proces.

C. VOORBEELDEN VAN DE MOGELIJKHEDEN EN PUNTEN VAN ZORG DIE KI MET ZICH MEEBRENGT

(121) In het volgende gedeelte geven wij voorbeelden van vormen van ontwikkeling en gebruik van KI die moeten worden aangemoedigd. Ook geven we voorbeelden van gevallen waarin de ontwikkeling, de installatie of het gebruik tegen onze waarden in kan gaan en specifieke punten van zorg met zich mee kan brengen. Er moet een evenwicht worden gevonden tussen wat er met KI moet en wat er kán worden gedaan, en er moet voldoende aandacht worden besteed aan wat er níet met KI moet worden gedaan.

1. Voorbeelden van de mogelijkheden van betrouwbare KI

(122) Betrouwbare KI kan een uitgelezen mogelijkheid bieden om de beperking te ondersteunen van urgente uitdagingen waarmee de samenleving te maken heeft, zoals een vergrijzende bevolking, toenemende sociale ongelijkheid en milieuvervuiling. Deze potentie is ook wereldwijd zichtbaar, bijvoorbeeld in de

duurzameontwikkelingsdoelstellingen van de VN[57]. In het volgende gedeelte gaan we in op de manier waarop een Europese KI-strategie waarmee enkele van deze uitdagingen kunnen worden aangegaan, kan worden gestimuleerd.

a. Klimaatactie en duurzame infrastructuur

(123) Het aanpakken van de klimaatverandering moet voor beleidsmakers over de hele wereld een topprioriteit zijn. Digitale transformatie en betrouwbare KI hebben een enorme potentie wat betreft het verminderen van het effect van mensen op het milieu en maken efficiënt en effectief gebruik van energie en natuurlijke hulpbronnen mogelijk[58]. Betrouwbare KI kan bijvoorbeeld worden gekoppeld aan big data om de energiebehoeften nauwkeuriger vast te stellen, zodat de energie-infrastructuur en het energieverbruik efficiënter kunnen worden[59].

(124) In sectoren als het openbaar vervoer kunnen KI-systemen voor intelligente vervoerssystemen[60] worden gebruikt om rijen te minimaliseren, routes te optimaliseren, visueel gehandicapten de mogelijkheid te geven zelfstandiger te zijn,[61] bij te dragen aan het koolstofvrij maken van de sector door energie-efficiënte motoren te optimaliseren, en de ecologische voetafdruk te verkleinen, voor een groenere samenleving. Momenteel komt er wereldwijd elke 23 seconden iemand om het leven bij een auto-ongeluk[62]. Met KI-systemen kan het aantal sterfgevallen drastisch worden verlaagd, bijvoorbeeld door betere reactietijden en betere naleving van de regels[63].

b. Gezondheid en welzijn

(125) Betrouwbare KI-technologieën kunnen worden (en worden al) gebruikt om behandelingen slimmer en gerichter te maken en om bij te dragen aan de preventie van levensbedreigende ziekten[64]. Artsen en medische beroepsbeoefenaars kunnen in potentie, zelfs voordat mensen ziek worden, een nauwkeurigere en gedetailleerdere analyse uitvoeren van de complexe gezondheidsgegevens van een patiënt en preventieve behandeling op maat verzorgen[65]. In het kader van de vergrijzende Europese bevolking kunnen KI en robotica

---

57 https://sustainabledevelopment.un.org/?menu=1300

58 Een aantal EU-projecten is gericht op de ontwikkeling van slimme netwerken en energie-opslag, die de potentie hebben om bij te dragen aan een geslaagde digitaal ondersteunde energietransitie, onder meer via op KI gebaseerde en andere oplossingen. Als aanvulling op het werk binnen deze individuele projecten heeft de Commissie het initiatief Bridge opgezet, waardoor bij de lopende Horizon 2020-projecten op het gebied van slimme netwerken en energie-opslag een gemeenschappelijke visie kan worden ontwikkeld op horizontale kwesties: https://www.h2020-bridge.eu/.

59 Zie bijvoorbeeld het project Encompass: http://www.encompass-project.eu/.

60 Nieuwe oplossingen op basis van KI helpen steden bij de voorbereiding op de toekomst van mobiliteit. Zie bijvoorbeeld het door de EU gefinancierde project Fabulos: https://fabulos.eu/.

61 Zie bijvoorbeeld het project PRO4VIP, dat onderdeel is van de strategie Europese visie voor 2020 voor de bestrijding van vermijdbare blindheid, met name als gevolg van ouderdom. Mobiliteit en oriëntatie was een van de prioriteitsgebieden van het project.

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 Het Europese project UB-Drive is bijvoorbeeld gericht op het aanpakken van de beschreven uitdagingen op het gebied van vervoer door bij te dragen aan het faciliteren van de geleidelijke automatisering van en samenwerking tussen voertuigen, zodat een veiliger, inclusiever en betaalbaarder vervoerssysteem mogelijk wordt: https://up-drive.eu/.

64 Zie bijvoorbeeld het project Revolver (Repeated Evolution of Cancer – Herhaalde ontwikkeling van kanker): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, of het project Murab, waarbinnen nauwkeurigere biopsieën worden uitgevoerd en dat is gericht op snellere diagnosticering van kanker en andere ziekten: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 Zie bijvoorbeeld het project Live Incite: www.karolinska.se/en/live-incite. Dit consortium van inkopers uit de gezondheidszorg daagt de sector uit om slimme KI- en andere oplossingen te ontwikkelen waarmee leefstijlinterventies in het perioperatieve proces mogelijk worden gemaakt. Het doel is de ontwikkeling van nieuwe innovatieve oplossingen op het gebied van e-gezondheid die patiënten op een gepersonaliseerde manier kunnen stimuleren om zowel voor als na een operatie de benodigde actie te ondernemen in hun leefstijl en zo het resultaat van de gezondheidszorg te optimaliseren.

waardevolle hulpmiddelen zijn ter ondersteuning van zorgverleners en de ouderenzorg.[66] De gesteldheid van patiënten kan er in realtime mee in de gaten worden gehouden, waardoor er levens kunnen worden gered[67].

(126) Betrouwbare KI kan ook op een bredere schaal ondersteuning bieden. Er kunnen bijvoorbeeld algemene tendensen in de gezondheidszorg- en behandelsector mee worden onderzocht en vastgesteld,[68] met eerdere vaststelling van ziekten, efficiëntere ontwikkeling van geneesmiddelen, gerichtere behandelingen[69] en uiteindelijk meer geredde levens tot gevolg.

c. Goed onderwijs en digitale transformatie

(127) Vanwege de nieuwe technologische, economische en ecologische veranderingen moet de samenleving proactiever worden. Overheden, marktleiders, onderwijsinstellingen en vakverenigingen hebben de verantwoordelijkheid om de burgers het nieuwe, digitale tijdperk in te loodsen door te zorgen dat zij over de vaardigheden beschikken die nodig zijn voor de banen van de toekomst. Betrouwbare KI-technologieën kunnen helpen om nauwkeuriger te voorspellen welke banen en beroepen door de technologie zullen worden verstoord, welke nieuwe functies er zullen ontstaan en welke vaardigheden er nodig zullen zijn. Op die manier kunnen overheden, vakverenigingen en sectoren worden geholpen bij het plannen van de (om)scholing van werknemers. Ook kunnen burgers die bang zijn overbodig te zullen worden, zo een manier krijgen om zich te ontwikkelen, zodat ze een nieuwe rol kunnen vervullen.

(128) Verder kan KI een geweldig hulpmiddel zijn voor het tegengaan van ongelijkheden in het onderwijs en het creëren van gepersonaliseerde en aanpasbare onderwijsprogramma's waarmee iedereen nieuwe kwalificaties, vaardigheden en competenties kan verwerven in overeenstemming met zijn of haar eigen leervermogen[70]. Met behulp van KI zou zowel de leersnelheid als de kwaliteit van het onderwijs kunnen worden verhoogd – van de basisschool tot het hoger onderwijs.

2. Voorbeelden van punten van zorg die KI met zich meebrengt

(129) Een punt van zorg in verband met KI ontstaat wanneer een van de componenten van betrouwbare KI wordt geschonden. Veel van de onderstaande zorgen vallen al binnen het toepassingsgebied van bestaande

---

66 Binnen het door de EU gefinancierde project Caresses houdt men zich bezig met robots voor de ouderenzorg, waarbij het accent ligt op de culturele gevoeligheid ervan: ze passen hun gedrag en hun manier van spreken aan aan de cultuur en de gewoonten van de oudere die ze bijstaan: http://caressesrobot.org/en/project/. Zie ook de KI-toepassing genaamd Alfred, een virtuele assistent die ouderen helpt om actief te blijven: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Daarnaast wordt er binnen het project Empattics (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems – Betere informatie voor patiënten en verbetering van de communicatiesystemen) onderzocht en beschreven hoe beroepsbeoefenaars in de gezondheidszorg en patiënten ICT-technologieën, met inbegrip van KI-systemen, gebruiken om samen interventies te plannen en om de voortgang van hun fysieke en geestelijke gesteldheid van patiënten te monitoren: www.empattics.eu.

67 Zie bijvoorbeeld de MyHealth Avatar (www.myhealthavatar.eu), die een digitale weergave van de gezondheidsstatus van een patiënt biedt. In het kader van dit onderzoeksproject zijn een app en een onlineplatform opgezet om digitale informatie over uw gezondheidsstatus op de lange termijn te verzamelen en voor u inzichtelijk te maken. Dit gebeurt in de vorm van een levenslange gezondheidsmetgezel ("avatar"). De MyHealth Avatar voorspelt ook uw risico op een beroerte, diabetes, hart- en vaatziekten en een te hoge bloeddruk.

68 Zie bijvoorbeeld het project Enrichme (www.enrichme.eu), waarbinnen de geleidelijke daling van de cognitieve capaciteit binnen de vergrijzende bevolking wordt aangepakt. Met behulp van een geïntegreerd platform voor omgevingsondersteund wonen en een mobiele dienstrobot voor monitoring en interactie op de lange termijn kunnen ouderen langer zelfstandig en actief blijven.

69 Zie bijvoorbeeld het gebruik van KI door Sophia Genetics, dat gebruikmaakt van statistische inferentie, patroonherkenning en automatisch leren om de waarde van genomica-en radiomicagegevens te maximaliseren: https://www.sophiagenetics.com/home.html.

70 Zie bijvoorbeeld het project MaTHiSiS, dat is gericht op het bieden van een oplossing voor leren op basis van affect in een comfortabele leeromgeving bestaande uit hoogwaardige technologische apparatuur en algoritmen: (http://mathisis-project.eu/). Zie ook de Watson Classroom van IBM en het platform van Century Tech.

wettelijke voorschriften die verplicht zijn en dus moeten worden nageleefd. Zelfs in situaties waar is aangetoond dat de wettelijke voorschriften zijn nageleefd, is het echter mogelijk dat hiermee niet alle ethische punten van zorg die kunnen ontstaan, zijn verholpen. Aangezien ons begrip van de geschiktheid van regels en ethische beginselen altijd een ontwikkeling doormaakt en in de loop van de tijd kan veranderen, is het mogelijk dat de onderstaande, niet-uitputtende lijst van punten van zorg in de toekomst wordt ingekort, uitgebreid, aangepast of bijgewerkt.

a. Personen herkennen en volgen met KI

(130) Met KI is almaar efficiëntere herkenning mogelijk van individuele personen door zowel publieke als particuliere entiteiten. Interessante voorbeelden van schaalbare KI-herkenningstechnologie zijn gezichtsherkenning en andere onvrijwillige herkenningsmethoden met behulp van biometrische gegevens (zoals leugendetectie, persoonlijkheidsbeoordeling via micro-expressies en automatische stemherkenning). Soms is herkenning van personen het gewenste resultaat en in lijn met ethische beginselen (bijvoorbeeld bij het opsporen van fraude, witwassen of terrorismefinanciering). Automatische herkenning brengt echter grote juridische en ethische zorgen met zich mee, aangezien deze op veel psychologische en sociaal-culturele niveaus onverwachte gevolgen kan hebben. Een evenredig gebruik van controletechnieken voor KI is nodig om de autonomie van Europese burgers in stand te houden. Voor de verwezenlijking van betrouwbare KI is het essentieel om helder af te bakenen of, wanneer en hoe KI mag worden gebruikt voor de automatische herkenning van personen en om verschil te maken tussen de herkenning van een persoon en het traceren en volgen van een persoon, alsook tussen gericht toezicht en grootschalig toezicht. De toepassing van dergelijke technologieën moet duidelijk gerechtvaardigd zijn op grond van bestaande wetgeving[71]. Indien de wettelijke basis voor een dergelijke activiteit "toestemming" is, moeten er praktische middelen[72] worden ontwikkeld waardoor de te geven zinvolle en gecontroleerde toestemming automatisch kan worden herkend door KI- of vergelijkbare technologieën. Hetzelfde geldt voor het gebruik van "anonieme" persoonsgegevens waarvan de anonimisering kan worden teruggedraaid.

b. Verborgen KI-systemen

(131) Mensen moeten altijd weten of ze rechtstreeks contact hebben met een ander mens of met een machine. Het is de verantwoordelijkheid van beroepsbeoefenaars op het gebied van KI om dit op betrouwbare wijze te verwezenlijken. Zij moeten er daarom voor zorgen dat mensen bewust worden gemaakt van het feit dat – of kunnen navragen en controleren of – zij met een KI-systeem te maken hebben (bijvoorbeeld door middel van heldere en transparante disclaimers). Opgemerkt zij dat er grensgevallen bestaan die de zaak extra ingewikkeld maken (zoals een door een mens ingesproken stem die door een KI-systeem wordt gefilterd). Bedacht moet worden dat de verwarring tussen mensen en machines verschillende gevolgen kan hebben, zoals hechting, beïnvloeding of vermindering van de waarde van het mens-zijn.[73] De ontwikkeling van op mensen lijkende robots[74] moet daarom worden onderworpen aan zorgvuldige ethische controle.

c. Schending van grondrechten door beoordeling van burgers met behulp van KI

(132) Samenlevingen moeten ernaar streven de vrijheid en autonomie van alle burgers te beschermen. Iedere vorm van beoordeling van burgers kan leiden tot verlies van deze autonomie en kan het beginsel van non-discriminatie in gevaar brengen. Beoordeling mag alleen worden gebruikt als deze duidelijk gerechtvaardigd is en als de maatregelen evenredig en rechtvaardig zijn. Normatieve beoordeling van burgers (algemene

71 Met betrekking hiertoe kan worden verwezen naar artikel 6 van de AVG, waarin onder andere is bepaald dat de verwerking alleen rechtmatig is indien deze een geldige rechtsgrond heeft.

72 Zoals blijkt uit de huidige mechanismen voor het geven van geïnformeerde toestemming op het internet, geven consumenten doorgaans toestemming zonder hier goed over na te denken. Deze mechanismen kunnen dus nauwelijks als praktisch worden beschouwd.

73 Madary & Metzinger (2016). Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. Frontiers in Robotics and AI, 3(3).

74 Dit geldt ook door KI-gestuurde avatars.

beoordeling van "morele persoonlijkheid" of "ethische integriteit") op *alle* aspecten en op grote schaal door overheidsinstanties of particuliere actoren vormt een bedreiging voor deze waarden, met name wanneer het gebruik ervan niet in overeenstemming is met de grondrechten, niet evenredig is of geen afgebakend en kenbaar gemaakt doel heeft.

(133) Beoordeling van burgers – op grote of kleinere schaal – wordt tegenwoordig al vaak gebruikt voor louter beschrijvende en domeinspecifieke beoordelingen (zoals schoolsystemen, e-learning en rijbewijzen). Zelfs bij deze smallere toepassingen moet er een volledig transparante procedure voor burgers beschikbaar worden gemaakt, met inbegrip van informatie over het proces, het doel en de methodologie van de beoordeling. Opgemerkt zij dat transparantie niet voldoende is om discriminatie te voorkomen, rechtvaardigheid te garanderen en alle problemen rondom beoordeling te verhelpen. Idealiter moet de mogelijkheid worden geboden om, indien mogelijk, zonder negatieve gevolgen te besluiten niet deel te nemen aan het beoordelingsmechanisme. Anders moeten er mechanismen worden ingesteld om de beoordelingen aan te vechten en te corrigeren. Dit is met name van belang in situaties waar de macht ongelijk tussen de partijen is verdeeld. In gevallen waarin dit nodig is om naleving van de grondrechten te verwezenlijken, moeten dergelijke onttrekkingsmogelijkheden in het ontwerp van de technologie worden gewaarborgd, wat noodzakelijk is in een democratische samenleving.

d. Dodelijke autonome wapensystemen

(134) Momenteel houdt een onbekend aantal landen en sectoren zich bezig met onderzoek naar en de ontwikkeling van dodelijke autonome wapensystemen, variërend van raketten met de mogelijkheid tot selectieve aanvallen tot lerende machines met de cognitieve vaardigheden om zonder menselijke interventie te besluiten wanneer, waar en met wie ze vechten. Dit brengt ethische punten van zorg met zich mee, zoals het feit dat er een oncontroleerbare wapenwedloop van historisch ongekende omvang zou kunnen ontstaan en militaire situaties waarbij er vrijwel geen sprake meer is van menselijke controle en waarbij de risico's van storingen niet worden aangepakt. Het Europees Parlement heeft opgeroepen tot de urgente ontwikkeling van een gemeenschappelijk, juridisch bindend standpunt waarin wordt ingegaan op ethische en juridische vragen op het gebied van menselijke controle, toezicht, verantwoording en de uitvoering van internationale mensenrechtenwetgeving, internationale humanitaire wetgeving en militaire strategieën.[75] Wijzend op het doel van de Europese Unie om de vrede te bevorderen, zoals vastgelegd in artikel 3 van het Verdrag betreffende de Europese Unie, staan wij achter de resolutie van het Parlement van 12 september 2018 en alle verwante inspanningen op het gebied van dodelijke autonome wapensystemen, en beogen wij deze te ondersteunen.

e. Potentiële punten van zorg op de langere termijn

(135) De ontwikkeling van KI is nog altijd domeinspecifiek en vereist dat goed opgeleide menselijke wetenschappers en technici de doelen ervan exact specificeren. Als we naar de toekomst kijken met een langere periode in gedachten, kunnen we echter speculeren over bepaalde kritieke punten van zorg op de langere termijn[76]. Uit een op risico gebaseerde benadering is op te maken dat we rekening moeten blijven houden met deze zorgen met het oog op mogelijke onbekende factoren en "zwarte zwanen".[77] Omdat dat deze punten van zorg grote gevolgen kunnen hebben en gezien de huidige onzekerheid rondom de bijbehorende ontwikkelingen, moeten er regelmatig controles worden uitgevoerd op deze punten.

---

75 Resolutie 2018/2752(RSP) van het Europees Parlement.

76 Sommigen zijn van mening zijn dat volledige kunstmatige intelligentie, kunstmatig bewustzijn, kunstmatige morele actoren, een superintelligentie of transformatieve KI (die momenteel niet bestaan) voorbeelden van dergelijke langetermijnzorgen kunnen zijn, maar deze worden door veel anderen als onrealistisch beschouwd.

77 Een zwarte zwaan is een zeer zeldzame gebeurtenis met grote gevolgen – zo zeldzaam dat deze mogelijk nog nooit is voorgekomen. Daarom kan gewoonlijk slechts met zeer weinig zekerheid worden vastgesteld hoe groot de kans erop is.

D. CONCLUSIE

(136) Dit document bevat de ethische richtsnoeren voor KI die zijn ontwikkeld door de deskundigengroep op hoog niveau inzake kunstmatige intelligentie (AI HLEG).

(137) Wij erkennen de positieve gevolgen die KI-systemen al teweeg hebben gebracht en zullen blijven brengen, zowel commercieel als maatschappelijk. Het gaat ons er echter evenzeer om dat de risico's en andere negatieve gevolgen die aan deze technologieën zijn verbonden, op correcte en evenredige wijze worden aangepakt in het kader van de KI-toepassing. KI is een technologie die zowel transformatief als verstorend is, en de ontwikkeling ervan is de afgelopen jaren mogelijk gemaakt door de beschikbaarheid van enorme hoeveelheden digitale gegevens, grote technologische vooruitgang op het gebied van rekenkracht en opslagcapaciteit en aanzienlijke wetenschappelijke en technische innovatie op het gebied van KI-methoden en -hulpmiddelen. KI-systemen zullen de samenleving en burgers blijven beïnvloeden op manieren die we ons nog niet kunnen voorstellen.

(138) In het kader daarvan is het belangrijk om KI-systemen te creëren die te vertrouwen zijn, want mensen zullen alleen volledig en vol overtuiging de vruchten ervan kunnen plukken als de technologie, met inbegrip van de achterliggende processen en mensen, betrouwbaar is. Bij het opstellen van deze richtsnoeren is betrouwbare KI daarom onze kernambitie geweest.

(139) Betrouwbare KI is opgebouwd uit drie componenten: de KI moet 1) wettig zijn, door eerbiediging van alle toepasselijke wet- en regelgeving te waarborgen, 2) ethisch zijn, door naleving van ethische beginselen en waarden te waarborgen, en 3) robuust zijn uit zowel technisch als sociaal oogpunt om te zorgen dat KI-systemen niet ongewild schade aanrichten, zelfs al zijn de bedoelingen goed. Elke component is nodig, maar op zichzelf niet voldoende om betrouwbare KI te bewerkstelligen. In het ideale scenario sluiten alle drie de componenten op elkaar aan en valt de werking ervan gedeeltelijk samen. Waar spanningen ontstaan, moeten we proberen de componenten op één lijn te brengen.

(140) In hoofdstuk I hebben we de grondrechten uiteengezet, evenals een bijbehorende reeks ethische beginselen die cruciaal zijn in een KI-context. In hoofdstuk II hebben we zeven kernvereisten beschreven waaraan KI-systemen moeten voldoen om betrouwbare KI te verwezenlijken. We hebben technische en niet-technische methoden voorgesteld die van pas kunnen komen bij de uitvoering ervan. Tot slot hebben we in hoofdstuk III een controlelijst voor betrouwbare KI gegeven die kan helpen bij de operationalisering van de zeven vereisten. In een laatste gedeelte hebben we voorbeelden gegeven van voordelige kansen en punten van zorg in verband met KI-systemen waarover we verdere discussie hopen te stimuleren.

(141) Europa heeft een uniek perspectief vanwege zijn focus op het centraal stellen van de burger bij al zijn inspanningen. Deze focus is verweven in het DNA van de Europese Unie middels de verdragen waarop deze is gebouwd. Dit document maakt onderdeel uit van een visie waarbij betrouwbare KI wordt bevorderd. Wij zijn ervan overtuigd dat betrouwbare KI de grondslag moet zijn op basis waarvan Europa een leidende positie kan verwerven op het gebied van innovatieve, geavanceerde KI-systemen. Deze ambitieuze visie zal ertoe bijdragen dat Europese burgers tot bloei komen, zowel individueel als collectief. Ons doel is het creëren van een cultuur van "betrouwbare KI voor Europa", zodat iedereen op dusdanige wijze kan profiteren van de voordelen van KI dat eerbiediging wordt gewaarborgd van onze basiswaarden: grondrechten, democratie en de rechtsstaat.

WOORDENLIJST

(142) Deze woordenlijst heeft betrekking op de richtsnoeren en is bedoeld als hulpmiddel om de in dit document gebruikte termen te begrijpen.

Systemen op basis van kunstmatige intelligentie of KI-systemen

(143) Systemen op basis van kunstmatige intelligentie (KI) zijn door mensen ontworpen[78] softwaresystemen (en mogelijk ook hardwaresystemen) die, met een complex doel, in de fysieke of digitale dimensie werken door via gegevensverzameling hun omgeving waar te nemen, de verzamelde gestructureerde of ongestructureerde gegevens te interpreteren, te redeneren op basis van de uit deze gegevens verkregen kennis of de verkregen informatie te verwerken en te beslissen met welke handeling(en) het gestelde doel het best kan worden bereikt. KI-systemen kunnen gebruikmaken van symbolische regels of een numeriek model leren en kunnen hun gedrag ook aanpassen door te analyseren welke invloed hun eerdere handelingen op de omgeving hebben.

(144) Als wetenschappelijke discipline omvat KI verschillende benaderingen en technieken, zoals automatisch leren (waarvan deep learning en reinforcement-leren specifieke voorbeelden zijn), automatisch redeneren (waaronder plannen, inroosteren, kennisrepresentatie en redeneren, zoeken en optimaliseren) en robotica (waaronder controle, waarneming, sensoren en actuatoren, alsook de integratie van alle andere technieken in cyber-fysieke systemen).

(145) Tegelijk met dit document wordt ook een afzonderlijk door de AI HLEG opgesteld document gepubliceerd, waarin de in dit document gehanteerde definitie van "KI-systemen" verder wordt uitgewerkt. Dit document is getiteld: "Een definitie van KI: de belangrijkste capaciteiten en wetenschappelijke disciplines".

Beroepsbeoefenaars op het gebied van KI

(146) Met beroepsbeoefenaars op het gebied van KI bedoelen we alle personen of organisaties die KI-systemen ontwikkelen (met inbegrip van onderzoek naar en ontwerp van deze systemen of de aanlevering van gegevens ervoor), installeren (met inbegrip van de toepassing) of gebruiken, met uitzondering van de personen of organisaties die KI-systemen gebruiken als eindgebruiker of consument.

Levenscyclus van een KI-systeem

(147) De levenscyclus van een KI-systeem omvat de ontwikkelingsfase (met inbegrip van onderzoek, ontwerp, aanlevering van gegevens en beperkte testen), de installatiefase (met inbegrip van de toepassing) en de gebruiksfase.

Controleerbaarheid

(148) Controleerbaarheid verwijst naar het vermogen van een KI-systeem om een controle van de algoritmen, gegevens en ontwerpprocessen van het systeem te ondergaan en vormt een van de zeven vereisten waaraan een betrouwbaar KI-systeem moet voldoen. Dat betekent niet noodzakelijkerwijs dat informatie over bedrijfsmodellen en intellectuele eigendom in verband met het KI-systeem altijd openbaar beschikbaar moet zijn. Door de traceerbaarheid en registratiemechanismen vanaf de vroege ontwerpfase van het KI-systeem te waarborgen kan worden bijgedragen aan de controleerbaarheid van het systeem.

Vertekening

(149) Vertekening is een neiging om een persoon, object of standpunt te bevoordelen of te benadelen. Vertekening kan op veel manieren in KI-systemen voorkomen. Bij gegevensgestuurde KI-systemen, zoals de systemen die worden geproduceerd middels automatisch leren, kan vertekening in de gegevensverzameling en de training bijvoorbeeld leiden tot een KI-systeem dat vertekening vertoont. Bij KI-systemen op basis van logica, zoals op

[78] Mensen ontwerpen KI-systemen rechtstreeks, maar kunnen ook KI-technieken gebruiken om hun ontwerp te optimaliseren.

regels gebaseerde systemen, kan vertekening ontstaan door de manier waarop een kennistechnicus de regels ziet die in een bepaalde situatie gelden. Vertekening kan ook ontstaan door onlineleren en aanpassing middels interactie, of door personalisering, waarbij gebruikers aanbevelingen of informatie krijgen die op hun smaak zijn afgestemd. Vertekening heeft niet noodzakelijkerwijs te maken met menselijke vooringenomenheid of met door de mens gestuurde gegevensverzameling, maar kan bijvoorbeeld ontstaan doordat een systeem slechts in een beperkt aantal situaties wordt gebruikt, waardoor het niet naar andere situaties kan worden uitgebreid. Vertekening kan goed of slecht en bedoeld of onbedoeld zijn. In bepaalde gevallen kan vertekening leiden tot discriminerende en/of onrechtvaardige resultaten. In dit document wordt dat onrechtvaardige vertekening genoemd.

Ethiek

(150) Ethiek is een academische discipline die een tak vormt van de filosofie. In algemene termen wordt binnen de ethiek ingegaan op vragen als "wat is een goede handeling?", "wat is de waarde van een mensenleven?", "wat is rechtvaardigheid?" en "wat is het goede leven?". De academische ethiek kent vier grote onderzoeksgebieden: i) meta-ethiek, waarbij het vooral gaat om de betekenis en duiding van normatieve zinnen en de vraag hoe de waarheidswaarde ervan kan worden bepaald (als die er is); ii) normatieve ethiek, het praktische middel om een morele handelwijze te bepalen door de normen voor goede en slechte handelingen te onderzoeken en een waarde aan specifieke handelingen toe te kennen; iii) descriptieve ethiek, gericht op een empirisch onderzoek van het morele gedrag en de morele overtuigingen van mensen; en iv) toegepaste ethiek, waarbij het gaat om wat we verplicht moeten (of wat we mogen) doen in een specifieke (vaak historisch nieuwe) situatie of een bepaald domein van (vaak historisch ongekende) mogelijke handelingen. Binnen de toegepaste ethiek wordt ingegaan op situaties uit het echte leven waarin onder tijdsdruk en vaak met beperkte rationaliteit beslissingen moeten worden genomen. KI-ethiek wordt over het algemeen beschouwd als een voorbeeld van toegepaste ethiek en is gericht op de normatieve problemen die voortkomen uit het ontwerp, de ontwikkeling, de toepassing en het gebruik van KI.

(151) Binnen ethische discussies worden vaak de termen "moreel" en "ethisch" gebruikt. De term "moreel" verwijst naar de concrete, feitelijke gedragspatronen, de gewoonten en de conventies die specifieke culturen, groepen of personen op een bepaald moment vertonen. De term "ethisch" verwijst naar een evaluatieve beoordeling van dergelijke concrete handelingen en gedragingen vanuit een systematisch, academisch perspectief.

Ethische KI

(152) In dit document wordt met ethische KI verwezen naar dusdanige ontwikkeling en installatie en dusdanig gebruik van KI dat naleving van de ethische normen, met inbegrip van grondrechten als bijzondere morele rechten, ethische beginselen en verwante kernwaarden, wordt gewaarborgd. Dit is het tweede van de drie kernelementen die noodzakelijk zijn voor de verwezenlijking van betrouwbare KI.

KI waarbij de mens centraal staat

(153) Bij de benadering van KI waarbij de mens centraal staat, wordt ernaar gestreefd te zorgen dat menselijke waarden centraal staan bij de manier waarop KI-systemen worden ontwikkeld, geïnstalleerd, gebruikt en gemonitord. Dit gebeurt door eerbiediging van de grondrechten te waarborgen, met inbegrip van de rechten die worden beschreven in de verdragen van de Europese Unie en het Handvest van de grondrechten van de Europese Unie. Al deze rechten zijn verenigd op grond van een gemeenschappelijke grondslag die is geworteld in respect voor de menselijke waardigheid en waarin de mens een unieke en onvervreemdbare morele status heeft. Bij deze benadering hoort ook het rekening houden met de natuurlijke omgeving en andere levende wezens die deel uitmaken van het menselijke ecosysteem, alsook een duurzame benadering om toekomstige generaties de mogelijkheid te geven tot bloei te komen.

Red teaming

(154) Red teaming is de praktijk waarbij een "red team" of onafhankelijke groep een kwaadwillige rol of een kwaadwillig standpunt inneemt en zo een organisatie uitdaagt om haar effectiviteit te vergroten. Red teaming wordt voornamelijk gebruikt om potentiële kwetsbaarheden op het gebied van beveiliging vast te stellen en aan te pakken.

Reproduceerbaarheid

(155) Reproduceerbaarheid beschrijft of een KI-experiment hetzelfde gedrag vertoont wanneer het onder gelijke omstandigheden wordt herhaald.

Robuuste KI

(156) Onder de robuustheid van een KI-systeem valt zowel de technische robuustheid (gepastheid in een bepaalde context, zoals het toepassingsgebied of de fase van de levenscyclus) als de robuustheid uit sociaal oogpunt (zorgen dat het KI-systeem naar behoren rekening houdt met de context en omgeving waarin het werkzaam is). Deze robuustheid is cruciaal om te zorgen dat er geen ongewilde schade kan ontstaan, zelfs al zijn de bedoelingen goed. Robuustheid is de derde van de drie componenten die noodzakelijk zijn voor de verwezenlijking van betrouwbare KI.

Belanghebbenden

(157) Met belanghebbenden bedoelen we iedereen die onderzoek doet naar KI of die deze ontwikkelt, installeert of gebruikt, alsook degenen die (direct of indirect) de gevolgen van KI ondervinden – met inbegrip van onder meer bedrijven, organisaties, onderzoekers, overheidsdiensten, instellingen, maatschappelijke organisaties, overheden, regelgevers, sociale partners, personen, burgers, werknemers en consumenten.

Traceerbaarheid

(158) De traceerbaarheid van een KI-systeem verwijst naar het vermogen om de gegevens en de ontwikkelings- en installatieprocessen van het systeem bij te houden, gewoonlijk door middel van gedocumenteerde opgeslagen identificatie.

Vertrouwen

(159) Wij gebruiken de volgende definitie uit de literatuur: "Vertrouwen wordt gezien als: 1) een reeks specifieke overtuigingen in verband met welwillendheid, bekwaamheid, integriteit en voorspelbaarheid (vertrouwensovertuigingen); 2) de bereidheid van één partij om in een riskante situatie op een ander te vertrouwen (vertrouwensintentie); of 3) de combinatie van deze elementen."[79] Hoewel "vertrouwen" gewoonlijk geen eigenschap is die aan machines wordt toegeschreven, willen we in dit document benadrukken hoe belangrijk het is er niet alleen op te kunnen vertrouwen dat KI-systemen in lijn zijn met wettelijke voorschriften en ethische normen en robuust zijn, maar ook dat dit vertrouwen kan worden toegeschreven aan alle mensen en processen die bij de levenscyclus van het KI-systeem betrokken zijn.

Betrouwbare KI

(160) Betrouwbare KI is opgebouwd uit drie componenten: de KI moet 1) wettig zijn, door eerbiediging van alle toepasselijke wet- en regelgeving te waarborgen, 2) ethisch zijn, door respect voor ethische beginselen en waarden te tonen en naleving ervan te waarborgen, en 3) robuust zijn uit zowel technisch als sociaal oogpunt, aangezien KI-systemen ongewild schade kunnen aanrichten, zelfs al zijn de bedoelingen goed. Betrouwbare KI heeft niet alleen betrekking op de betrouwbaarheid van het KI-systeem zelf, maar omvat ook de betrouwbaarheid van alle processen en actoren die deel uitmaken van de levenscyclus van het systeem.

Kwetsbare personen en groepen

[79] Siau, K., Wang, W. (2018), Building Trust in Artificial Intelligence, Machine Learning, and Robotics, *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. blz. 47–53.

(161) Vanwege de heterogeniteit ervan bestaat er geen algemeen aanvaarde of breed gedragen definitie van kwetsbare personen. Wat een kwetsbare persoon of groep is, hangt van de context af. Tijdelijke levensgebeurtenissen (zoals jeugd of ziekte), marktfactoren (zoals ongelijkheid wat betreft beschikking over informatie of marktmacht), economische factoren (zoals armoede), factoren in verband met iemands identiteit (zoals geslacht, religie of cultuur) en andere factoren kunnen een rol spelen. In artikel 21 van het Handvest van de grondrechten van de EU over non-discriminatie worden de volgende gronden genoemd, die, naast andere gronden, een referentiepunt kunnen vormen: geslacht, ras, kleur, etnische of sociale afkomst, genetische kenmerken, taal, godsdienst of overtuigingen, politieke of andere denkbeelden, het behoren tot een nationale minderheid, vermogen, geboorte, een handicap, leeftijd en seksuele gerichtheid. In andere rechtsartikelen wordt ingegaan op de rechten van specifieke groepen, naast de bovenstaande rechten. Een dergelijke lijst is nooit uitputtend en kan in de loop van de tijd veranderen. Een kwetsbare groep is een groep personen die één of meerdere kenmerken van kwetsbaarheid gemeen hebben.

Dit document is opgesteld door de leden van de deskundigengroep op hoog niveau inzake KI,

die hieronder in alfabetische volgorde worden genoemd.

Pekka Ala-Pietilä, voorzitter van de AI HLEG
KI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – corapporteur
Zalando
Mária Bieliková
Slowaakse Technische Universiteit in Bratislava
Cecilia Bonefeld-Dahl – corapporteur
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE-initiatief Ethics of Intelligent and Autonomous Systems (Ethiek van intelligente en autonome systemen) & Sorbonne Université
Mark Coeckelbergh
Universiteit van Wenen
Virginia Dignum – corapporteur
Universiteit van Umeå
Luciano Floridi
Universiteit van Oxford
Jean-Francois Gagné – corapporteur
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Bureau voor de grondrechten
Sami Haddadin
Munich School of Robotics and MI
Gry Hasselbalch
De denkdoetank DataEthics & Universiteit van Kopenhagen
Fredrik Heintz
Universiteit van Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Universiteit van Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wenen
Robert Kroplewski
Advocaat & adviseur voor de Poolse overheid
Elisabeth Ling
RELX
Pierre Lucas
Orgalim – de technologische sectoren van Europa
Ieva Martinkenaite
Telenor
Thomas Metzinger – corapporteur
JGU Mainz & Europese Vereniging van Universiteiten
Catelijne Muller
ALLAI Nederland & EESC
Markus Noga
SAP
Barry O'Sullivan, vicevoorzitter van de AI HLEG
University College Cork
Ursula Pachl
BEUC
Nicolas Petit – corapporteur
Universiteit van Luik
Christoph Peylo
Bosch
Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
Staf Europacollege & CEPS
Francesca Rossi
IBM
Cristina San José
Europese Bankfederatie
George Sharkov
Digital SME Alliance
Philipp Slusallek
Duits onderzoekscentrum voor KI (DFKI)
Françoise Soulié Fogelman
KI-consultant
Saskia Steinacker – corapporteur
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – corapporteur
TU Delft
Thiébaut Weber
EVV
Cecile Wendling
AXA
Karen Yeung – corapporteur
Universiteit van Birmingham

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe en Karen Yeung hebben als rapporteurs gefungeerd voor dit document.

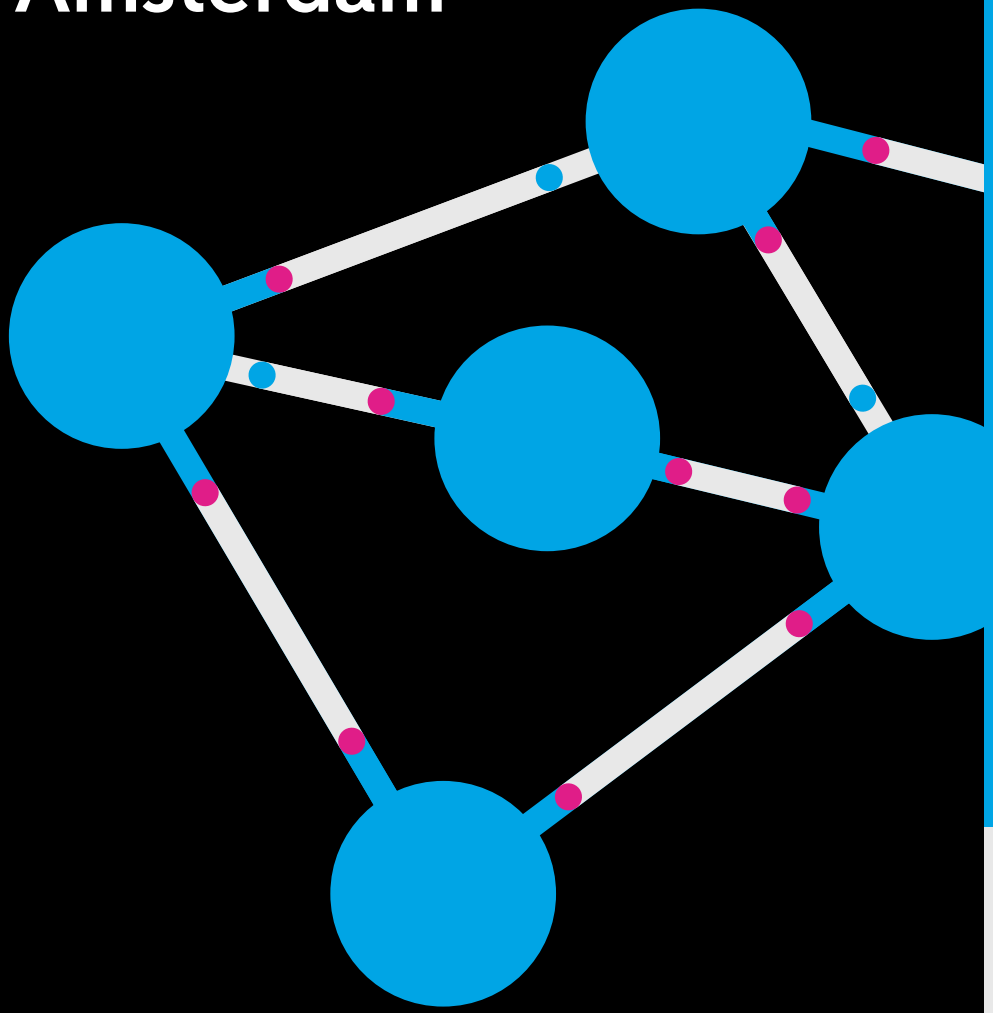
Pekka Ala-Pietilä is voorzitter van de AI HLEG. Barry O'Sullivan is vicevoorzitter en coördineert het tweede product van de AI HLEG.

Nozha Boujemaa, die tot 1 februari 2019 vicevoorzitter was en het eerste product coördineerde, heeft ook bijgedragen aan de inhoud van dit document.

Nathalie Smuha heeft redactionele ondersteuning geboden.

ISBN 978-92-76-11999-9 KK-02-19-841-NL-N

City of Amsterdam



Grip on algorithms

Approach and tools for a responsible use of algorithms in Amsterdam

Preface

Technological developments are changing our city and the ways in which we provide our services. We need to think about how the City of Amsterdam wants to approach technology. Technology can solve social issues and implement laws and regulations. Algorithms have the ability to develop themselves, which is why we need to maintain a grip on them. Everyone knows that the misuse of algorithms in the Dutch childcare benefits scandal led to a loss of confidence in the government.

In order to get a grip on algorithms, we initiated the Algorithm Lifecycle Approach programme. The Algorithm Lifecycle Approach, initiated by the Digital City Agenda and the Amsterdam Intelligence Agenda, offers transparency, direction and accountability for residents. It also ensures the democratic values of a constitutional state. We do this by checking existing algorithms and adjusting them where necessary.

The Algorithm Lifecycle Approach relates to the Amsterdam character traits: free-thinking, open-minded and inclusive. Based on this approach and experiments, we continuously try to ensure that everyone benefits from the opportunities of new technology in the city, while at the same time guaranteeing everyone's digital rights. This results in successes such as the world's first Algorithm Register. We achieved this by cooperating with Helsinki, but also through initiatives such as supplementary procurement conditions and the audit framework put in place for this purpose.

Transparency and accountability are the decisive factors for the work we do as Digital City. We need to take responsibility for there to be actual transparency. We are therefore proud of the governance framework developed as a basis for transparency on the algorithms we use. At the same time, practice also shows that we need to keep practising to strengthen that crucial supervision. The right to reply, accountability and actual transparency are only possible if we and all people of Amsterdam understand and comply with duties and responsibilities. This is a prerequisite for building residents' trust in our digital services and for us to provide the best possible service. Raising awareness about algorithms and sharing the lessons learned based on this approach are thus decisive steps for the Digital City.

This guideline provides transparency and helps in both taking and giving responsibility. It is the starting point for working together towards a responsible Digital City where every resident feels at home.

Document structure

This guideline contains the lessons learned and products from the City of Amsterdam's Algorithm Lifecycle Approach. We have worked on that approach under the Digital City Agenda and the Amsterdam Intelligence Agenda. The guideline is intended for administrators, municipal councillors and municipal officers, but may also benefit other governments and organisations that use algorithms for their services and that want to do so in a responsible and transparent manner.

First, we describe our approach. This is followed by lessons learned and products developed for the seven tools that manage and audit algorithms during their lifecycle and assess their risks. From research into and roadmaps for an Algorithm Register to contractual terms that can be downloaded for the procurement of algorithm systems from suppliers. We share all these as downloadable files so you can get started on them yourself.

For questions or more information, please contact Siham El Yassini via algoritmen@amsterdam.nl

Contents page

Introduction

Each and every day, the City of Amsterdam aims to keep the city liveable for all residents. Sometimes we use algorithms to help us with our service provision. Instead of a person, a computer then performs certain actions. Algorithms help us to carry out our tasks more quickly, in a more targeted manner or simply better. Examples:

- scan cars driving through the streets to check whether a parked car has a right to be parked.
- being able to timely intervene in crowded places.
- directing citizen reports to the appropriate departments.
- identifying debts at an early stage and alerting the relevant care provider.

We want to take the lead to encourage the use of algorithms locally, as cities can benefit immensely from the use of algorithms. However, the use of algorithms also requires cities to be transparent about how algorithms are used. This is how we can minimise algorithms' risks for residents to the maximum extent possible. It is important that algorithms are fair and explainable. The algorithms that the City of Amsterdam uses to prepare decisions must be explainable in a easy-to-understand language. However, algorithms are written in computer language, which means it is not always clear to everyone exactly what considerations apply.

The following examples show the importance of this:

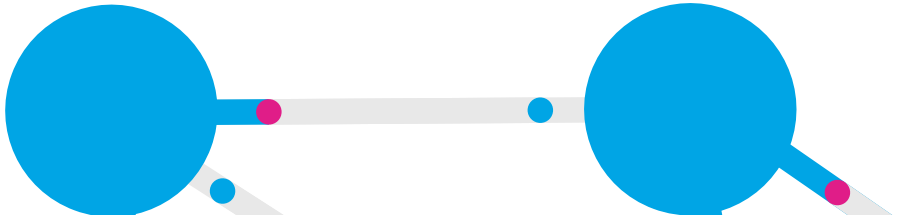
- the SyRI judgment of The Hague District Court
- the Dutch childcare benefits scandal (Toeslagenaffaire) (in Dutch)
- the latest report of the Netherlands Court of Audit and the Rotterdam Court of Audit (the latter is available in Dutch only)

The definition of an algorithm

There are currently many different definitions for an 'algorithm' or 'algorithmic application'. Efforts are under way both nationwide and within the EU to develop usable and clear definitions.

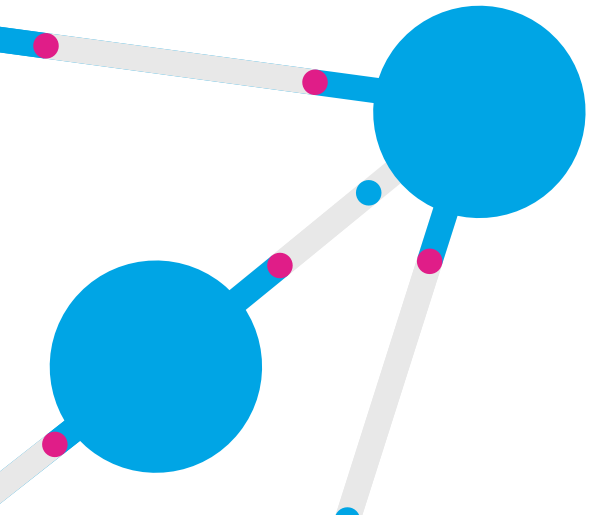
Until then, we apply the following definition: *software that, through the use of data analysis, statistics or self-learning logic, makes automated predictions or decisions or gives advice that lead to an impact on citizens or businesses*.

This definition, it is added, corresponds for the most part to the one mentioned by the European Commission in its proposal to regulate AI. The definition is broad and precise at the same time. This definition fits simple computational models as well as 'machine-learning models' and other forms of Artificial Intelligence (AI). We include such an algorithm in the Algorithm Register if we use it in an activity that affects citizens or businesses. The other tools then apply as well.



Approach and policy on algorithms

We initiated the Algorithm Lifecycle Approach to get a better grip on algorithms and to create transparency in the algorithms that we use. The approach forms part of the Digital City Agenda and the Amsterdam Intelligence Agenda. Within the Algorithm Lifecycle Approach, we are working on different tools to make the use of algorithms more fair and more transparent for Amsterdam people. These are tools that can manage algorithms during their lifecycle, assess their risks and audit them.



1 Algorithm Register

The City of Amsterdam's Algorithm Register was launched on 28 September 2020. The Algorithm Register lists:

- which algorithms we use.
- why we use them.
- how they work.

There is an administrative agreement for the Algorithm Register. The Algorithm Register creates transparency in the algorithms that we use and is a tool to for conversation with different stakeholders. The register is included in the coalition agreement as a tool. The Algorithm Register will change from a beta version to a full-fledged register in the course of 2023. We will continue to develop the register towards the nationwide Algorithm Register.

2 Contractual terms

We have developed contractual terms for the algorithms that we purchase from suppliers, which describe which information we require from them. This will allow us to share that information with residents. The contractual terms were established in November 2020. An AI Pact is currently being developed which may replace the Standard Clauses For Procurement Of Trustworthy Algorithmic Systems.

3 Objection procedure

Residents, businesses or institutions can object to a municipal decision, such as a fine or imposed taxes. We are currently examining whether residents receive clear and complete information when objecting to a decision in which an algorithm had been involved. The Municipal Management Team (GMT) adopted this on 20 January 2022.

4 Governance

On behalf of the City of Amsterdam, we have very specifically defined the following in governance:

- the measures needed to avoid risks when using algorithms.
- the information we need to keep.
- the responsible party in case an algorithm does something it is not supposed to do

The Municipal Management Team adopted this on 20 January 2022. It will be on the Municipal Executive's agenda in Q4 2022.

Approach and policy on algorithms

5 Audit

The service processes in which we use algorithms can be audited. An audit verifies that the appropriate measures are in place to prevent risks, such as those of discrimination or a violation of rights. The City of Amsterdam will make use of the Audit Framework of the Netherlands Court of Audit for the coming period and adjust the Amsterdam model. We decide at a later stage which of the two frameworks will eventually be taken into use.

6 Bias analysis

We have developed a standard for bias analysis to avoid bias when using algorithms. The Municipal Management Team adopted this bias analysis on 20 January 2022. We continue to further develop this standard, but we mandate it in the algorithm development.

7 Impact assessment

We have established a human rights impact assessment. We use the assessment to analyse the impact of the use of algorithms on human rights. It enables us to take appropriate measures. The Municipal Management Team adopted this bias analysis on 20 January 2022. In addition, we may use other impact assessments such as the Government's Human Rights and Algorithms Impact Assessment.

Additional tools: collaboration or intended collaboration with partners

We like to look beyond borders and develop these tools openly. We do this in cooperation with different government bodies and knowledge institutions such as TU Delft and the Amsterdam University of Applied Sciences. It saves other government bodies having to reinvent the wheel. We also ask citizens to participate in the development process.

Moreover, we share our best practices and insights on a European level. We are involved in the setup of four initiatives.

The Algorithm Lifecycle Approach not only deploys tools, it also addresses new issues such as:

- Information management and archiving.
- Standards.
- Further development of existing tools such as the Algorithm Register.

Tool 1: Algorithm Register

The Algorithm Register is an overview of the algorithms used by the City of Amsterdam in its municipal services. An algorithmic application is software that in an automated manner:

- makes predictions
- makes decisions
- gives advice

which impacts citizens and businesses, by means of data-analysis, statistics or self-learning logic. For each algorithm, you first find general information on the intent and operation of algorithms. What follows is more technically detailed information. Your response helps us to make the algorithms that we use better, fairer and more responsible.

Go to the Algorithm Register

The register is a beta version: it is still under development. We also want to move to a national register. This is why it is so important to make a national data standard. To then be able to merge different registers, for example, by having a standard for interoperability (to enable exchange) and representation (website display).

Siham El Yassini, Coordinator Algorithms Amsterdam:
"The most prominent manifestation of the Algorithm Lifecyle Approach is the Algorithm Register. The Algorithm Register explains for each algorithm what it does, how it does it and whether it does it in an unbiased way. Including an algorithm in the register is an intensive process. Our starting point is to screen the algorithms with key stakeholders. Citizens, entrepreneurs, knowledge institutions, technicians. In doing so, we want to provide a platform to jointly discuss the development of our algorithms."

Read more:

Study on citizens' information need on the use of algorithms by government bodies.

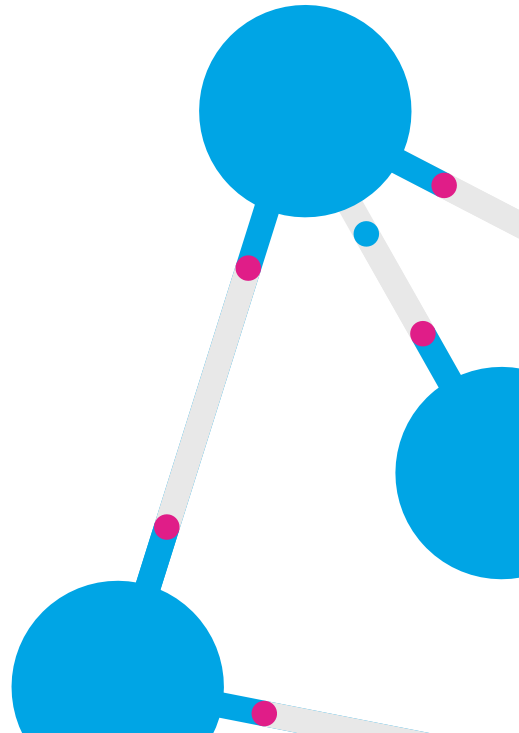
On behalf of the Consortium on Public Control of Algorithms, PON & Telos have conducted a study on the way in which citizens want to receive information about the use of algorithms by government bodies. And which information they wish to be able to see.

The study

Whitepaper on the Algorithm Register Amsterdam, Helsinki and Saidot (September 2020)

A collection of lessons learned on AI use for public services made visible and clear in an Algorithm Register. This whitepaper, written in collaboration with Saidot and Helsinki, contributed to international awareness of the Amsterdam Algorithm Register and many new international relationships with the Amsterdam Algorithm Life Cycle Team.

Go to the whitepaper



Tool 2: Contractual terms for algorithmic applications

Amsterdam has developed supplementary AI procurement conditions. These terms and conditions detail what information third parties must provide about the algorithm the city uses. The provider of algorithmic applications may be an external party or an in-house supplier. We use the information provided to us by means of these contractual terms for the procurement of algorithms to further ensure citizens' trust in our services. This is possible because we can provide information on how artificial intelligence has been used (technology, procedure and explainability) to provide a service in the city.

Firstly, the supplementary procurement conditions for AI build on the ethical guidelines developed by the European Commission's High Level Expert Group on Artificial Intelligence (AI). Secondly, they build on our digital ethics policies as one of the founders of the Cities Digital Rights Coalition. We made these supplementary AI procurement conditions available as a template for reuse by other government or non-government organisations. It serves as a practical tool to both promote the use of AI, and be transparent about our services at the same time.

We have rewritten the initial draft version of the supplementary AI procurement conditions as much as possible to ensure that their contents are as consistent as possible with the European AI Act still under negotiation. Finally, the Amsterdam AI procurement conditions contain elements that stem from Dutch jurisdiction and are not usable in other European countries. Also, the latest draft of the supplementary AI procurement conditions may be widely used by public organisations.

This version, which is in line with the draft version of the AI Act, is not an official EU document.

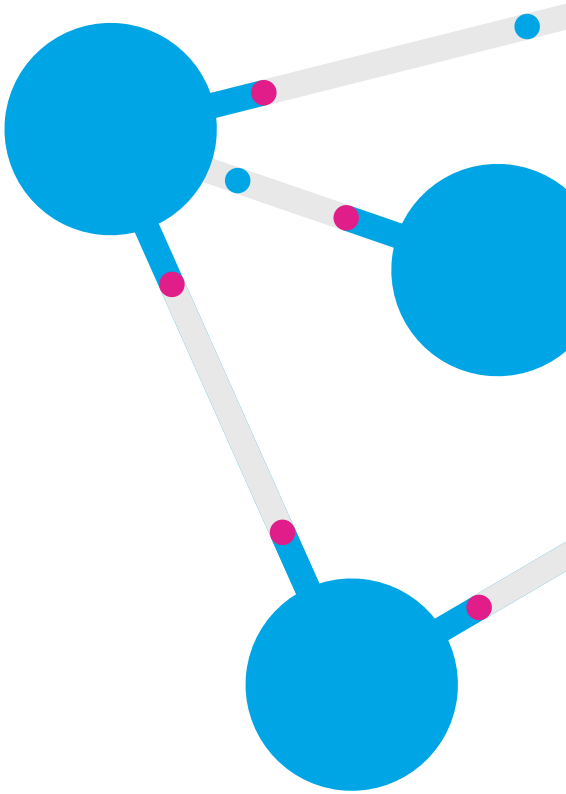
Partners: European Commission, DG GROW & DG CONNECT, Pels & Rijcken, Association of Netherlands Municipalities (VNG), Ministry of Foreign Affairs and Kingdom Relations and Consortium on Control on Public Algorithms.

Download the Standard Clauses For Procurement Of Trustworthy Algorithmic Systems plus explanatory notes
(available in both a Dutch and English version)

The draft EU version in line with the AI Act is available on this website

Go to the Digital Public Buyer's Community

From 2023, the European Commission's Digital Public Buyer's Community will provide an offline and online community of practice. This will provide additional support to this group in the form of workshops, webinars, information exchange and advice. Members of this group will be notified of the platform's launching date.



Je maakt afspraken over:

1. Transparante informatie

Zonder dat de leverancier informatie deelt die concurrentiegevoelig is.

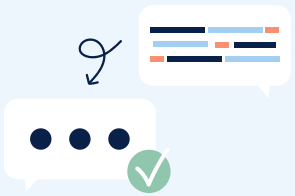
2. Kwaliteit van het algoritme

De leverancier zorgt voor betrouwbare algoritmes waarin de burger niet benadeeld wordt.

3. Datagebruik

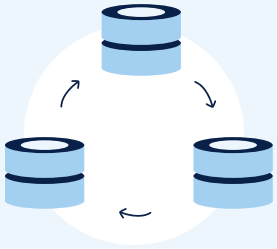
Houdt zelf de rechten op de gebruikte data.

Voordelen:



Leg gemakkelijk besluiten uit aan burgers.

Duidelijke afspraken, altijd op dezelfde manier.



De data wordt gedeeld, zodat we die kunnen hergebruiken.

Betere kwaliteitscontrole op algoritmes.

Siham El Yassini, Coordinator Algorithms Amsterdam :
“Together with Procurement, we have set up contractual terms for the procurement of algorithms. This is a tough process, which Amsterdam cannot really undertake on its own. You need negotiating power, especially in dealing with Big Tech, the big boys from Silicon Valley. Nationwide, or even at a European level.”

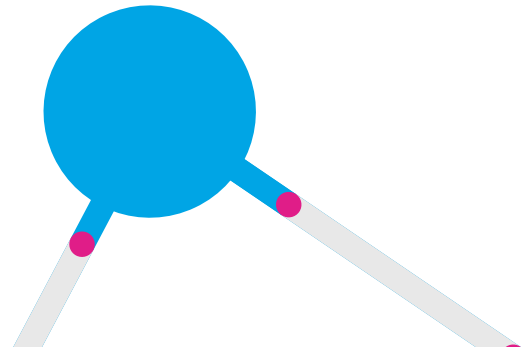
Tool 3: Objections procedure

Sometimes we use algorithms in drafting or making a decision. This could involve a direct use of algorithms. For example, when scanning vehicle registration numbers to check whether parking fees have been paid. It could also involve indirect use. For example, in a risk assessment that may or may not warrant further enquiries about someone. An algorithm is often technically complex. It is therefore not always clear to the party objecting the decision or the municipal officer handling the objection what impact the algorithm had on the decision. And whether this is important in the assessment of the objection. This may have implications for legal protection when using algorithms in an administrative law environment.

The guideline on the objections procedure has recommendations for officials handling an objection in which the algorithm formed the basis of the initial decision. The recommendations are to ensure an accessible, permanent and effective objection procedure. To make it clear for citizens:

- how they can object to a decision based on algorithms.
- where they can find information about algorithms.
- how a careful reconsideration is made.

The guideline on the objections procedure helps officials with handling an objection.



Tool 4: Governance for a responsible application of algorithms

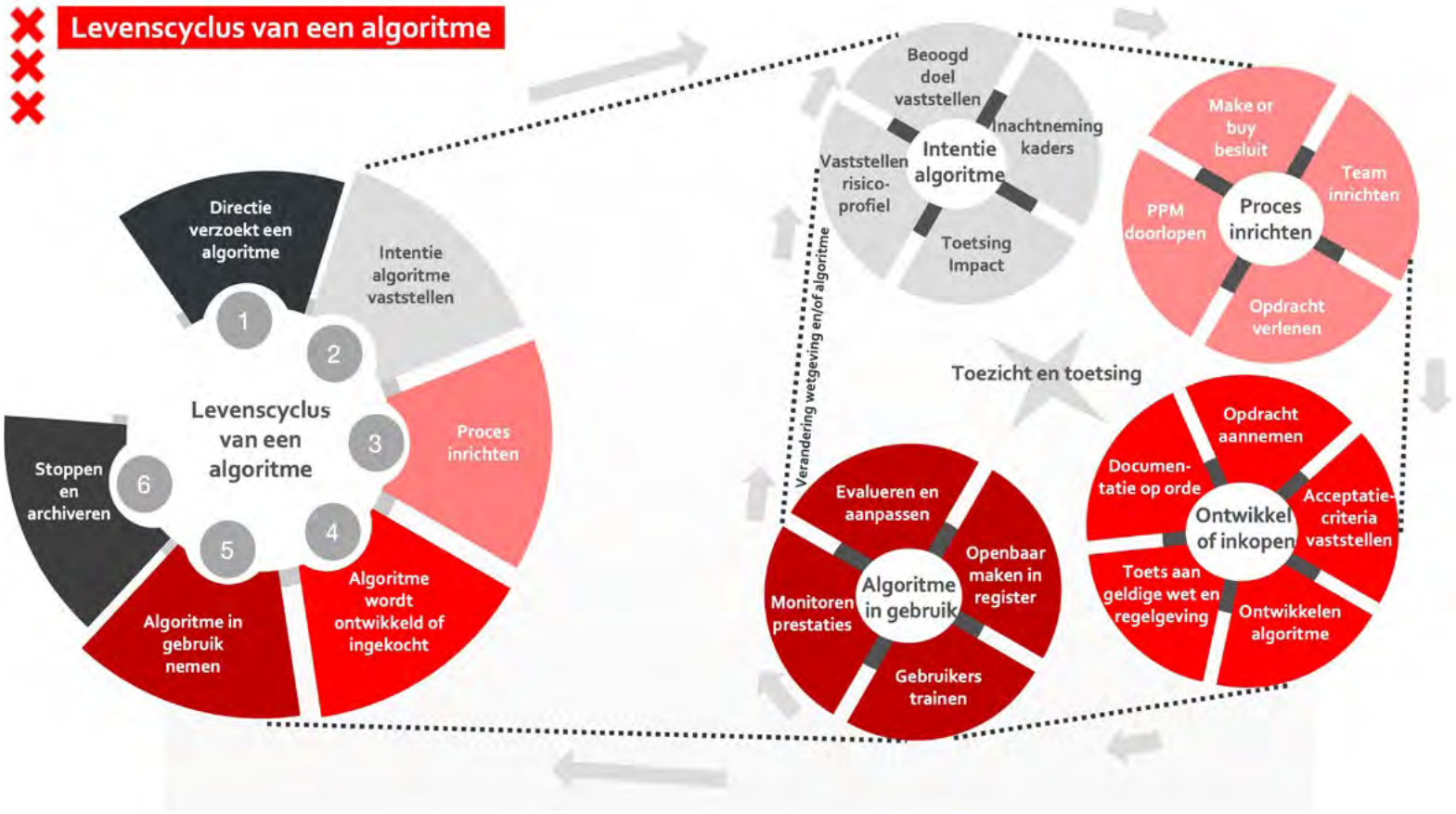
Governance establishment and lifecycle model.
The governance description outlines:

- defined duties and responsibilities.
- the measures needed to avoid risks when using algorithms.
- the information we need to lay down.
- the responsible party in case an algorithm does not meet its purpose.

The Algorithm Lifecycle Approach is aimed at developing, managing and archiving algorithms in use by the municipality. The tools focus on a grip on and development of the algorithm's 'lifecycle' from beginning to end. Determining responsibilities is a best practice. This starting point is decisive in guaranteeing algorithms' quality and transparency. But we still have a way to go. We sometimes use algorithms that have not, or only partly, been developed by us. Consider, the LAA algorithm used to combat address fraud. As a municipality, we are the user of an algorithm from the government.

Challenges:

- The use of algorithms by departments is not always visible. Algorithms may be 'hidden' in systems or mathematical models. Identifying algorithms within the organisation becomes a challenge.
- We are currently streamlining the established tools to hold a grip on algorithms. It avoids having to overload the organisation with questions. This could include the link between the GDPR tools and the Algorithm Lifecycle tools.
- In the near future, an awareness campaign will start within the civil service organisation on the development and use of algorithms.



Independent advice and supervision by the Amsterdam Personal Data Committee

The City of Amsterdam Executive has set up the Amsterdam Personal Data Committee. This committee advises the City of Amsterdam Executive on data processing. This also includes the ethical assessment in the use of algorithms. The committee's working method helps the transparency involved in processing personal data when algorithms are used by the municipality. The Amsterdam Personal Data Committee does so particularly by organising public meetings and by issuing opinions.
The Amsterdam Personal Data Committee advises on the use, thereby taking the following into account:

- laws and regulations.
- social and ethical insights.
- technological developments.

The Amsterdam Personal Data Committee does so in case of deviation from the ethical principles for data and algorithms or questions about the use of algorithms. The Amsterdam Personal Data Regulations were expanded in 2021 to include algorithms, data ethics, digital human rights and the exposure of personal data.

Tool 5: Audit

DThe service processes in which we use algorithms can be audited. An audit verifies that the appropriate measures are in place to prevent risks, such as those of discrimination or a violation of rights. Each year, the CIO office commissions an audit on algorithms. That audit is carried out to assess the use of algorithms against the different frameworks and laws and regulations. We use the audits as a tool to improve as we learn. We also learn from non-internal audits:

- In 2018, an audit was carried out for SIA. The audit framework for Amsterdam was then used for the first time.
- The Amsterdam Metropolitan Court of Audit is currently performing an audit. The key question for this audit: to what extent is the Amsterdam algorithm framework sufficient for a responsible application of algorithms. And what lessons can we learn from the use of algorithms in everyday Amsterdam practices. The Amsterdam Metropolitan Court of Audit published its audit design on 28 July 2022. The expected publication date of the audit is scheduled for Q3 2023.

More information (in Dutch)

Tool 6: Bias analysis

How do we make sure that we develop and apply fair algorithms? How do we deal with possible unforeseen biases? To this end, we have drawn up a bias analysis document, which has been adopted by our Municipal Management Team. The document serves as a first draft and will in practice be used together with the Fairness Handbook.
The bias analysis includes the following components:

- Defining the ('sensitive') groups to be studied
- Drafting hypotheses on features that may lead to indirect biases
- Selecting metrics that fit the project
- Analysing direct bias
- Analysing indirect bias
- Analysing bias on non-measurable variables
- Weighing and reviewing biases found with responsible management
- Mitigating biases where necessary
- Drafting conclusions.
- The Fairness Handbook provides guidance and a guideline for fair algorithms.

Algorithms can produce decisions and outcomes that discriminate against individuals and against demographic groups. That is why we need to build and apply algorithms properly. However, checking and assessing the fairness of models is an enormous task. It requires collaboration between different stakeholders, including data scientists, subject-matter experts and end users.

We want to minimise the chances that algorithms can harm citizens as much as possible. That is why we issued the bias analysis and the Fairness Handbook. That handbook offers a practical set of tools which can be used to assess the fairness of a model. The handbook provides a basic explanation on algorithmic justice and possible biases and serves all those dealing with data or algorithms at work. It explains how biases and other problems in the development cycle of the model can cause various forms of damage, which in turn may affect individuals or disadvantaged groups in society.

Both the bias analysis and the Fairness Pipeline offer a step-by-step plan. We can use that plan to evaluate a model for biases and minimise these problems.

Go to the Fairness Handbook

Tool 7: Impact assessment

The City of Amsterdam has drawn up a Human Rights Impact Analysis, including a bias analysis. This tool was drafted based on practical lessons learned and existing tools, and also because there were no tools previously that could be deployed on this issue. In governance, we argue that, in addition to the Human Rights Impact Assessment, other human rights impact assessments may also be used, such as the government's Human Rights and Algorithms Impact Assessment (IAMA). The government's IAMA may serve as a tool. Utrecht University has developed this tool. The IAMA is a type of guideline that supports government organisations in making decisions about the development and deployment of algorithms. Step by step, it describes discussion points that must be addressed before the algorithm is implemented. By shedding light on the course of a careful decision-making and implementation process for algorithms, IAMA can help prevent situations such as the Dutch childcare benefits scandal (Toeslagenaffaire). The tool ensures that all the different parties responsible for the development or deployment of an algorithmic system enter into discussions with each other.

Download the IAMA (in Dutch)

Lessons learned

We began by developing a number of practical policy tools. We did so to get a better grip on the algorithms that we use. We have since been doing this together with a lot of other government bodies. The support of the Ministry of the Interior and Kingdom Relations via the Innovation Fund was the decisive factor in this. The City of Amsterdam Algorithm Life-cycle Approach shows that, in practice, there is a demand for the exchange of knowledge and experiences between national and international governments and institutions. All the more reason to have an open and iterative process. In such a process, the open-source tools can be reused and adapted to third-party needs. The key to success is a national government that assumes a hands-on and encouraging attitude. Such attitude requires the alignment of policy, the legal aspects and societal and social perceptions as much as possible.

A number of examples include:

- The importance of an Algorithm Register whereby we always consider the point of view of our citizens. The register offers proper information to citizens, which can thus help to build confidence in the government. And democracy. The register only has that function if it contains all the algorithms that affect personal life. Meaning, not only from governments, but also from semi-public organisations and the business community.
- In addition to transparency, the deciding factor for a successful Algorithm Register is accountability. Accountability starts by recognising the rapid development and application of algorithms in today's society. On the one hand, it requires a clear framework that is practically usable and can evolve both nationally and internationally: when is something an algorithm? On the other hand, it is recommended to be flexible and have an open design. This will help to apply the experiences and lessons learned from various studies by knowledge institutions, watchdogs and other cities and countries in the future. Governance implementation and alignment with internal organisational processes are important in this regard.
- The responses to the Algorithm Register show that there is a greater need to understand how 'digital public space' affects people's lives. And how government organisations can arrange this. The first recommendation is to design the register in such a way that it is adaptable for other types of technologies. Starting with the national sensor register. See also the appeal with the Association of Netherlands Municipalities. The second recommendation is to help other government bodies and semi-public organisations. We have been receiving national and international requests from governments and the academic world for more information about the Amsterdam Algorithm Register, for example, from the Consortium on Public Control of Algorithms. This shows the importance of national coordination for effective deployment of resources. An open, iterative approach is best for lasting and best possible register alignment. Also, for a longer period in a changing legal, technological and sociological environment.

Other lessons learned include:

- Including new algorithms in the register is easier than including existing algorithms that are already in use.
- We have learned that we need to streamline the tools to avoid regulatory burden.
- Ownership is a decisive factor. Who is the algorithm from? What laws and regulations support the algorithm? By including the algorithm in the register, we also make it clear internally who in fact is in charge of the algorithm.
- During the Alpha Audit, we concluded that the Amsterdam tool did not keep pace with practice. We decided to use the nationwide framework. We also decided to explore the best fit: the nationwide framework or an adjusted version of the Amsterdam framework.
- Recognising (high-risk) algorithms in existing systems is difficult.
- We are accustomed to emphasising that which is not allowed. How is proactive insight into how to actually use algorithms possible?
- These tools will have to become part of a broader approach on the use of technology in society. An Algorithm Register exclusively for the purpose of one municipality gives a limited perspective. As a resident and critical expert, you should be able to check algorithms of at least all public authorities.

Publication details

For any questions or contact:
algoritmen@amsterdam.nl

Published by:
City of Amsterdam,
December 2022

Author:
Siham El Yassini, Aik van Eemeren,
Lydia Prinsen

Editor:
Daniël Boon-Gerrits

Design:
Vorm de Stad

Partner institutions:
Consortium on Public Control of Algorithms, consisting of:

The Amsterdam University of Applied Sciences
The Association of Netherlands Municipalities (VNG)
The Association of Water Boards
The City of The Hague
The City of Rotterdam
The City of Utrecht
Delft University of Technology
European Commission: DG GROW, DG CONNECT, Living-in.eu
Eurocities
The Ministry of the Interior and Kingdom Relations
The Ministry of Infrastructure and Water Management
The Netherlands Court of Audit
The police
The Province of Limburg
The Province of Noord-Brabant
The Province of Zuid-Holland
Saidot
Pels Rijcken

Opinion of the Data Ethics Commission

daten
ethik
kommission

Opinion of the Data Ethics Commission

daten
ethik
kommission

Content overview

Table of contents

Executive Summary

Our society is experiencing profound changes brought about by digitalisation. Innovative data-based technologies may benefit us at both the individual and the wider societal levels, as well as potentially boosting economic productivity, promoting sustainability and catalysing huge strides forward in terms of scientific progress. At the same time, however, digitalisation poses risks to our fundamental rights and freedoms. It raises a wide range of ethical and legal questions centring around two wider issues: the role we want these new technologies to play, and their design. If we want to ensure that digital transformation serves the good of society as a whole, both society itself and its elected political representatives must engage in a debate on how to use and shape data-based technologies, including artificial intelligence (AI).

Germany's Federal Government set up the Data Ethics Commission *(Datenethikkommission)* on 18 July 2018. It was given a one-year mandate to develop ethical benchmarks and guidelines as well as specific recommendations for action, aiming at protecting the individual, preserving social cohesion, and safeguarding and promoting prosperity in the information age. As a starting point, the Federal Government presented the Data Ethics Commission with a number of key questions clustered around three main topics: algorithm-based decision-making (ADM), AI and data. In the opinion of the Data Ethics Commission, however, AI is merely one among many possible variants of an algorithmic system, and has much in common with other such systems in terms of the ethical and legal questions it raises. With this in mind, the Data Ethics Commission has structured its work under two different headings: **data** and **algorithmic systems** (in the broader sense).

In preparing its Opinion, the Data Ethics Commission was inspired by the following **guiding motifs**:

- Ensuring the human-centred and value-oriented design of technology
- Fostering digital skills and critical reflection in the digital world
- Enhancing protection for individual freedom, self-determination and integrity
- Fostering responsible data utilisation that is compatible with the public good
- Introducing risk-adapted regulation and effective oversight of algorithmic systems
- Safeguarding and promoting democracy and social cohesion
- Aligning digital strategies with sustainability goals
- Strengthening the digital sovereignty of both Germany and Europe.

1

General ethical and legal principles

Humans are morally responsible for their actions, and there is no escaping this moral dimension. Humans are responsible for the goals they pursue, the means by which they pursue them, and their reasons for doing so. Both this dimension and the societal conditionality of human action must always be taken into account when designing our technologically shaped future. At the same time, the notion that technology should serve humans rather than humans being subservient to technology can be taken as incontrovertible fact. Germany's constitutional system is founded on this **understanding of human nature**, and it adheres to the tradition of Europe's cultural and intellectual history.

Digital technologies have not altered our ethical framework – in terms of the basic values, rights and freedoms enshrined in the German Constitution and in the Charter of Fundamental Rights of the European Union. Yet the new challenges we are facing mean that we need to reassert these values, rights and freedoms and perform new balancing exercises. With this in mind, the Data Ethics Commission believes that the following ethical and legal principles and precepts should be viewed as indispensable and socially accepted benchmarks for action.

Human dignity
Human dignity is a principle that presupposes the unconditional value of every human being, prohibiting such practices as the total digital monitoring of the individual or his or her humiliation through deception, manipulation or exclusion.

Self-determination
Self-determination is a fundamental expression of freedom, and encompasses the notion of informational self-determination. The term "digital self-determination" can be used to express the idea of a human being a self-determined player in a data society.

Privacy
The right to privacy is intended to preserve an individual's freedom and the integrity of his or her personal identity. Potential threats to privacy include the wholesale collection and evaluation of data about even the most intimate of topics.

Security
The principle of security relates not only to the physical and emotional safety of humans but also to environmental protection, and as such involves the preservation of vitally important assets. Guaranteeing security entails compliance with stringent requirements, e.g. in relation to human/machine interaction or system resilience to attacks and misuse.

Democracy

Digital technologies are of systemic relevance to the flourishing of democracy. They make it possible to shape new forms of political participation, but they also foster the emergence of threats such as manipulation and radicalisation.

Justice and Solidarity

In view of the vast amounts of power being accumulated using data and technologies, and the new threats of exclusion and discrimination, the safeguarding of equitable access and distributive justice is an urgent task. Digitalisation should foster participation in society and thereby promote social cohesion.

Sustainability

Digital developments also serve sustainable development. Digital technologies should contribute towards achieving economic, ecological and social sustainability goals.

Ethics cannot be equated on a one-to-one basis with the law. In other words, not everything that is relevant from an ethical perspective can and should be enshrined in legislation; conversely, there are provisions of the law that are motivated purely by pragmatic considerations. Nevertheless, the law must, at all times, be heedful of the potential ethical implications of the legal provisions in force, as well as living up to ethical standards. The Data Ethics Commission holds the view that **regulation is necessary, and cannot be replaced by ethical principles**. This is particularly true for issues with heightened implications for fundamental rights that require the central decisions to be made by the democratically elected legislator. Regulation is also an essential basis for building a system where citizens, companies and institutions can trust that the transformation of society will be guided by ethical principles.

At the same time, regulation must not unduly inhibit technological and social innovation and dynamic market growth. Overly rigid laws that attempt to regulate every last detail of a situation may place a stranglehold on progress and increase red tape to such an extent that innovation by German companies can no longer keep pace with the rate of technological development on the international stage.

Yet legislation is only one of a range of tools that can be used to lend tangible shape to ethical principles. The **synergistic use of various governance instruments** at different levels (multi-level governance) is vital in view of the complexity and dynamism of data ecosystems. These instruments include not only legislative measures and standardisation, but also various forms of co-regulation or self-regulation. Technology and technological design can moreover function as governance instruments themselves, and the same applies to business models and options for steering the economy. Governance in the broader sense also encompasses policy-making decisions in the fields of education and research. It is important to consider each of the aforesaid governance instruments not only in a national context, but also (and in particular) in their **European and international** contexts.

In the view of the Data Ethics Commission, all of the key questions presented by the Federal Government belong to one of two different perspectives: questions that concentrate mainly on data (the **"data perspective"**) and questions that are primarily focused on algorithmic systems (the **"algorithms perspective"**). These two perspectives should not be regarded as competing views or two sides of the same coin; instead, they represent two different **ethical discourses, which both complement each other and are contingent upon each other**. These different ethical discourses are typically also reflected in different governance instruments, including in different acts of legislation.

2

Data

The **data perspective** focuses on digital data, which are used for machine learning, as a basis for algorithmically shaped decisions, and for a plethora of further purposes. This perspective considers data primarily with a view to their **origin** and to the potential **impact** their processing may have on certain parties who are involved with the data, such as by being the data subject, as well as on society at large. From an ethical and legal point of view, it is important to identify **standards for data governance**; typically, however, **rights** that parties involved with the data can enforce against others will play an even more significant role. A central distinction in this context is that between personal and non-personal data, since it determines whether the provisions of data protection law apply.

General standards for data governance

In the opinion of the Data Ethics Commission, responsible data governance must be guided by the following data ethics principles:

- **Foresighted responsibility:** Possible future cumulative effects, network effects and effects of scale, technological developments and changing actor constellations must be taken into account when gauging the potential impact of collecting, processing and forwarding data on individuals or the general public.

- **Respect for the rights of the parties involved:** Parties who have been involved in the generation of data – whether as data subjects or in a different role – may have rights in relation to such data, and these rights must be respected.

- **Data use and data sharing for the public good:** As a non-rivalrous resource, data can be duplicated and used in parallel by many different individuals for many different purposes, thereby furthering the public good.

- **Fit-for-purpose data quality:** Responsible use of data includes ensuring a high level of data quality that is fit for the relevant purpose.

- **Risk-adequate level of information security:** Data are vulnerable to external attacks, and it is difficult to recover them once they have gone astray. The standard of information security applied must therefore be commensurate with the potential for risk inherent to the situation in question.

- **Interest-oriented transparency:** Controllers must be prepared and in a position to account for their data-related activities. This requires appropriate documentation and transparency and, if necessary, a corresponding liability regime in place.

Data rights and corresponding obligations

For self-determined navigation in the data society, parties must have, and be able to enforce, certain data-related rights against others. First and foremost among these rights are those relating to an individual's **personal data**, which derive from the right to informational self-determination that is enshrined as a fundamental right, and which are guaranteed by the applicable data protection law. Digital self-determination in the data society also includes the self-determined economic exploitation of one's own data, and it includes self-determined management of **non-personal data**, such as non-personal data generated by one's own devices. The Data Ethics Commission takes the view that, in principle, a right to digital self-determination in the data society also applies to companies and **legal entities** and – at least to some extent – to groups of persons (collectives).

Data are often generated with contributions from different parties who are acting in different roles – be it as the data subject, be it as the owner of a data-generating device or be it in yet another role. In the opinion of the Data Ethics Commission such contributions to the generation of data should not lead to exclusive ownership rights in data, but rather to **data-specific rights of co-determination and participation**, which in turn may lead to corresponding obligations on the part of other parties. The extent to which an individual should be entitled to data rights of this kind, and the shape they should take, depends on the following general factors:

a) the nature and scope of that party's **contribution to data generation**,

b) the **weight of that party's legitimate interest** in being granted the data right,

c) the weight of any possibly **conflicting interests** on the part of the other party or of third parties, taking into account any potential compensation arrangements (e.g. protective measures, remuneration),

d) the **interests of the general public**, and

e) the **balance of power** between the parties involved.

Data rights may allow their holders to pursue a number of different goals, in particular the following:

- requiring that a controller **desist from data use** (up to a right to require erasure of the data),
- requiring that a controller **rectify the data**,
- requiring that a controller grant **access to data** (up to full data portability), or
- requiring an **economic share** in profits derived with the help of the data.

For each type of data right (desistance, rectification, access, economic share) there exists a **separate set of conditions** defining, e.g., what counts as a party's legitimate interest in being granted the data right. For determining whether a party has a right to require desistance from a particular data use, key considerations include the potential for harm associated with said use and the circumstances under which the party in question had contributed to generating the data. Potential for harm may also be relevant when a request is made to rectify data, but the benchmark is lower in this respect. Where a party requests access to data, there is a graded spectrum of interests that count as a legitimate interest to be granted such access, which is particularly relevant within existing value creation systems. Only under very narrowly defined conditions may a party have an independent claim to an economic share in profits derived by others. The **rights granted to data subjects** under the EU's General Data Protection Regulation (GDPR) are a particularly important manifestation of these data rights, aimed specifically at protecting the natural persons to whom the data pertain; they are also to some extent a standardised manifestation given that they hinge on the qualification of data as personal data.

Considering these principles, the Data Ethics Commission wishes to submit the following key recommendations for action:

Standards for the use of personal data

1

The Data Ethics Commission recommends that **measures be taken against ethically indefensible uses of data**. Examples of these uses include total surveillance, profiling that poses a threat to personal integrity, the targeted exploitation of vulnerabilities, addictive designs and dark patterns, methods of influencing political elections that are incompatible with the principle of democracy, vendor lock-in and systematic consumer detriment, and many practices that involve trading in personal data.

2

Data protection law as well as other branches of the legal system (including general private law and unfair commercial practices law) already provide for a range of instruments that can be used to prevent such ethically indefensible uses of data. However, in spite of the widespread impact and enormous potential for harm, too little has been done to date in terms of harnessing the power of these instruments, particularly against the market giants. The various factors contributing to this **enforcement gap** must be tackled systematically.

3

As well as steps to make front-line players (e.g. supervisory authorities) more aware of the existing options, there is an urgent need for the **legislative framework in force to be fleshed out more clearly and strengthened in certain areas**. Examples of recommended measures include the blacklisting of data-specific unfair contract terms, the fleshing out of data-specific contractual duties of a fiduciary nature, new data-specific torts, the blacklisting of certain data-specific unfair commercial practices and the introduction of a much more detailed legislative framework for profiling, scoring and data trading.

4

In order to allow supervisory authorities to take action more effectively, these authorities need significantly better human and material resources. Attempts should be made to strengthen and formalise cooperation between the different data protection authorities in Germany, thereby ensuring the uniform and coherent application of data protection law. If these attempts fail, consideration should be given to the **centralisation of market-related supervisory activities** within a federal-level authority that is granted a broad mandate and that cooperates closely with other specialist supervisory authorities. The authorities at *Land* level should remain responsible for supervisory activities relating to the public sector, however.

5

The Data Ethics Commission believes that **"data ownership"** (i.e. exclusive rights in data modelled on the ownership of tangible assets or on intellectual property) would not solve any of the problems we are currently facing, but would create new problems instead, and **recommends refraining from their recognition**. It also advises against granting to data subjects copyright-like rights of economic exploitation in respect of their personal data (which might then be managed by collective societies).

6

The Data Ethics Commission also argues that **data should not be referred to as "counter-performance"** provided in exchange for a service, even though the term sums up the issue in a nutshell and has helped to raise awareness among the general public. Regardless of the position that data protection authorities and the European Court of Justice will ultimately take with regard to the prohibition under the GDPR of "tying" or "bundling" consent with the provision of a service, the Data Ethics Commission believes that consumers must be offered **reasonable alternatives** to releasing their data for commercial use (e.g. appropriately designed **pay options**).

7

Stringent requirements and limitations should be imposed on the use of data for **personalised risk assessment** (e.g. the "black box" premiums in certain insurance schemes). In particular, the processing of data may not intrude on intimate areas of private life, there must be a clear causal relationship between the data and the risk, and the difference between individual prices charged on the basis of personalised and non-personalised risk assessments should not exceed certain percentages (to be determined). There should also be stringent requirements in respect of transparency, non-discrimination and the protection of third parties.

8

The Data Ethics Commission advises the Federal Government not to consider the issues falling under the heading of **"digital inheritance"** as having been settled by the Federal Court of Justice's 2018 ruling. The ephemeral spoken word is being replaced in many situations by digital communications that are recorded more or less in their entirety, and the possibility that these records will be handed over to a deceased's heirs adds a whole new dimension of privacy risk. A range of mitigating measures should be taken, including the imposition of new obligations on service providers, quality assurance standards for digital estate planning services and national regulations on post-mortem data protection.

9

The Data Ethics Commission recommends that the Federal Government should invite the social partners to work towards a common position on the legislative provisions that should be adopted with a view to **stepping up the protection of employee data**, based on examples of best practices from existing collective agreements. The concerns of individuals in non-standard forms of employment should also be taken into account during this process.

10

In view of the benefits that could be gained from **digitalising healthcare**, the Data Ethics Commission recommends swift expansion of digital infrastructures in this sector. The expansion of both the range and the quality of digitalised healthcare services should include measures to better allow patients to exercise their rights to informational self-determination. Measures that could be taken in this respect include the introduction and roll-out of an electronic health record, building on a participatory process that involves the relevant stakeholders, and the further development of procedures for reviewing and assessing digital medical apps in the insurer-funded and consumer-funded health markets.

11

The Data Ethics Commission calls for action against the significant enforcement gap that exists with regard to statutory **protection of children and young people** in the digital sphere. Particular attention should be paid to the development and mandatory provision of technologies (including effective identity management) and default settings that not only guarantee reliable protection of children and young people but that are also family-friendly, i.e. that neither demand too much of parents or guardians nor allow or even encourage excessive surveillance in the home environment.

12

Standards and guidelines on the handling of the personal data of **vulnerable and care-dependent persons** should be introduced to provide greater legal certainty for professionals in the care sector. At the same time, consideration should be given to clarifying in the relevant legal provisions on living wills that these may also include dispositions with regard to the future processing of personal data as far as such processing will require the care-dependent person's consent (e.g. for dementia patients who will not be in a position to provide legally valid consent).

13

The Data Ethics Commission believes that a number of binding requirements should be introduced to ensure the **privacy-friendly design of products and services**, so that the principles of privacy by design and privacy by default (which the GDPR imposes on controllers) will already be put into practice upstream, by manufacturers and service providers themselves. Such requirements would be particularly important with regard to consumer equipment. In this context, standardised icons should also be introduced so that consumers are able to take informed purchase decisions.

14

Action must also be taken at a number of different levels to provide manufacturers with adequate **incentives to implement features of privacy-friendly design**. This includes effective legal remedies that can be pursued against parties along the entire distribution chain to ensure that also manufacturers can be held accountable for inadequate application of the principles of privacy by design and privacy by default. Consideration should also be given, in particular, to requirements built into tender specifications, procurement guidelines for public bodies and conditions for funding programmes. The same applies to **privacy-friendly product development**, including the training of algorithmic systems.

15

While debates on data protection tend (quite rightly) to centre around natural persons, it is important not to ignore the fact that **companies and legal persons must also be granted protection**. The almost limitless ability to pool together individual pieces of data can be used as a means of obtaining a comprehensive picture of a company's internal operating procedures, and this information can be passed on to competitors, negotiating partners, parties interested in a takeover bid and so on. This poses a variety of threats – *inter alia* to the digital sovereignty of both Germany and Europe – in view of the significant volumes of data that flow to third countries. Many of the Data Ethics Commission's recommendations for action therefore also apply on a *mutatis mutandis* basis to the data of legal persons. The Data Ethics Commission believes that action must be taken by the Federal Government to **step up the level of data-related protection afforded to companies**.

Improving controlled access to personal data

16

The Data Ethics Commission identifies enormous potential in the use of data for research purposes that serve a public interest (e.g. to improve healthcare provision). Data protection law as it currently stands acknowledges this potential, in principle, by granting far-reaching privileges for the processing of personal data for research purposes. Uncertainty persists, however, in particular as regards the scope of the so-called research privilege for secondary use of data, and the scope of what counts as "research" in the context of product development. The Data Ethics Commission believes that appropriate **clarifications in the law** are necessary to rectify this situation.

17

The fragmentation of research-specific data protection law, both within Germany itself and among the EU Member States, represents a potential obstacle to data-driven research. The Data Ethics Commission therefore recommends that **research-specific regulations should be harmonised**, both between federal and *Land* level and between the different legal systems within the EU. Introducing a notification requirement for research-specific national law could also bring some improvement, as could the establishment of a European clearing house for cross-border research projects.

18

In the case of research involving particularly sensitive categories of personal data (e.g. health data), **guidelines** should be produced with information for researchers on how to obtain consent in a legally compliant manner, and **innovative consent models should be promoted and explicitly recognised by the law**. Potential options include the development and roll-out of digital consent assistants or the recognition of so-called meta consent, alongside further endeavours to clarify the scope of the research privilege for secondary use of data.

19

The Data Ethics Commission supports, in principle, the move towards a **"learning healthcare system"**, in which healthcare provision is continuously improved by making systematic and quality-oriented use of the health data generated on a day-to-day basis, in keeping with the principles of evidence-based medicine. If further progress is made in this direction, however, greater efforts must be made at the same time to protect data subjects against the significant potential for discrimination that exists when sensitive categories of data are used; this might involve **prohibiting the exploitation of such data** beyond the defined range of purposes.

20

The development of procedures and standards for data **anonymisation** and **pseudonymisation** is central to any efforts to improve controlled access to (formerly) personal data. A legal presumption that, if compliance with the standard has been achieved, data no longer qualify as personal, or that "appropriate safeguards" have been provided in respect of the data subject's rights, would improve legal certainty by a long way. These measures should be accompanied by rules that – on pain of criminal penalty – prohibit the de-anonymisation of anonymised data (e.g. because new technology becomes available that would allow the re-identification of data subjects) or the reversal of pseudonymisation, both in the absence of narrowly defined grounds for doing so. Also research in the field of **synthetic data** shows enormous promise, and more funding should be funnelled into this area.

21

Fundamentally speaking, the Data Ethics Commission believes that **innovative data management and data trust schemes** hold great potential, provided that these systems are designed to be robust, suited to real-life applications and compliant with data protection law. A broad spectrum of models falls under this heading, ranging from dashboards that perform a purely technical function (**privacy management tools**, PMT) right through to comprehensive data and consent management services (**personal information management services**, PIMS). The underlying aim is to empower individuals to take control over their personal data, while not overburdening them with decisions that are beyond their capabilities. The Data Ethics Commission recommends that research and development in the field of data management and data trust schemes should be identified as a funding priority, but also wishes to make it clear that adequate protection of the rights and legitimate interests of all parties involved will require additional **regulatory measures at EU level**. These regulatory measures would need to secure central functions without which operators cannot be active, since their scope for action would otherwise be very limited. On the other hand, it is also necessary to protect individuals against parties that they assume to be acting in their interests, but that, in reality, are prioritising their own financial aims or the interests of others. In the event that a feasible method of protection can be found, data trust schemes could serve as vitally important mediators between data protection interests and data economy interests.

22

As far as the right to **data portability** enshrined in Article 20 GDPR is concerned, the Data Ethics Commission recommends that industry-specific codes of conduct and standards on data formats should be adopted. Given that the underlying purpose of Article 20 GDPR is not only to make it more straightforward to change provider, but also to allow other providers to access data more easily, it is important to evaluate carefully the market impact of the existing right to portability and to analyse potential mechanisms by which it can be prevented that a small number of providers increase yet further their market power. Until the findings of this evaluation are available, expansion of the scope of this right (for example to cover data other than data provided by the data subject, or real-time porting of data) would seem premature and not advisable.

23

In certain sectors, for example messenger services and social networks, **interoperability or interconnectivity obligations** might help to reduce the market entry barriers for new providers. Such obligations should be designed on an asymmetric basis, i.e. the stringency of the regulation should increase in step with the company's market share. Interoperability and interconnectivity obligations would also be a prerequisite for building up or strengthening, within and for Europe, certain basic services of an information society.

Debates around access to non-personal data

24

Access by European companies to appropriate non-personal data of appropriate quality is a key factor for the growth of the European data economy. In order to benefit from enhanced **access to data**, however, stakeholders must have a sufficient degree of data-awareness and have the data skills that are necessary to make use of the data. Also, access to data proves to be disproportionately advantageous to stakeholders that have already built up the largest reserves of data and that have the best data infrastructures at hand. The Data Ethics Commission therefore wishes to stress that the factors referred to should always receive due attention when discussing whether and how to improve data access, in keeping with the **ASISA principle** *(Awareness – Skills – Infrastructures – Stocks – Access)*.

25

The Data Ethics Commission therefore supports the efforts already initiated at European level to promote and improve **data infrastructures** in the broadest sense of the term (e.g. platforms, standards for application programming interfaces and other elements, model contracts, EU Support Centre), and recommends to the Federal Government that these efforts should continue to be matched by corresponding efforts at national level. It would also be advisable to set up an ombudsman's office at federal level to provide assistance and support in relation to the negotiation of data access agreements and dispute settlement.

26

The Data Ethics Commission ascribes enormous importance to a holistically conceived, sustainable and strategic **economic policy** that outlines effective methods of preventing not only the exodus of innovative European companies or their acquisition by third-country companies, but also an excessive dependence on third-country infrastructures (e.g. server capacities). A balance must be struck in this context between much-needed international cooperation and networking on the one hand, and on the other a resolute assumption of responsibility for sustainable security and prosperity in Europe against the backdrop of an ever-evolving global power dynamic.

27

Also from the perspective of boosting the European data economy, the Data Ethics Commission does not see any benefit in introducing new exclusive rights ("data ownership", "data producer right"). Instead, it recommends affording **limited third-party effects to contractual agreements** (e.g. to restrictions on data utilisation and onward transfer of data by a recipient). These third-party effects could be modelled on the new European regime for the protection of trade secrets. The Data Ethics Commission also recommends the adoption of legislative solutions enabling European companies to cooperate in their use of data, for example by using data trust schemes, without running afoul of anti-trust law (**"data partnerships"**).

28

The data accumulated in existing value creation systems (e.g. production and distribution chains) are often of enormous commercial significance, both inside and outside that value creation system. In many cases, however, the provisions on data access that appear in the contractual agreements concluded within a value creation system are unfair and/or inefficient, or lacking entirely; in certain cases, there is no contractual agreement at all. Efforts must therefore be made to **raise awareness among businesses** in sectors far outside what is commonly perceived as the "data economy", and to provide practical guidance and support (e.g. model contracts).

29

The Data Ethics Commission furthermore recommends cautious **adaptations of the current legislative framework**. The first stage in this process should be to make explicit reference in Section 311 of the [German] Civil Code (*Bürgerliches Gesetzbuch*, BGB) to the special relationship that exists between a party that has contributed to the generation of data in a value creation system and the controller of the data, clarifying that such parties may have certain quasi-contractual duties of a fiduciary nature. These duties should normally include a duty to enter into negotiations about fair and efficient

data access arrangements. Consideration should also be given to whether additional steps should be taken, which could range from blacklisting particular contract terms also for B2B transactions, to formulating default provisions for data contracts, to introducing sector-specific data access rights.

30

The Data Ethics Commission believes that **open government data (OGD) concepts** hold enormous potential, and recommends that these concepts should be built on and promoted. It also recommends a series of measures to promote a **shift in mindset among public authorities** (something that has not yet fully taken place) and to make it easier in practice to share data on the basis of OGD concepts. These measures include not only the establishment of the relevant **infrastructures** (e.g. platforms), but also harmonisation and improvement of the existing **legal framework** that is currently fragmented and sometimes inconsistent.

31

Nevertheless, the Data Ethics Commission identifies a degree of tension between efforts to promote OGD (relying on principles such as "open by default" and "open for all purposes"), and efforts to enhance data protection and the protection of trade secrets (with legally enshrined concepts such as "privacy by default"). The Data Ethics Commission submits that, in cases of doubt, **priority should be given to the duty of protecting** individuals and companies who have entrusted their data to the State (often without being given any choice in the matter, e.g. tax information). The State must deliver on this duty by implementing a range of different measures, which may include technical as well as legal safeguards against misuse of data.

32

In particular, it would be beneficial to develop **standard licences and model terms and conditions** for public-sector data sharing arrangements, and to make their use mandatory (at least on a sector-specific basis). These standard licenses and model terms and conditions should include clearly defined safeguards for the rights of third parties who are affected by a data access arrangement.

Provision should also be made against data being used in a way that ultimately harms public interests, and also against still greater accumulation of data and market power on the part of the big players (which would be likely to undermine competition) and against the taxpayer having to pay twice.

33

As regards **open-data concepts in the private sector**, priority should be given to **promoting and supporting voluntary data-sharing arrangements**. Consideration must be given not only to the improvement of infrastructures (e.g. data platforms), but also to a broad range of potential incentives; these might include certain privileges in the context of tax breaks, public procurement, funding programmes or licensing procedures. Statutory data access rights and corresponding obligations to grant access should be considered as fall-back options if the above measures fail to deliver the desired outcomes.

34

Generally speaking, the Data Ethics Commission believes that a cautious approach should be taken to the introduction of statutory data access rights; ideally such rights should be developed only on a **sector-by-sector basis**. Sectors in which the level of demand should be analysed include the media, mobility or energy sectors. In any case, before a statutory data access right or even a disclosure obligation is introduced, a full impact assessment needs to be carried out, examining and weighing up against each other all possible implications; these include implications for data protection and the protection of trade secrets, for investment decisions and the distribution of market power, as well as for the strategic interests of German and European companies compared to those of companies in third countries.

35

The Data Ethics Commission recommends considering enhanced obligations of private enterprises to grant access to data **for public interest and public-sector purposes** (business-to-government, B2G). A cautious and sector-specific approach is, however, recommended in this respect as well.

3

Algorithmic systems

The **algorithms perspective** focuses on the architecture of data-driven algorithmic systems, their dynamics and the systems' impacts on individuals and society. The ethical and legal discourse in this area typically centres around the relationship between humans and machines, with a particular focus on automation and the outsourcing of increasingly complex operational and decision-making processes to "autonomous" systems enabled by AI. The algorithms perspective differs from the data perspective in that the data processed by the system might have no connection whatsoever with the persons affected by it; in particular, individuals may suffer ethically indefensible implications even if all of the data used (e.g. to train an algorithmic system) are non-personal. The current debates on "algorithmic oversight" or liability for AI are of central importance in this respect.

General standards for algorithmic systems

The Data Ethics Commission distinguishes between three different levels of algorithmic involvement in human decision-making, based on the distribution of tasks between the human and the machine in the specific case in question:

a) **algorithm-based** decisions are human decisions based either in whole or in part on information obtained using algorithmic calculations,

b) **algorithm-driven** decisions are human decisions shaped by the outputs of algorithmic systems in such a way that the human's factual decision-making abilities and capacity for self-determination are restricted,

c) **algorithm-determined** decisions trigger consequences automatically; no provision is made for a human decision in the individual case.

In the opinion of the Data Ethics Commission, the following principles should be observed to ensure the responsible use of algorithmic systems.

- **Human-centred design:** Systems must be centred around the human who uses them or who is affected by their decisions; they must prioritise his or her fundamental rights and freedoms, basic needs, physical and emotional well-being and skills development.

- **Compatibility with core societal values:** The process of system design must take account of the system's impact on society as a whole, and in particular its effects on the democratic process, on the citizen-centred nature of state action, on competition, on the future of work and on the digital sovereignty of Germany and Europe.

- **Sustainability:** Considerations relating to the availability of human skills, participation, environmental protection, sustainable resource management and sustainable economic activity are becoming increasingly important factors in the design and use of algorithmic systems.

- **Quality and performance:** Algorithmic systems must work correctly and reliably so that the goals pursued with their help can be achieved.

- **Robustness and security:** Robust and secure system design involves not only making the system secure against external threats, but also protecting humans and the environment against any negative impacts that may emanate from the system.

- **Minimisation of bias and discrimination:** The decision-making patterns upon which algorithmic systems are based must not be the source of systematic bias or the cause of discriminatory decisions.

- **Transparent, explainable and comprehensible systems:** It is vitally important to ensure not only that the users of algorithmic systems understand how these systems function and can explain and control them, but also that the parties affected by a decision are provided with sufficient information to exercise their rights properly and challenge the decision if necessary.

- **Clear accountability structures:** Questions of the allocation of responsibility and accountability including possible liability arising with the use of algorithmic systems must be unambiguously resolved.

System criticality

The level of **criticality of an algorithmic system** dictates the specific requirements it must meet, in particular with regard to transparency and oversight. System criticality is determined by assessing an algorithmic system's potential for harm, on the basis of a two-pronged investigation into the **likelihood that harm will occur** and the **severity of that harm**.

The **severity** of the harm that could potentially be sustained, for example as a result of a mistaken decision, depends on the significance of the legally protected rights and interests affected (such as the right to privacy, the fundamental right to life and physical integrity, the prohibition of discrimination), the level of potential harm suffered by individuals (including non-material harm or loss of utility that are hard to calculate in monetary terms), the number of individuals affected, the total figure of the harm potentially sustained and the overall harm sustained by society as a whole, which may go well beyond a straightforward summation of the harm suffered by individuals. The **likelihood** that harm will be sustained is also influenced by the properties of the system in question, in particular the role of the algorithmic system components in the decision-making process, the complexity of the decision, the effects of the decision and the reversibility of these effects. The severity and likelihood of the predicted harm may also be contingent on whether the algorithmic systems are operated by the State or by private enterprises and, particularly in a business context, on the market power wielded by the system's operator.

In conclusion, the Data Ethics Commission wishes to make the following recommendations for action on the basis of these principles:

Risk-adapted regulatory approach

36

The Data Ethics Commission recommends adopting a **risk-adapted regulatory approach** to algorithmic systems. The principle underlying this approach should be as follows: the greater the potential for harm, the more stringent the requirements and the more far-reaching the intervention by means of regulatory instruments. When assessing this potential for harm, the **sociotechnical system as a whole** must be considered, or in other words all the components of an algorithmic application, including all the people involved, from the development phase – for example the training data used – right through to its implementation in an application environment and any evaluation and adjustment measures.

37

The Data Ethics Commission recommends that the potential of algorithmic systems to harm individuals and/or society should be determined uniformly on the basis of a **universally applicable model**. For this purpose, the legislator should develop a **criteria-based assessment scheme** as a tool for determining the criticality of algorithmic systems. This scheme should be based on the general ethical and legal principles presented by the Data Ethics Commission.

38

Among other things, the **regulatory instruments and the requirements that apply to algorithmic systems** should include corrective and oversight mechanisms, specifications of transparency, explainability and comprehensibility of the systems' results, and rules on the allocation of responsibility and liability for using the systems.

39

The Data Ethics Commission believes that a useful first stage in determining the potential for harm of algorithmic systems is to distinguish between **five levels of criticality**. Applications that fall under the lowest of these levels (Level 1) are associated with zero or negligible potential for harm, and it is unnecessary to carry out special oversight of them or impose requirements other than the general quality requirements that apply to products irrespective of whether they incorporate algorithmic systems.

40

Applications that fall under Level 2 are associated with **some potential for harm**, and can and should be regulated on an as-needs basis; regulatory instruments used in this connection may include ex-post controls, an obligation to produce and publish an appropriate risk assessment, an obligation to disclose information to supervisory bodies or also enhanced transparency obligations and access rights for individuals affected.

41

In addition, the introduction of licensing procedures may be justified for applications that fall under Level 3, which are associated with **regular** or **significant potential for harm**. Applications that fall under Level 4 are associated with **serious potential for harm**; the Data Ethics Commission believes that these applications should be subject to enhanced oversight and transparency obligations. These may extend all the way through to the publication of information on the factors that influence the algorithmic calculations and their relative weightings, the pool of data used and the algorithmic decision-making model; an option for "always-on" regulatory oversight via a live interface with the system may also be required.

42

Finally, a complete or partial ban should be imposed on **applications with an untenable potential for harm** (Level 5).

43

The Data Ethics Commission believes that the measures it has proposed should be implemented in a new EU Regulation on algorithmic systems enshrining general **horizontal requirements (Regulation on Algorithmic Systems, EU-ASR)**. This horizontal regulation should incorporate the fundamental requirements for algorithmic sytems that the Data Ethics Commission developed. In particular, it should group together general substantive rules – informed by the concept of system criticality – on the admissibility and design of algorithmic systems, transparency, the rights of individuals affected, organisational and technical safeguards and supervisory institutions and structures. This horizontal instrument should be fleshed out in **sectoral instruments** at EU and Member State level, with the concept of system criticality once again serving as a guiding framework.

44

The process of drafting the EU-ASR (as recommended above) should incorporate a debate on how best to demarcate the respective scopes of this Regulation and the **GDPR**. A number of factors should be taken into account in this respect; firstly, algorithmic systems may pose specific risks to individuals and groups even if they do not involve the processing of personal data, and these risks may relate to assets, ownership, bodily integrity or discrimination. Secondly, the regulatory framework introduced for the future horizontal regulation of algorithmic systems may need to be more flexible and risk-adapted than the current data protection regime.

Instruments

45

The Data Ethics Commission recommends the introduction of a **mandatory labelling scheme** for algorithmic systems of enhanced criticality (Level 2 upwards). A mandatory scheme of this kind would oblige operators to make it clear whether (i.e. when and to what extent) algorithmic systems are being used. Regardless of system criticality, operators should always be obliged to comply with a mandatory labelling scheme if there is a risk of confusion between human and machine that might prove problematic from an ethical point of view.

46

An individual affected by a decision should be able to exercise his or her right to "meaningful **information** about the logic involved, as well as the scope and intended consequences" of an algorithmic system (cf. GDPR) not only in respect of fully automated systems, but also in situations that involve any kind of **profiling**, regardless of whether a decision is taken on this basis later down the line. The right should also be expanded in the future to apply to the algorithm-based decisions themselves, with differing levels of access to these decisions according to system criticality. These measures may require the clarification of certain legislative provisions or a widening of regulatory scope at European level.

47

It certain cases, it may be appropriate to ask the operator of an algorithmic system to provide an **individual explanation** of the decision taken, in addition to a general explanation of the logic (procedure) and scope of the system. The main objective should be to provide individuals who are affected by a decision with comprehensible, relevant and concrete information. The Data Ethics Commission therefore welcomes the work being carried out under the banner of "Explainable AI" (efforts to improve the explainability of algorithmic systems, in particular self-learning systems), and recommends that the Federal Government should fund further research and development in this area.

48

In view of the fact that, in certain sectors, society as a whole may be affected as well as its individual members, also particular **parties who are not individually affected** by an algorithmic system should be entitled to access certain types of information about it. It is likely that rights of this kind would be granted primarily for journalistic and research purposes; in order to take due account of the operator's interests, they would need to be accompanied by adequate protective measures. The Data Ethics Commission believes that consideration should also be given to the granting of unconditional rights to access information in certain circumstances, in particular when algorithmic systems with serious potential for harm (Level 4) are used by the State.

49

It is appropriate and reasonable to impose a legal requirement for the operators of algorithmic systems with at least some potential for harm (Level 2 upwards) to produce and publish a proper **risk assessment**; an assessment of this kind should also cover the processing of non-personal data, as well as risks that do not fall under the heading of data protection. In particular, it should appraise the risks posed in respect of self-determination, privacy, bodily integrity, personal integrity, assets, ownership and discrimination. It should encompass not only the underlying data and logic of the model, but also methods for gauging the quality and fairness of the data and the model accuracy, for example the bias or the rates of (statistical) error (overall or for certain sub-groups) exhibited by a system during forecasting/category formation.

50

To provide controllers and processors with greater legal clarity, further work must be done in terms of fleshing out the requirements to **document and log** the data sets and models used, the level of granularity, the retention periods and the intended purposes. In addition, operators of sensitive applications should be obliged in future to document and log the program runs of software that may cause lasting harm. The data sets and models used should be described in such a way that they are comprehensible to the employees of supervisory institutions carrying out oversight measures (as regards the origin of the data sets or the way in which they are pre-processed, for example, or the optimisation goals pursued using the models).

51

System operators should be required by the standard-setting body to guarantee a minimum level of **quality, from both a technical and a mathematical-procedural perspective**. The procedural criteria imposed must ensure that algorithmically derived results are obtained in a correct and lawful manner. For this purpose, quality criteria could be imposed, in particular as regards corrective and control mechanisms, data quality and system security. For example, it would be appropriate to impose quality criteria on the relationship between algorithmic data processing outcomes and the data used to obtain these outcomes.

52

The Data Ethics Commission believes that a necessary first step is to clarify and flesh out in greater detail the scope and legal consequences of Article 22 GDPR in relation to the use of algorithmic systems in the context of human decision-making. As a second step, the Data Ethics Commission recommends the introduction of additional **protective mechanisms for algorithm-based and algorithm-driven decision-making systems**, since the influence of these systems in real-life settings may be almost as significant as that of algorithm-determined applications. The prohibitory principle followed to date by Article 22 GDPR should be replaced by a more flexible and risk-adapted regulatory framework that provides adequate guarantees as regards the protection of individuals (in particular where profiling is concerned) and options for these individuals to take action if mistakes are made or if their rights are jeopardised.

53

Consideration should be given to expanding the **scope of anti-discrimination legislation** to cover specific situations in which an individual is discriminated against on the basis of automated data analysis or an automated decision-making procedure. In addition, the legislator should take effective steps to prevent **discrimination on the basis of group characteristics** which do not in themselves qualify as protected characteristics under law, and where the discrimination often does not currently qualify as indirect discrimination on the basis of a protected characteristic.

54

In the case of algorithmic systems with regular or significant (Level 3) or even serious potential for harm (Level 4), it would be useful – as a supplement to the existing regulations – for these systems to be covered by **licensing procedures or preliminary checks** carried out by supervisory institutions, in the interests of preventing harm to individuals who are affected, certain sections of the population or society as a whole.

Institutions

55

The Data Ethics Commission recommends that the Federal Government should expand and realign the competencies of existing supervisory institutions and structures and, where necessary, set up new ones. Official supervisory tasks and powers should primarily be entrusted to the **sectoral supervisory authorities** that have already built up a wealth of expert knowledge in the relevant sector. Ensuring that the competent authorities have the financial, human and technical **resources** they need is a particularly important factor in this respect.

56

The Data Ethics Commission also recommends that the Federal Government should set up a **national centre of competence for algorithmic systems**; this centre should act as a repository of technical and regulatory expertise and assist the sectoral supervisory authorities in their task of monitoring algorithmic systems to ensure compliance with the law.

57

The Data Ethics Commission believes that initiatives involving the development of technical and statistical **quality standards for test procedures and audits** (differentiated according to critical application areas if necessary) are worthy of support. Test procedures of this kind – provided that they are designed to be adequately meaningful, reliable and secure – may make a vital contribution to the future auditability of algorithmic systems.

58

In the opinion of the Data Ethics Commission, particular attention should be paid to innovative forms of **co-regulation and self-regulation**, alongside and as a complement to forms of state regulation. It recommends that the Federal Government should examine various models of co-regulation and self-regulation as a potentially useful solution in certain situations.

59

The Data Ethics Commission believes that an option worth considering might be to require operators by law (inspired by the "comply or explain" regulatory model) to sign a declaration confirming their willingness to comply with an **Algorithmic Accountability Code**. An independent commission with equal representation – which must be free of state influence – could be set up to develop a code of this kind, which would apply on a binding basis to the operators of algorithmic systems. Appropriate involvement of civil society representatives in the drafting of this code must be guaranteed.

60

Voluntary or mandatory evidence of protective measures in the form of a specific **quality seal** may also serve as a guarantee to consumers that the algorithmic system in question is reliable, while at the same time providing an incentive for developers and operators to develop and use reliable systems.

61

The Data Ethics Commission takes the view that companies and authorities operating critical algorithmic systems should be obliged in future to appoint a **contact person**, in the same way that companies of a specific size are currently obliged to appoint a data protection officer. Communications with the authorities should be routed through this contact person, and he or she should also be subject to a duty of cooperation.

62

To ensure that official audits of algorithmic systems take due account of the interests of civil society and any companies affected, suitable **advisory boards should be set up within the sectoral supervisory authorities**.

63

In the opinion of the Data Ethics Commission, technical standards adopted by **accredited standardisation organisations** are a generally useful measure, occupying an intermediate position between state regulation and purely private self-regulation. It therefore recommends that the Federal Government should engage in appropriate efforts towards the development and adoption of such standards.

64

The system of granting **competitors, competition associations or consumer associations the right to file an action** has been an important feature of the German legal landscape for many years, and could play a key role in civil society oversight of the use of algorithmic systems. In particular, private rights of this kind could allow civil

society players with a legitimate mandate to enforce compliance with legal provisions in the area of contract law, fair trading law or anti-discrimination law, without needing to rely on the authorities to take action and without needing to wait for individuals to authorise them.

Special topic: Algorithmic systems used by media intermediaries

65

Given the specific risks posed by media intermediaries that act as **gatekeepers to democracy**, the Data Ethics Commission recommends that options should be examined for countering these risks, also with regard to influencing EU legislation (→ see Recommendation 43 above). A whole gamut of risk mitigation measures should be considered, extending through to ex-ante controls (e.g. in the form of a licensing procedure).

66

The national legislator is under a constitutional obligation to protect the democratic system from the dangers to the free, democratic and pluralistic formation of opinions that may be created by providers that act as gatekeepers by establishing a binding normative framework for **media**. The Data Ethics Commission believes that the small number of operators concerned should be obliged to use algorithmic systems that allow users (at least as an additional option) to access an unbiased and balanced selection of posts and information that embodies pluralism of opinion.

67

The Federal Government should consider measures that take due account of the risks typically encountered in the media sector in respect of all media intermediaries and also in respect of providers that do not act as gatekeepers or whose systems are associated with a lower potential for harm. These measures might include mechanisms for **enhancing transparency** (for example by ensuring that information is available about the technical procedures used to select and rank news stories, **introducing labelling obligations for social bots**) and establishing a right to post countering responses on timelines.

Use of algorithmic systems by state bodies

68

The State must, in the interests of its citizens, make use of the best available technologies, including algorithmic systems, but must also exercise particular prudence in all of its actions in view of its obligation to preserve fundamental rights and act as a role model. As a general rule, therefore, the use of algorithmic systems by public authorities should be assessed on the basis of the criticality model as **particularly sensitive**, entailing at the very least a comprehensive risk assessment.

69

In the areas of **law-making** and the **dispensation of justice**, algorithmic systems may at most be used for peripheral tasks. In particular, algorithmic systems must not be used to undermine the functional independence of the courts or the democratic process. By way of contrast, enormous potential exists for the use of algorithmic systems in connection with **administrative** tasks, in particular those relating to the provision of services and benefits. The legislator should take due account of this fact by giving the green light to a greater number of partially and fully automated administrative procedures. Cautious consideration should therefore be given to expanding the scope of both Section 35a of the German Administrative Procedures Act (*Verwaltungsverfahrensgesetz*, VwVfG) (which is couched in overly restrictive terms) and the corresponding provisions of statutory law. All of these measures must be accompanied by adequate steps to protect citizens.

70

Decisions taken by the State on the basis of algorithmic systems must still be **transparent**, and it must still be possible to provide **justifications** for them. It may be necessary to clarify or expand the existing legislation on freedom of information and transparency in order to achieve these goals. Furthermore, the use of algorithmic systems does not negate the principle that decisions made by public authorities must generally be justified individually; on the contrary, this principle may impose limits on the use of overly complex algorithmic systems. Finally, greater priority should be accorded to open-source solutions, since the latter may significantly enhance the transparency of government actions.

71

From an ethical point of view, there is no general right to non-compliance with rules and regulations. At the same time, however, automated "total" enforcement of the law raises a number of different ethical concerns. As a general rule, therefore, systems should be designed in such a way that a human can override **technical enforcement** in a specific case. The balance struck between the potential transgression and the automated (and perhaps preventive) enforcement measure must at all times meet the requirements of the proportionality principle.

Liability for algorithmic systems

72

Liability for damages, alongside criminal responsibility and administrative sanctions, is a vital component of any ethically sound regulatory framework for algorithmic systems. It is already apparent today that algorithmic systems pose challenges to liability law as it currently stands, *inter alia* because of the complexity and dynamism of these systems and their growing "autonomy". The Data Ethics Commission therefore recommends that the current provisions of liability law should undergo in-depth checks and (where necessary) revisions. The scope of these checks and revisions should not be restricted on the basis of too narrowly defined technological features, such as machine learning or artificial intelligence.

73

The proposal for a future system under which legal personality would be granted to high-autonomy algorithmic systems, and the systems themselves would be liable for damages (**"electronic person"**), should **not be pursued further**. As far as this concept is, by some protagonists, based on a purported equivalence between human and machine it is ethically indefensible. And as far as it boils down to introducing a new type of company under company law it does not, in fact, solve any of the pertinent problems.

74

By way of contrast, if harm is caused by autonomous technology used in a way functionally equivalent to the employment of human auxiliaries, the operator's liability for making use of the technology should correspond to the otherwise existing vicarious **liability regime of a principal for such auxiliaries** (cf. in particular Section 278 of the German Civil Code). For example, a bank that uses an autonomous system to check the creditworthiness of its customers should be liable towards them to at least the same extent that it would be had it used a human employee to perform this task.

75

As the debate currently stands, it appears highly likely that appropriate amendments will need to be made to the **Product Liability Directive** (which dates back to the 1980s), and a connection established to new product safety standards; in addition, certain changes may need to be made to the rules relating to **fault-based liability** and/or new bases of **strict liability** may need to be introduced. In each case, it will be necessary to determine the liability regime that is most appropriate for particular types of products, digital content and digital services, and the exact shape that this regime should take (once again depending on the criticality of the relevant algorithmic system). Consideration should also be given to innovative liability concepts currently being developed at European level.

4

A European path

The Data Ethics Commission examined a great many different questions in the course of its work, and discussions on these questions have raised new ones in turn; this alone should serve to indicate that this Opinion can serve only as one out of many building blocks in the larger edifice of a **debate on ethics, law and technology** that will continue for many years to come. The Data Ethics Commission takes the view that it is important to remember that ethics, law and democracy must serve as a shaping force for change, both in the broader sense and more specifically in the field of technology. To achieve this goal, interdisciplinary discourse in politics and society is required, and care must be taken to ensure that any rules and regulations adopted are open enough to retain their regulatory clout and their ability to adapt, even in the face of fast-paced changes to technologies and business models. These rules and regulations must be enforced effectively by means of appropriate instruments, procedures and structures, and these latter must make it possible to intervene promptly in response to infringements or undesirable developments.

In the global contest for future technologies, Germany and Europe are being confronted with value systems, models of society and cultures that differ widely from our own. The Data Ethics Commission supports the **"European path"** that has been followed to date: the defining feature of European technologies should be their consistent alignment with European values and fundamental rights, in particular those enshrined in the European Union's Charter of Fundamental Rights and the Council of Europe's Convention for the Protection of Human Rights and Fundamental Freedoms.

The Data Ethics Commission believes that the State has a particular responsibility to develop and enforce ethical benchmarks for the digital sphere that reflect this value system. In order to deliver on this promise to citizens, it must act from a position of political and economic strength on the global stage; excessive dependence on others turns a nation into a rule taker rather than a rule maker, resulting in the citizens of this nation being subject to resulting in the citizens of this nation being subject to requirements imposed by players elsewhere in the world, or by private corporations that are, for the most part, exempt from democratic legitimacy and oversight. Embarking on **efforts to safeguard the digital sovereignty of Germany and Europe in the long term** is therefore not only a politically far-sighted necessity, but also an expression of ethical responsibility.

Part A

Introduction

Guiding motifs

Our society is experiencing profound changes brought about by digitalisation. Innovative data-based technologies may benefit us at both the individual and the wider societal levels, as well as potentially boosting economic productivity, promoting sustainability and catalysing huge strides forward in terms of scientific progress. At the same time, however, digitalisation poses risks to our fundamental rights and freedoms. It raises a wide range of ethical and legal questions centring around two wider issues: the role we want these new technologies to play, and their design. If we want to ensure that digital transformation serves the good of society as a whole, both society itself and its elected political representatives must engage in a debate on how to use and shape data-based technologies, including artificial intelligence (AI).

Germany's Federal Government set up the Data Ethics Commission *(Datenethikkommission)* on 18 July 2018. It was given a one-year mandate to develop ethical benchmarks and guidelines as well as specific recommendations for action, aiming at protecting the individual, preserving social cohesion, and safeguarding and promoting prosperity in the information age. As a starting point, the Federal Government presented the Data Ethics Commission with a number of key questions clustered around three main topics: algorithm-based decision-making (ADM), AI and data. In the opinion of the Data Ethics Commission, however, AI is merely one among many possible variants of an algorithmic system, and has much in common with other such systems in terms of the ethical and legal questions it raises. With this in mind, the Data Ethics Commission has structured its work under two different headings: **data** and **algorithmic systems** (in the broader sense).

In preparing its Opinion, the Data Ethics Commission was inspired by the following **guiding motifs**:

- Ensuring the human-centred and value-oriented design of technology
- Fostering digital skills and critical reflection in the digital world
- Enhancing protection for individual freedom, self-determination and integrity
- Fostering responsible data utilisation that is compatible with the public good
- Introducing risk-adapted regulation and effective oversight of algorithmic systems
- Safeguarding and promoting democracy and social cohesion
- Aligning digital strategies with sustainability goals
- Strengthening the digital sovereignty of both Germany and Europe.

1. Mission and basic understanding

Our society is experiencing profound changes brought about by digitalisation. Innovative data-based technologies may benefit us at both the individual and the wider societal levels, as well as potentially boosting economic productivity, promoting sustainability and catalysing huge strides forward in terms of scientific progress; in some cases, this has already happened. The digital transformation offers tremendous opportunities for all countries, in particular for Germany as a closely networked and high-tech economy, but it means that German companies are coming under increasing competitive pressure on the international market. At the same time, it is already becoming apparent that digitalisation poses risks to our fundamental rights and freedoms. It raises a wide range of ethical and legal questions centring around two wider issues: the role we want these new technologies to play, and their design. If we want to ensure that digital transformation serves the good of individuals and society as a whole, both society itself and its elected political representatives must engage in a debate on how to shape the design of data-based technologies, including AI.

On 18 July 2018, the Federal Government set up the Data Ethics Commission (*Datenethikkommission*) and named its 16 members (→ see Annex, 2). Christiane Wendehorst and Christiane Woopen were appointed as co-spokespersons. The Data Ethics Commission was given a one-year mandate to develop ethical benchmarks and guidelines, aiming at protecting the individual, preserving social cohesion, and safeguarding and promoting prosperity in the information age. It was also asked to put forward specific recommendations for action and suggestions for possible legislation with a view to allowing these ethical guidelines to be observed, implemented and supervised. As a starting point, the Federal Government presented the Data Ethics Commission with a number of key questions (→ see Annex 1) clustered around three main topics: (I) algorithm-based decision-making, (II) AI and (III) data.

In this context, "**AI**" is understood by the Data Ethics Commission to be a catch-all term for technologies and related applications based on digital methods which involve the machine processing of potentially very large and heterogeneous data sets in a complex procedure that mimics human intelligence; the results obtained from such a procedure may be applied in an automated way. Some of the most important methods underpinning AI (as just one aspect of a much wider computer science landscape) include sub-symbolic pattern recognition, machine learning, computer-based knowledge representation and knowledge engineering, which in turn encompasses heuristic search methods, inference techniques and action planning.

The Data Ethics Commission however believes that it would be wrong to restrict the ethical and legal debate to AI alone. It is merely one among many possible variants of an algorithmic system, and thus represents a subset of this field. Both AI systems and other types of algorithmic systems share a number of features that may give rise to ethical problems, meaning that regulations focused on AI alone would tackle only part of the problem. The feature of self-learning, which is in the foreground in AI, brings with it specific challenges, and due consideration must be given to them at the risk assessment stage; at the same time, however, there are many other features besides self-learning that require special attention. The following arguments therefore relate to **algorithmic systems of all kinds**.

Applications are rarely based on a single algorithm, and examining algorithms in isolation is rarely meaningful. Any ethical appraisal must be based on the **sociotechnical system as a whole**, or in other words all the components of an algorithmic application, including all the people involved, from the development phase – for example the training data used – right through to its implementation in an application environment and any evaluation and adjustment measures.

2. Working method

Between September 2018 and September 2019, the Data Ethics Commission met on a monthly basis. It discussed examples of use cases for new technologies in a range of different sectors, and analysed them in terms of both the technology involved and the ethical and legal issues raised. The findings obtained from this work and from fundamental debates made it possible to identify overarching topics and questions, which were used as a starting point for the development of an ethical appraisal framework and the drafting of specific recommendations for future political and legislative action. As early as October 2018, in response to a policy paper by the Federal Government, the Data Ethics Commission put forward two specific recommendations for points that should be included in the Artificial Intelligence Strategy, and these recommendations were taken up by the Federal Government. In November 2018, the Data Ethics Commission issued another recommendation, calling for the roll-out of an electronic health record, building on a participatory process.[1]

The Data Ethics Commission involved the public in two public conferences. The first took place on 7 February 2019 at the Federal Ministry of Justice and Consumer Protection (*Bundesministerium der Justiz und für Verbraucherschutz*), and centred around the issue of "Self-determination and external determination in the age of artificial intelligence". The second – an international round table under the title "Towards ethical shaping of our digital future" – was held on 9 May 2019 at the Federal Ministry of the Interior, Building and Community (*Bundesministerium des Innern, für Bau und Heimat*). Both events allowed the Data Ethics Commission to engage in in-depth discussions with experts and stakeholders as well as members of the public and interested citizens.[2]

On 14 November 2018, during the Federal Government's *Digitalklausur*, an exchange of views took place between the Federal Chancellor, all the members of the Federal Government, and the two co-spokespersons of the Data Ethics Commission. Ad-hoc discussions were also held with individual members of the Federal Government. In addition, the Data Ethics Commission organised expert hearings and consultation meetings with other institutions and bodies working on related topics, including the Study Commission "Artificial Intelligence", the Commission of Experts on Competition Law 4.0, the Federal Government's Digital Council, the Advisory Council for Consumer Affairs and many more.

One of the defining features of the Data Ethics Commission is that its work and advisory activities are fully independent and free from any external political influence. All of the viewpoints outlined in this report reflect either the personal opinions expressed by the Data Ethics Commission's individual members, or the opinions that emerged from internal discussions within its institutional members. The Data Ethics Commission has adopted all of the recommendations in this report by consensus.

1 Both documents are available on the Data Ethics Commission's website (at *www.datenethikkommission.de*).
2 Further information on the public conferences, including video recordings, can be found on the Data Ethics Commission's website (at *www.datenethikkommission.de*).

3. Objectives and scope of the report

The goal pursued by the Data Ethics Commission in publishing this report is to further the development of our **ethical and legal framework** in order to confront the challenges posed by digital technologies. The main concern is to ensure that the fundamental conditions are in place for the free democratic basic order to be preserved, and for the potential that exists to be leveraged so that sustainability-oriented goals can be achieved and our social market economy can flourish.

Given the increase in the volume of personal data being collected and the use of automated methods to process these data for different purposes, one of the main priorities of the Data Ethics Commission is to reconcile the need to protect the **individual's fundamental rights and freedoms** – including self-determination and integrity – with the need to promote progress, prosperity, the safeguarding of democracy and the shaping of a society that is fit for the future. Protecting individuals against data misuse and discrimination and guaranteeing the security of all parties involved are tasks that fall squarely within the remit of a State governed by the rule of law, and effective regulations must be adopted and institutions set up for this purpose. At the same time, however, the State must facilitate the emergence of innovative business models that safeguard future prosperity for everyone.

The Data Ethics Commission believes that digitalisation – in particular the rapidly increasing availability of data and the use of complex algorithmic systems, including AI – holds **enormous potential** for technical and social innovation and for achievement of the UN's Sustainable Development Goals. Promising avenues for action include promoting health, humanising the world of work, designing sustainable cities and communities, providing a decent education and implementing effective climate protection measures. At the same time, however, we must not forget the **major risks** that may face individuals, society as a whole and the free democratic basic order in connection with the extensive use of digital technologies. These risks include the possibility of high-granularity profiling (using techniques such as online tracking, voice analysis during remote job interviews, or even the diagnosis of pathological mental conditions on the basis of social media posts), the potential for these profiles to be exploited for the purpose of controlling and manipulating people (either on a small scale through individual pricing or on a larger scale by manipulating democratic opinion-building processes through "micro-targeting"), the potential for discrimination against different social groups, and the ability to delegate human responsibility to machines. With these factors in mind, the Data Ethics Commission believes that we must actively shape our future in such a way as to realise the potentials while avoiding the risks.

The Data Ethics Commission advocates for a multi-step approach to achieving these goals. The first step is an ethical reflection on the value of human activity in an environment shaped by technology, and a reaffirmation of the **key ethical principles and precepts** upon which our society is founded (→ Part B). In the view of the Data Ethics Commission, the key questions can be divided into questions that concentrate mainly on data (the "data perspective") and questions that are primarily focused on algorithmic systems (the "algorithms perspective"). These two perspectives represent ethical discourses which both complement each other and are contingent upon each other, and which are also each reflected in different **governance instruments** (→ Part D).

In the section devoted to the data perspective (→ Part E), the Data Ethics Commission outlines general ethical principles for **data governance** (→ E 1), in particular ethical principles governing **data rights and data obligations** (→ E 2); these serve as the basis for a series of specific recommendations for action regarding the use of data and data access (→ E 3 to 5). In the section devoted to the algorithms perspective (→ Part F), the Data Ethics Commission sets out general ethical requirements for the **design of algorithmic systems** (→ F 2) and the **risk-adapted regulation** of these systems (→ F 3). The instruments and institutions that would be required to implement regulations of this kind are examined in detail and summarised in recommendations to the legislator (→ F 4 to 8). A shared basic understanding of technical parameters and relationships (→ Part C) serves as an essential foundation for considerations of this kind. The report ends with a plea for the Federal Government to follow a "European path" (→ Part G).

As per its mission, the Data Ethics Commission's recommendations are targeted primarily at the German **Federal Government** and its associated institutions. At certain points, however, the target audience is widened to include other stakeholders, for example Länder and municipalities, research institutions or enterprises. The Federal Government is always the secondary target audience of any such recommendations, given the underlying recommendation for it to encourage and support these other stakeholders in their efforts. All of the recommendations should also be viewed in the context of the institutions and rules that have been or will be put in place at EU and international level, and in the context of further developments in these arenas. In cases where the Data Ethics Commission suggests that a recommendation should be implemented at **EU or international level**, it should be interpreted as a recommendation to the German Federal Government to make a vigorous and future-oriented contribution to the debate taking place within Europe and across the globe.

Part B

Ethical and legal principles

1. The fundamental value of human agency

Given the fast-paced development of digital technologies, including self-learning algorithmic systems ("artificial intelligence") which incorporate certain functions that can outperform the abilities of humans, the elementary question is raised **whether human agency poses an ethically relevant value in and of itself** which transcends considerations of effectiveness and efficiency, and which is inherently preferable to the functioning of machine systems. This question is all the more pressing as the momentum and internal logic of international competition are, for the most part, dictated solely by the goal of maximising economic efficiency.

Human agency derives its basic value from its moral significance. A human being can provide reasons for one's actions and decide whether or not to perform them, and must bear responsibility for these actions. It is only by taking action that individuals can develop and realise their full potential in accordance with their capabilities, preferences and understanding of a meaningful life. This **dimension of meaning** lends a value to human activity that could never be claimed for the functioning of technical systems. Technology can only ever be the means to achieving a goal that humans have set. Even if – hypothetically speaking – humans were to decide that algorithmic systems could set themselves goals, allowing them to do so would be a goal that had been set by humans. The use of technical systems may therefore be a component of human activity, and may even be ethically required in certain cases, but it will never be possible for technical systems to replace the moral dimension of human agency completely. Human agency and the human drive to develop as a living being are characterised by their multi-dimensional nature. Although the conceptions of man espoused by different cultures and different faiths vary significantly, they all incorporate the dimension of the living and of moral responsibility, and despite all the differences in the respective answers, they all embrace the question of the meaning of life whereas technical systems merely function.

We must weigh up many different criteria when identifying cases in which preference should be given to human activity over the use of algorithmic systems. As a basic principle, a higher level of effectivity should be prioritised only with regard to the performance of certain limited functions. **Effectiveness should not rule supreme.** It must not place material restrictions on the ability of humans to take action as a form of self-development, and it must take second place to the basic ethical dimension of a meaningful and flourishing life, both as an individual and as a member of society. For example, even if it were possible for a human to be cared for more effectively by a robot than by another human, care by a robot cannot be allowed to replace the human element of attention and affection for the person needing that care. At the same time, however, the use of robots to perform care-related tasks alongside humans may be deemed expedient if it makes the situation significantly safer for the person receiving care. Yet the effectiveness gains of technical systems must take a back seat if they entail an intrusion into the privacy or personal integrity of the individual, for example because they force an employee to modify all of his or her work processes in order to maximise effectiveness. People must be allowed to retain their subjectivity rather than morphing into objects that are "acted upon" by machines.

Humans are morally responsible for their actions, and there is no escaping this moral dimension. Humans are responsible for the goals they pursue, the means by which they pursue them, and their reasons for doing so. This dimension must always be taken into account when designing our technologically shaped future. At the same time, the notion that technology should serve humans rather than humans being subservient to technology can be taken as incontrovertible fact. Germany's constitutional system is founded on this **understanding of the human being**, and it adheres to the tradition of Europe's cultural and intellectual history.

2. Relationship between ethics and law

Exponential technical developments relating to the collection and use of digital data and the deployment of algorithmic systems and artificial intelligence are increasingly shaping the life of every individual and all aspects of our social coexistence. These developments give rise to far-reaching and profound questions, and the answers to these questions must be guided by the **fundamental legal and ethical principles** that a democratic society undertakes to uphold.

The benchmarks and guiding principles underpinning the processes by which society shapes and has to shape various sectors – the economy, education, public spaces, healthcare, finance, transport and energy – are fundamentally ethical in nature. Although liberal systems are characterised by a high degree of moral pluralism, a common ethical framework is nevertheless established in constitutional law, and more especially in fundamental rights as far as the relationship between the State and the individual is concerned. The significance of this ethical and legal framework in relation to an individual case and in the event of conflict between differing values or fundamental rights is not always clear-cut. Yet this does not relativise the binding nature and **fundamental importance of the ethical foundation of our community**. Instead, it merely goes to prove once again the crucial importance of an open and ongoing debate on the future shape of our society, and serves as a basis for democratic decision-making processes that acknowledge the possibility of different answers within the framework of the Constitution.

Ethics cannot be equated on a one-to-one basis with the law. In other words, not everything that is relevant from an ethical perspective can and should be enshrined in legislation; conversely, there are provisions of the law that are motivated purely by pragmatic considerations and are not ethically imperative. Nevertheless, legislation must, at all times, be heedful of its potential ethical implications and must live up to ethical standards – at the very least, the requirements outlined in constitutional law.

The Data Ethics Commission holds the view that regulation is necessary, and cannot be replaced by ethical principles and guidelines in cases where the constitutionally developed **principle of materiality** requires the enactment, in the form of parliamentary legislation, of democratically legitimate rules that can be enforced against anyone. Internet governance is also the governance of society. As algorithmic systems, including artificial intelligence, become an increasingly normal feature of the daily lives we lead together in society, we must also develop and enforce rules to govern them. This calls for an ongoing public debate, and also – particularly in cases where fundamental rights are at threat – parliamentary debate and legislative initiatives. Given past experiences of law enforcement in the Internet sphere, and in view of the experience that power tends to be accumulated in the hands of a few large corporations in certain sectors of markets dominated by digital technologies, a systematic move away from enforceable rules and towards voluntary regulation would appear to be a mistake.

At the same time, regulation must not unduly inhibit technological and social innovation and dynamic market growth. Overly rigid laws that attempt to regulate every last detail of a situation may place a stranglehold on progress and increase red tape to such an extent that innovative processes in Germany can no longer keep pace with the rate of technological development on the international stage. On the other hand, regulatory frameworks can and must protect fundamental rights and freedoms and create legal certainty. This is an essential first stage in building a system within which citizens, companies and institutions can trust in the fact that the transformation of society will be guided by ethical principles. In addition, the “toolbox-like” nature of the legal system with its options for regulating matters at many different levels, ranging from acts and ordinances right down to codes, self-governance options and voluntary obligations, makes it suitable for creating a framework that is adaptable and can keep up with technological progress.

However, the **need for guidance goes far beyond the regulatory sphere**. With this in mind, many different stakeholders – such as professional groups, companies and advisory boards at national, regional and international level – have responded to the manifold upheavals by drafting ethical codes or sets of guiding ethical principles, in some cases with an ensuing public debate.

The Data Ethics Commission welcomes the diversity of stakeholders taking action and the number of voices being heard in the discussion on how the process of digitalisation can be shaped in an ethical way, since this highlights the indispensability of public debate and **for everyone to take responsibility for the flourishing of our future lives together**. In keeping with the mission assigned to it in the coalition agreement, the Data Ethics Commission has based its recommendations for a "framework on how to develop data policy and deal with algorithms, artificial intelligence and digital innovations" not only on the precepts of constitutional law, but also on cross-cutting ethical principles that apply to differing degrees in all areas of society; these principles are briefly outlined below.[1]

1 By following this approach, the Data Ethics Commission is adhering to the same basic principles endorsed by the European Group on Ethics in Science and New Technologies (EGE) in its opinion: EGE: Statement on Artificial Intelligence, Robotics and "Autonomous" Systems, 2018, (available at: *http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf*).

3. General ethical and legal principles

3.1 Human dignity

Human dignity, which from an ethical viewpoint is synonymous with the unconditional value of every human being and which is enshrined as a "fundamental constitutional principle" in the constitutional order, is of foundational and supreme importance. It follows from the principle of human dignity that every individual merits respect, regardless of his or her attributes and achievements. Protecting the value which is inherent in every human being and which does not need to be acquired also implies that human beings are not ranked in a classifying system across various spheres of life and activities ("super scoring") or labelled like an object with a price and treated accordingly. The fact that each human is an individual rather than a pattern made up of data points must also be borne in mind at all times in situations where human behaviour is measured and these measurements are processed by algorithmic systems. Algorithmic systems must therefore always be designed in such a way that they can cater to each human's claim to individuality.

Acknowledging human dignity involves recognising that humans must always be "superior to technology", i.e. that they must not be completely or irrevocably subordinated to technical systems. The opportunities for configuration and intervention may be localised at different levels in each specific application, but the **principle of human sovereignty of action** must be upheld. Humans hold responsibility in human/machine interactions, and must not be regarded as defective beings that need to be optimised or perfected by the machine. Instead, human use of algorithmic systems aims at realising human ideas and objectives more effectively and rapidly and with fewer errors.

Protecting human dignity also involves ensuring that the **human as a relational being** is not misled by technology about the nature of a relationship; for example, it would be wrong for a human to be systematically deceived into thinking that he or she is speaking with another human when it is actually a bot. The **psychological integrity of the individual** is a particularly important factor in protecting human dignity. This rules out the use of data-driven systems for manipulative purposes, particularly when the systems draw on comprehensive and highly granular personality profiles. It also rules out the use of algorithmic systems to **discriminate systematically against individuals or groups**, for example by "downgrading" them, preventing them from using certain services for ethically untenable reasons or systematically misleading them as they participate in the democratic discourse.

3.2 Self-determination

The opportunity for self-determination is inextricably linked with human dignity. Humans **express their freedom** by determining their life goals and the way they lead these lives, as a basis for determining, developing and enacting the very essence of their self. A society that takes freedom seriously must put in place a framework within which its citizens can develop freely and respect each other's freedom, despite all their differences. For example, if people are to lead a self-determined life and develop in freedom, technical systems must not restrict and control human avenues for action without an ethically meaningful reason. Self-determination must not be viewed solely through an individualistic lens – humans are relational beings whose life unfolds through social interactions with others, on the basis of manifold reciprocal links and influences.

The rules that govern these interactions are shaped over time by the **cultural and socionormative framework** that serves as a basis for our life together in society. They are also shaped by law in a democratic society, especially where imbalances of power and information prevail.

The more information third parties have collected about an individual, the more difficult it becomes for that individual to act unselfconsciously in social situations or even to reinvent himself or herself completely. Steps must be taken to ensure that data collection and evaluation practices do not result in personal and social profiles being routinely created in multiple locations, thereby "cementing" a particular version of the individual. Self-determination therefore also encompasses the **right to develop and alter one's own identity**, and the possibility of starting one's life afresh. The right to self-determination thus also includes each individual's right to decide how he or she is perceived in public and to prevent public misrepresentations.

Another vital aspect of self-determination is that people must not only be allowed to assume **responsibility**, but must do so and do justice to the task. Responsibility always lies with a human – institutionally enshrined, if necessary – never with a machine. Even if a technical system is used to apply inferences based on automated evaluations (i.e. whether or not a loan should be granted), the responsibility for developing and using this system in an ethically sound manner must lie with humans.

An important manifestation of self-determination is **informational self-determination**. It includes the individual's right to determine who can collect and use which personal data, when they may do so and for what purpose. Informational self-determination allows an individual to protect his or her freedom of action and privacy to the extent he or she deems important, and also to determine as what personality he or she wants to be perceived and treated in public.

In this era of digitalisation, the special importance of individuals as self-determined actors in the data society goes beyond their informational self-determination. The term **digital self-determination** refers to this; it encompasses the skills needed by an individual to determine for himself or herself the content that should be used as a basis for interacting with his or her environment, and how he or she can unfold his or her own personality in an interactive way. Under certain circumstances, it may also include the self-determined economic exploitation of an individual's own data assets and the self-determined governance of non-personal data, for example the data generated when operating certain devices. Digital self-determination always goes hand in hand with digital accountability.

The Data Ethics Commission takes the view that **businesses and legal persons** should also be entitled to a right to digital self-determination. Legal persons cannot invoke the concept of human dignity granted by Article 1 paragraph 1 of the [German] Basic Law (*Grundgesetz*, GG) and protected in the framework of the general right of personality, and are therefore barred from referring to the associated core area of personality development, which enjoys absolute protection. Article 2 paragraph 1 of the Basic Law, in conjunction with Article 19 paragraph 3 of the Basic Law, does however grant legal persons a protected right of personality that also incorporates a right to informational self-determination.

The ability of consumers to take self-determined action and conscious consumption decisions is a vital prerequisite for optimum resource allocation and maximisation of the public good at macroeconomic level. Any erosion of the **skills needed by consumers** to exercise their right of self-determination, for example because of the excessive use of decision-making assistants and the associated habituation effects, raises ethical questions regarding external determination and the freedom of individuals to take decisions, and also regarding the ability of a small number of market-dominant firms to exert control over society.

3.3 Privacy

The **protection of human dignity and self-determination** are closely and materially linked with the protection of privacy. The individual's right to determine who may access which personal information relating to him or her, and when and for what purpose they may do so (informational self-determination, see section 3.2 above), is justified by the supreme ethical importance of the ability to prevent intrusions into one's private sphere and also to appear in public in the certainty that one's privacy is protected. Efforts to protect human dignity must include legislative measures to regulate the responsible use of personal data.

A further aspect of privacy is the need to **preserve the integrity of an individual's personal identity**. For example, this integrity may be violated if an algorithmic system – using data collected for entirely different purposes – "calculates" the personality of an individual together with his or her preferences and proclivities, and the system operator then uses these calculations for its own purposes, regardless of or even contrary to the individual's will.

Given that different spheres of society are being shaped more and more by data-driven technologies, it is important for us to **increase the amount of attention we pay to the use of data**. Many people are willing to make their personal data available for public or semi-public use because they will receive certain products and services in return, or because they wish to contribute to the public good. Merely telling the public to think twice before disclosing personal data is not effective. Instead, effective regulations must be adopted so that people can rely on the fact that their data will be used responsibly, and that steps will be taken to prevent any ethically unacceptable uses.

3.4 Security

Algorithmic systems also give rise to crucial security questions. The context of use may promote or jeopardise user security. Security is relevant from an ethical and legal perspective because of the role it plays in **protecting high-ranking values**, such as an individual's physical and mental health and his or her privacy, or public security, peace, and free and equal democratic elections.

Security can relate to collecting and using data, which means that the concept also has a bearing on the **protection of privacy**. The major data scandals that have hit the headlines in recent years have made it clear that privacy breaches and the use of personal data for manipulative purposes can have far-reaching – and sometimes political – consequences.

Consideration must also be given to the **physical and emotional safety** of an individual who operates and uses an algorithmic system. Stringent requirements apply in this respect, e.g. in connection with human/machine interactions. If a robot carer is used, for example, it must be ensured that neither the person receiving care nor the person providing care suffer any harm in terms of their physical and mental integrity.

Algorithmic systems may also have an impact on **environmental safety**. Malfunctions of algorithmically controlled public infrastructures, e.g. traffic or energy and water supply infrastructures, may cause enormous amounts of damage.

Algorithmic systems may also be innately unsafe, causing malfunctions or even functioning as **gateways for malicious attacks and manipulation**. Even beyond inherent system vulnerabilities of this kind, it must not be forgotten that an algorithmic system could be misused for harmful purposes.

3.5 Democracy

Digital technologies are in a complex manner **systemically relevant** for the development of fundamental rights (in particular freedom of expression and information, (informational) self-determination, confidentiality of telecommunications, freedom of assembly and association, freedom of occupation and right to property), for democracy, for the safeguarding of diversity, for an open societal debate and for free and equal elections. For example, social media sites serve as a low-threshold opportunity for every citizen to participate in debates on the shape of our future, and as such should in principle be welcomed. At the same time, however, there is a risk that they may be used for manipulation and radicalisation. The State should take decisive action to counter these risks by adopting rules and setting up institutions capable of preventing undesirable developments and misuse.

It is also an undeniable fact that the rise of the Internet has been accompanied by an economic decline in journalism and its privately funded plurality. Yet the electronic public sphere cannot in any way be considered a valid replacement for the role played by journalism in a democracy, namely that of a "fourth estate" or "watchdog of democracy" – i.e. an instance that exercises control of power and claim to truth on the basis of systematic and independent investigations and criticism. Under certain circumstances, powerful **media intermediaries playing a gatekeeper function** may exert a controlling influence over the democratic formation of will, posing a significant threat to democracy that – based on ethical considerations and the provisions of constitutional law – must be countered through legislative means.

Education and training must also play a prominent role in safeguarding the free democratic basic order, since they influence, in a wide variety of ways, the participation of citizens in the shaping of society – a process that is of critical and fundamental importance for democracy, these citizens' understanding and appraisal of socially relevant interrelationships and developments, and – ultimately – their level of confidence in a future that can be shaped and that is founded on values. Education and training must impart not only technical and mathematical skills, but also skills in the fields of ethics, law, economics and the social sciences.

3.6 Justice and solidarity

Observance of the principles of justice by society and its institutions is another fundamental factor that allows us to live together in peace, prosperity, freedom and democracy. Data and technology have placed enormous influence – both economic clout and the societal sway that results from the former – in the hands of a small number of large companies, and this has raised new questions about a fair economic order. The availability of large volumes of data and the digitalisation of processes e.g. in the workplace and the healthcare sector raises other questions relating to **equitable access and distributive justice**, however, for example in relation to income and the provision of healthcare; these developments may mean that scarce resources can be distributed more fairly, but they may also mean that individual groups of people suffer disadvantage or discrimination.

There is also a close link between justice and opportunities for participation. **Stronger participatory processes**, also supported by digital tools, can play an important role in promoting social innovations during a time of technology-induced social upheavals. Finally, questions of justice arise in connection with situations where the use of algorithmic systems – in particular self-learning algorithmic systems – means that individuals or groups of people suffer discrimination for no justifying reason.

A clear **assignment of responsibility and accountability** is an indispensable feature of a democratic State under the rule of law. An adequate level of transparency and explainability is an essential prerequisite for auditing algorithmic systems appropriately on the basis of their real potential for harm. Opportunities for seeking legal recourse and, if necessary, holding another party accountable, i.e. liable, must also be available under certain conditions.

In the world as it stands today, **access to digital resources** via the Internet is a fundamental requirement for digital and thus also social participation. As part of its public provision remit, the State is obliged to ensure that its citizens can access up-to-date Internet infrastructure anywhere in the country and to an adequate extent, using either a fixed or a mobile connection. As part of its educational remit, it must provide its citizens with the skills needed for self-determined navigation of the digital world and for accurate appraisal of the opportunities and risks of Internet use.

Opportunities for participation promote **social cohesion**, which is also based on a fundamental attitude of societal solidarity and integration of the latter into the institutional framework. Digital technologies may strengthen solidarity, but may also weaken or destroy it. When algorithmic systems are used in certain spheres of society such as the insurance sector or the provision of opportunities for social participation, care must be taken to avoid a systematic weakening of solidarity, which may, in some cases, be caused by very subtle effects. For example, it is perfectly possible for data-driven differentiation and unequal treatment that appears plausible and justified in individual cases to lead overall to a reduction in solidarity with certain groups of people, some of whom may be particularly reliant on society's support.

3.7 Sustainability

Digital technologies offer huge potential in terms of more efficient resource management and innovative business models. This economic aspect generally attracts the lion's share of attention in general debates on the topic. To date, however, less interest has been shown in the question of whether digital technologies can also contribute to economic sustainability. Consideration must also be given to issues relating to ecological and social sustainability. The UN has adopted **17 Sustainable Development Goals relating to economic, social and ecological aspects**, which apply to all the UN Member States and should be achieved by 2030. Digital technologies may make it easier to do so; this is the aim pursued by the International Telecommunication Union (ITU) with its "AI for Good" initiative, for example. Similarly, the German Advisory Council on Global Change (*Wissenschaftliche Beirat der Bundesregierung Globale Umweltveränderungen*) recently outlined its vision of an AI-based and highly granular network of environmental sensors that would allow unprecedented "comprehensive and real-time monitoring of the natural Earth systems, their condition and development", as a vital building block in a future digital sustainability policy.

Yet digital technologies do not only conserve resources; they also consume them, for example through the ever-rising demand for electricity and the reliance of digital products on certain rare earth elements that are only available in limited quantities and in certain countries. Rare-earth mining causes enormous damage to the environment. This raises questions with regard to sustainable economic and ecological development, and also **questions of international justice** concerning the use of natural resources and global responsibility for future generations.

Human knowledge and human skills are also resources whose sustainability must be safeguarded. The development of digital technologies and the concomitant reduction in the tasks that need to be performed by humans will mean that individuals gain certain new skills but lose other **competences of the human being**. A debate must be held on our responsibility towards the next generation, and measures are required to preserve and develop certain skills and avenues for independent action.

As noted elsewhere in this Opinion, there is a need for regular and comprehensive **technological impact assessments**, and these assessments must also consider the sustainability of new technologies in their various manifestations. It is incumbent upon the legislator to ensure that responsibility for sustainability is incorporated into the rules that govern the data economy and algorithmic systems, for example through the introduction of an obligation to disclose the entire energy footprint of an energy-hungry blockchain system.

The pursuit of sustainability goals set by the United Nations should be a particular focus of **public investments** into the data economy and algorithmic systems. When allocating government funding, priority should be given not to economic gains which are only short-term in nature, but to the development of data and algorithmic systems for purposes such as recording and monitoring environmental impacts and developments, or systems for optimising and reducing energy and resource consumption. In addition, more should be done to promote sustainability-oriented social innovations that foster social creativity and participation.

Part C

Technical foundations

Data-intensive IT applications have a lasting impact on our living and working environment, our economy, our scientific endeavours and our society. As well as being permanently tethered to our smartphones, we use search engines on a daily basis, rely on recommendation software, send text or voice messages to our family and friends, regulate the temperature in our home remotely and allow navigation devices to guide us from one place to another. We are able to do so because of a series of technological developments that have occurred over the past few decades. Some of the fundamental technical concepts underpinning these developments are described below; the aim is not to provide a comprehensive account but to highlight key points as a basis for identifying any resulting problems and starting points for potential governance approaches.

1. Status quo

Entirely new fields of application have been opened up thanks to the improved performance and miniaturisation of the physical components of IT systems (hardware) that are used to store and process data, along with continual enhancements to both wired and wireless connectivity. Smartphones, tablets and wearables are gradually infiltrating our workplaces and homes, along with sensors, actuators and, in some cases, "autonomous" systems such as robots. In many locations, the Internet is "always on" thanks to mobile access, making it possible – e.g. in combination with various sensors in smartphones, such as geolocators, gyrosensors, cameras, microphones, etc. – not only to input text, but also to upload image, video and audio recordings to the Internet at any time and from almost anywhere. This penetration of technology makes it possible not only to communicate and use social networking sites, but also to link devices to the Internet of Things (IoT).

It has become impossible to draw a clear dividing line between the analogue and the digital worlds; the former contains more and more components that transfer information into the latter, while digital information is becoming ever more widely available in the analogue world, bringing the two closer and closer together and creating a **hybrid world**.

Data volumes are increasing exponentially thanks to comprehensive arrays of sensors, the IoT and the falling price of storage capacity. Specialised tools are needed to process such large volumes of data. At the same time, the accumulation of so much data (together with the availability of high-performance hardware) has promoted the widespread use of machine learning procedures, and some of these have achieved impressive results, for example in the field of speech and image recognition.

Speech recognition and video processing have now seen such huge leaps forward in terms of performance that there is potential for the **boundaries between reality and computer-generated information** to become blurred. When this happens, people are no longer sure whether or not they are talking to a speech bot, or whether they are watching a normal video recording or a "deep fake", i.e. a synthesised human image saying things that the real person never actually said.

2. System elements

2.1 Data

2.1.1 Definition and properties of data

In keeping with the Data Ethics Commission's mission, this report concentrates on data that are **digital and machine-readable**. These data are made up of a stream of binary electrical impulses, which may be transient (signals that only exist for an instant, e.g. a control impulse for a technical system) or persistent (stored on a medium).

Data are multifaceted. The word "data" is an umbrella term that encompasses an enormous range of manifestations. For example, data can be categorised on the basis of data type (e.g. binary, nominal, ordinal, metric and textual data), the process used to generate the data (e.g. survey data, sensor data), the sector in which the data are collected (e.g. financial data, weather data) or their function in a digital system (e.g. login data, training data). They can be further categorised on the basis of their level of processing. Data that have not yet been processed are referred to as "raw data". Processed data are referred to as "structured" or "unstructured", depending on the level of structuring (normalisation). Data can function as the input into a system or the output from a system, and an output may, in turn, function as an input into another system. Data can also represent digital assets, such as multimedia content or units of cryptocurrency. A further distinction of enormous legal significance is that between personal and non-personal data.

The terms "data" and "information" are not always synonymous. To make sense of the **binary electrical impulses** that form the basis for digital data, i.e. to transform data into "information", it is necessary to know their **context** and **semantics** (meaning). One possible context would be the origin of a generated signal – knowing which precise sensor emitted a signal, for example. The term "semantics" refers to the information contained in a certain sequence of binary signals; for example, a "4" that appears in a survey may equally well represent the number of children in a household or the number of tubes of toothpaste bought in the past six months. Potential sources of context and semantics include metadata, domain tables, ontologies, identifiers and other technical specifications that supplement data values. Whenever the term "data" is used in the remainder of this report, familiarity with the context and semantics will always be implied.

Data are of varying quality. The purpose of most data – or, more accurately, the information contained therein – is to reflect reality as accurately as possible. This can for example be done by assigning attributes that are exhibited by entities in real life to the correct entities in the digital world (information objects). There are also many types of data that are intended to express the likelihood of something happening in reality (either now or in the future). Some types of data are intended to construct a hypothetical reality, while others have no relation to reality whatsoever. In all of these cases, the pool of data **may contain errors**. A distinction should be made between these errors and cases in which the data do what is expected of them but are **unsuitable** for achieving a specific goal, for example performing a particular analysis (e.g. the data are insufficiently granular, or outdated, or incomplete in some way).

The quality of the data used is of decisive importance for data-driven systems, since even a perfect algorithm cannot deliver high-quality results if it receives poor data as an input (i.e. inaccurate or inadequate data). Data quality is not an absolute value; the relevant data quality dimensions and their quality level depend on the specific use (see Figure 1).

Figure 1: Example of different use-specific quality requirements

2.1.2 Data management

Data are not some pre-existing entity – they are created. The process of collecting, preparing and processing data involves many different human decisions that have implications for the future use of the data. For example, the potential that might have been gained from data may be irretrievably lost if they are stored without any context or semantics. Careful **data management** is necessary to avoid situations of this kind.

Before collating data from different sources, it is vital to ensure that the collation will be possible from both a technical and a semantic perspective ("interoperability"). The data from these different sources must be mapped against each other in a way that reflects their semantics. In cases where interoperability is particularly important, efforts should be made to achieve **standardisation** of the technical specifications (formats, descriptive metadata, etc.). Reference data play an important role in this respect, i.e. standardised schemes or ontologies, some of which fall under the remit of national or international institutions (e.g. the International Classification of Diseases (ICD) published by the WHO).

2.1.3 Big data and small data

The term **"big data"** does not refer to a separate type of data, but instead to a new methodological approach for the identification of relationships. Laney[1] famously used the "three Vs" – volume, velocity and variety – to define this approach while it was still in its incipient stages; large volumes of varied data, potentially from a variety of sources, are generated at high velocity (often in real time). Special technologies are needed to process these large volumes of rapidly changing data that vary in terms of both their nature and their quality. The analysis of large data sets ("big data") is particularly well suited to situations where it is necessary to identify the most promising of a large number of potential correlations. In the field of medical research, for example, it is helpful to start with big data methods that identify a number of likely candidates from a long list of environmental factors that might increase risk for a disease, before going on to perform costly and high-precision experiments or studies that investigate only these candidates. A specific problem associated with this approach is that it initially shows only **correlations** rather than causalities, and completely unsuitable candidates may therefore be identified.

1 Doug Laney: 3D Data Management. Controlling Data Volume, Velocity, and Variety, META Group Inc., 2001.

In many areas, the volumes of data available will never be large enough to allow analysis using big data methods (for example, the client base of a small or medium-sized company may never exceed 200 customers, and the number of political parties in one country rarely reaches three figures). Suitable **"small data"** analytical methods can also be used to extract a great deal of knowledge and information from data. The quantity of data is not what matters; instead, the decisive factor is the availability of suitable tools that make it possible to combine data of an adequately high quality in quantities that are sufficient for the task at hand, as a basis for effective data analysis.

2.2 Data processing

2.2.1 Algorithms

From a data protection point of view, the term **"processing"** refers to the entire sequence of actions from data generation and extraction through to storage and any transformation of the actual data (Article 4(2) GDPR). By way of contrast, the mathematical and technical sciences mainly deploy the term to refer to the use of data. The following arguments are based on the latter of these two understandings of the term.

Any method of digital data processing follows the **IPO (input, processing, output) model** – data enter a system as an input, are processed, and then leave it as an output. Any form of internal processing within an IPO system is based on an algorithm, or in other words an operational processing sequence that specifies a procedure as a series of different processing steps, with the aim of achieving the desired result through successive transformations of the data inputs. Algorithms have been around since the time of Euclid, who specified a method for easily calculating the greatest common divisor of two natural numbers. The word "algorithm" is derived from the name of the Arabian mathematician al-Khwarizmi (formerly Latinised as "Algorithmi"), who published a collection of calculation rules for solving algebraic equations in 830 AD or thereabouts.

It is hard to overestimate the importance of the term "algorithm" in modern computer science. To solve a particular problem by processing data, an algorithm must not only be implemented correctly, but also used productively. This presumes a knowledge of the algorithm. In many cases, however, the algorithm that will ultimately deliver the desired result is not yet known, and the first and most important task is to **find a suitable algorithm**. For many situations of practical relevance, the processing specifications can be derived directly (i.e. deduced) from specialist knowledge, known models or legislative provisions. In other situations, our understanding of the context is not yet sophisticated enough to allow it to be described using more or less simple mathematical formulae.

If this framework of understanding is absent, various strategies can be applied to identify an algorithm. These include random chance, trial and error or data-based **inference**. The latter approach follows the principle of induction – an attempt is made to infer a general rule from individual cases (i.e. the data). If a general rule is found that can be used to solve the question, it can be assumed to be a suitable algorithm. It is worth remembering that there may well be several suitable rules, and furthermore that the result of this process of induction may not necessarily be correct. The result inferred from the individual cases may be partially or wholly inaccurate.

2.2.2 Statistical inference

A central concern of statistics is the drawing of inferences from data. **Statistical inference procedures** can be applied to data sets to investigate problems that lack a known inherent logic. More importantly, however, they can also be used for problems where random chance forms an integral part of the process to be modelled. Examples would be estimating the probability that it will rain on the following day, or identifying high-probability prospects for a particular product. There are many different statistical inference methods to choose among, starting with various forms of regression (linear regression, logistic regression or regularisation (ridge regression)), moving through support-vector machines (SVM), Bayesian networks and rule learners (such as Aprioiri, CART and random forest), and ending up with neural networks (NN). All of these procedures are suitable for extracting information from the available data. Some of them are specifically designed to solve regression questions, for example estimating the future height of a child based on the height of his or her parents, whereas others, such as SVM, CART and NN, are used for classification-type question, e. g. pregnant/not pregnant, dog/cat. Whether or not they represent a suitable means of answering a question depends on many factors, including the data volume and type.

Besides methods for induction, statistics offers a broad set of **tools for measuring the quality of the results (estimations) obtained**. These measurements can be used to estimate potential errors and to monitor actual errors in practice. Thus an estimate of a child's future height can be stated as 175 cm with a deviation range of +/– 4 cm. If a pregnancy test yields a positive result, this result might be deemed to be 93% accurate. A pregnancy test is a good example of the need to monitor two different parameters: the number of false positives (e. g. when the women is not pregnant but the pregnancy test is positive) and the number of false negatives (e. g. when the woman is pregnant but the pregnancy test is negative). The ideal statistical procedure would never result in any of these errors. In practice, it is necessary to weigh up the severity of the two errors and decide which false rate should be minimised. Is it worse for a woman to find out at a later date that she is, in fact, pregnant after being told that she is not, or for a woman to be told that she is pregnant when this is not true? The two error types cannot be minimised at the same time, since it is generally the case that the lower the frequency of one, the higher the frequency of the other. A balance must be struck, and this will look different depending on the context.

The quality characteristics of the methods themselves are used as a basis for assessing the quality of the results. It is even possible to **guarantee the quality** of the results obtained using certain methods; for example, estimation procedures that use a uniformly minimum-variance unbiased estimator (UMVUE) ensure that the best possible results are obtained using the data available. If a regression using UMVUE-based parameters supplies a result stating that the expected height of a child is 175 cm +/− 4 cm, no other estimator would have achieved a smaller error margin. Similarly, if a support-vector machine is used, the model determined on the basis of the relevant data (provided that a model can be found at all) is guaranteed to be the best possible model for the method in question. In certain cases, well-founded procedures for assessing the quality of either the model itself or the estimates generated using the model are yet to be developed – this applies, in particular, to the method class of neural networks. Quality indications can also be provided for neural networks, however. Measurements of how well a model functions using data that were previously unknown are particularly important. The model is taught using one data set (training data) and assessed for quality using a different data set (test data). This approach can be used to identify models that do not reflect the general rule because they have learned their training data too thoroughly. Cases of this kind are referred to as overfitting; an overfitted model will achieve significantly better quality values for the training data than for the test data.

Many statistical procedures can be solved analytically. This means that the question can be formulated as a mathematical equation or a system of equations and solved through transformations (even though this often requires a great deal of skill). However, a direct analytical solution is impossible for many other methods (for example if additional conditions such as a regularisation term are applied, see below). In these cases, use can be made of **optimisation procedures** that approximate the solution through many small steps. Optimisation procedures are not necessarily optimal; for example, the calculated result may be only a local optimum and not the global optimum (or one of them).

Different classes of problems: analytical procedures and optimisation procedures

A direct analytical solution is possible for tasks such as “Find the value of y for the equation $y=4 \cdot x+3$ where $x=3$”.

A solution of this kind is not possible for the task “Solve the linear equation $a \cdot x_1 + b \cdot x_2 + \cdots + h \cdot x_8 = y$, in which as many parameters as possible a, b, ..., g, h are equal to 0”.

An additional regularisation term is applied for this purpose: $\min((a \cdot x_1 + b \cdot x_2 + \cdots + h \cdot x_8 - y) + \text{sum}(\text{parameter} \neq 0))$.

Optimisation procedures are used to find solutions.

2.2.3 Machine learning

The boundary between traditional statistics and **machine learning**, a term first defined by Mitchell,[2] is difficult to delineate. The scales tip towards machine learning at the latest when optimisation procedures (→ see section 2.2.2 above for further details) are used to solve inductive inference problems.

The different approaches to **estimation or "learning" strategies** that fall under the heading of machine learning can be differentiated on the basis of the formulation of the optimisation problem to be solved. A distinction is made between a number of different learning procedures:

- **Supervised learning:** Supervised learning procedures require knowledge of the correct output (the "O" in the IPO model) for each piece of information used as input (the "I"). Height is a classic example: before inferring the height of a child (output) from the height of his or her parents (input), it is necessary to know the height of the child in advance. It is also necessary to know the correct result of a pregnancy test, the actual weather that follows a weather forecast, the properties of the soil predicted by a soil analysis, etc. In practice, the real challenge often lies in obtaining the correct output information and assessing its quality. This output information is frequently referred to as a **label**. The majority of machine learning algorithms currently in use were trained using supervised learning procedures.
- The decisive questions with regard to these learning procedures are how to formulate the actual optimisation problem, which regularisation terms to use and how to define the loss function (i.e. are all errors treated the same, or are there different weightings and levels of severity, e. g. when comparing false negatives for patients with cancer who are incorrectly diagnosed as healthy and false positives for healthy patients who are incorrectly diagnosed with cancer?).

Quality of labels

Labels can also contain errors. Several levels of complexity can be defined for data labelling:

1. Labels whose accuracy can be verified when the data are collected. Example: only one correct and relevant value exists for physical systems or properties such as the speed of an object, the temperature of a room or an individual's date of birth. In principle, therefore, these values can be ascertained as labels by an algorithm.
2. Labels whose accuracy cannot be verified when the data are collected and may, in certain cases, not be verifiable at a later date.
3. Labels with a construed and non-verifiable relationship to the real world. Example: concepts such as social milieus or character types have been developed with a view to achieving a better understanding and analytical grasp of humans and their behaviour. These concepts are abstractions that are not necessarily an accurate representation of the "truth" (in so far as it exists).

2 Tom Mitchell: Machine Learning, McGraw-Hill, 1997.

Identifying an optimisation goal

A public transport company is planning to alter its bus routes to reflect recent changes in the city where it operates; many residents have moved to peripheral areas, large inner-city brownfield sites have been developed, and gentrification has brought about huge changes in the composition of the population in various districts. The project manager has collected data in the form of passenger and usage figures, and is attempting to optimise the routes served so that the city's needs can be met as effectively as possible without needing to use extra buses. He is aware that a range of different goals or constraints could be imposed on the optimisation, such as using fewer buses, using fewer drivers or avoiding the creation of new routes. For example, depending on how the optimisation problem is formulated, it might be possible to achieve a solution whereby densely populated neighbourhoods are served by more bus lines compared to other districts, but anyone living in a suburb is forced to put up with longer travel times or a lower frequency of buses. Since the project manager himself lives in the affluent commuter belt, he has a personal preference for an optimisation strategy that minimises the longest travel time. A strategy of this kind would result in faster connections to all areas of the city, including the outlying districts. His line manager is unimpressed by both of these models. He believes that the goal should be to transport as many passengers as possible. This puts short-distance routes with plenty of passengers at an advantage, but is bad news for longer routes with more than four stops. It should be readily apparent from the above that decisions on the optimisation function can have social impacts. Many questions are raised, including the following: Who should decide on the goal of optimisation? Who else should have a say in the decision? How can the matter be debated with the general public, and is it necessary and meaningful to do so? Should certain groups/neighbourhoods have access to legal remedies if they feel that they have been placed at an unfair disadvantage compared to others?

- **Reinforcement learning** involves assessing an agent's actions and imposing a punishment or reward. An agent selects from a pool of different actions and performs whichever action it has selected; this action changes the state of the system and functions as an optimisation input. In addition to the state (or change in state) of the system that is brought about by the agent's actions, there must also be a clearly defined reward function. In the case of supervised learning, the correct and optimal solution is available for every input; this is not necessarily true in the case of reinforcement learning. Instead, the optimisation goal pursued is that of finding the action strategies that lead to the best end state with reference to the optimisation problem. Actions that deliver only short-term improvements may need to be rejected to achieve this goal. Alongside the optimisation problem itself and the relevant loss factor, the reward function plays a particularly important role in this learning strategy.
- **Unsupervised learning** involves searching for structures in a particular quantity of input data. There is no need for the correct structures to be known or for a reward function to exist. A precise definition of the structure being searched for is required, however. For example, a search can be carried out for clusters (i.e. groups in the data) by imposing the requirement that the difference between all the data points in a cluster should be minimised while the difference between the clusters should be maximised. The optimisation problem for unsupervised learning is identified on this basis. Unsupervised learning is also referred to as **data mining**.

Decisive factors include not only the learning procedures but also the availability of sufficient volumes of data that are adequately high in quality and broad in scope, since close approximation of an optimisation goal cannot otherwise be achieved. In many cases, the volume, quality or scope of the data are lacking in some way, meaning that other avenues must be pursued to ensure that good outcomes can nevertheless be obtained using machine learning techniques.

For example, **synthetic data** can be used, i.e. data that are generated artificially rather than being collected directly in the real world and that boast several advantages over real-world data.[3] They can be produced in any quantity, which is particularly important when dealing with simulations for which real-world data cannot yet be generated. When they are created, steps can be taken to ensure that the entire range of possible values is included in the synthetic data, e. g. in order to test how a technical system would behave when confronted with unusual data combinations. Their quality can be measured, and if necessary it can be guaranteed in individual cases that the properties of a set of real-world reference data are retained; alternatively, distortions occurring in sets of real-world data can be pinpointed and removed in order to avoid discrimination. If the set of synthetic data contains no references to persons, it is anonymous and does not fall within the scope of the GDPR. Synthetic data can also be used to train algorithms or test systems; there is, however, a risk that the algorithm will be influenced by properties of the artificially generated data that have no counterpart in reality. Separate functional testing must therefore be carried out before the algorithm is used for practical applications.

A middle course is frequently adopted in the form of **augmentation**. This involves creating new data from the real-world data so that a greater range of situations can be covered at the training stage; the pool of data is enlarged, but the relationship to the real-world data is preserved. The term "augmentation" describes the process of generating new data that deviate slightly from the original data. For example, a characteristic feature of augmented images is that they have been shifted, rotated or distorted in some way.

2.2.4 Artificial intelligence

In the current parlance, the field of machine learning – and more specifically neural networks – is referred to as **artificial intelligence (AI)**, but this term often gives rise to confusion. Machine learning is only one specific procedure that falls under the heading of "weak AI" and that is used to solve well-specified tasks. By way of contrast, "strong AI" methods are expected not just to tackle a single task, but to handle a broad spectrum of tasks, potentially without human intervention. Despite the hopes raised by the term "artificial intelligence", machine learning methods are not capable of such feats.

Historically speaking, the concept of **artificial intelligence** first appeared in the Dartmouth Proposal, published back in 1956 in the USA,[4] to refer to a broad area of research within the field of computer science. The decades since AI first emerged as a field of research have been marked by repeated cycles of unrealistic expectations followed by disillusionment. AI left the ivory towers and made inroads into the economy and everyday life (both workplaces and homes) at the latest in the 1970s and 1980s, in the form of "expert systems", and research efforts in Germany stepped up a gear in the 1980s.

Achievements that can be chalked up to AI research include not only machine learning techniques, but also a large number of other vitally important methods, such as procedures for **pattern recognition, knowledge representation, inferences, action planning and user modelling**. Applications for these procedures include speech, image and dialogue comprehension, robotics and multi-agent systems.

3 Jörg Drechsler/Nicola Jentzsch: Synthetische Daten: Innovationspotenzial und gesellschaftliche Herausforderungen [Synthetic data: potential for innovation and societal challenges], Stiftung Neue Verantwortung, May 2018 (available at: *https://www.stiftung-nv.de/sites/default/files/synthetische_daten.pdf*).

4 John McCarthy/Marvin Minsky/Nathaniel Rochester/Claude Shannon: A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1955.

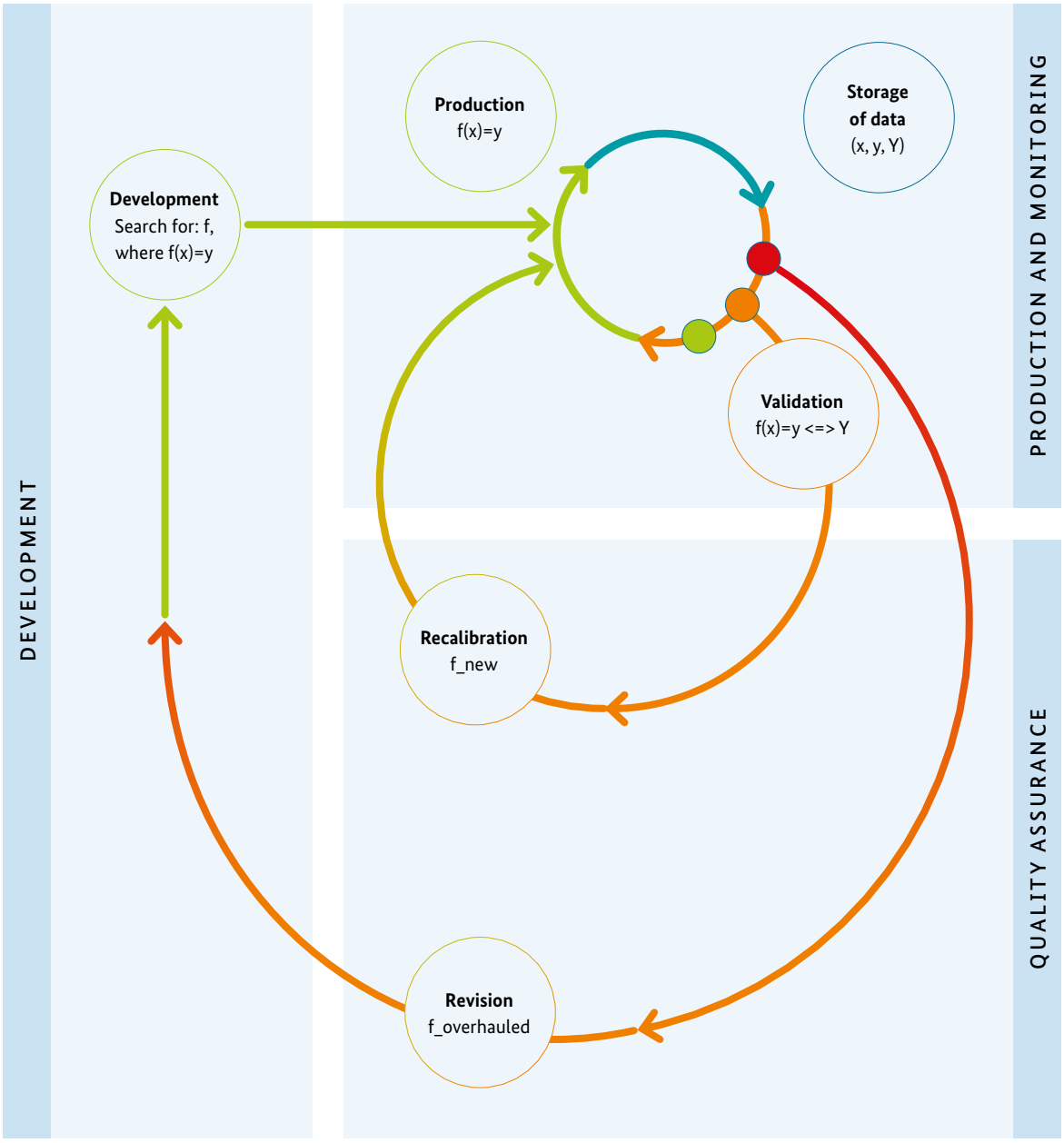
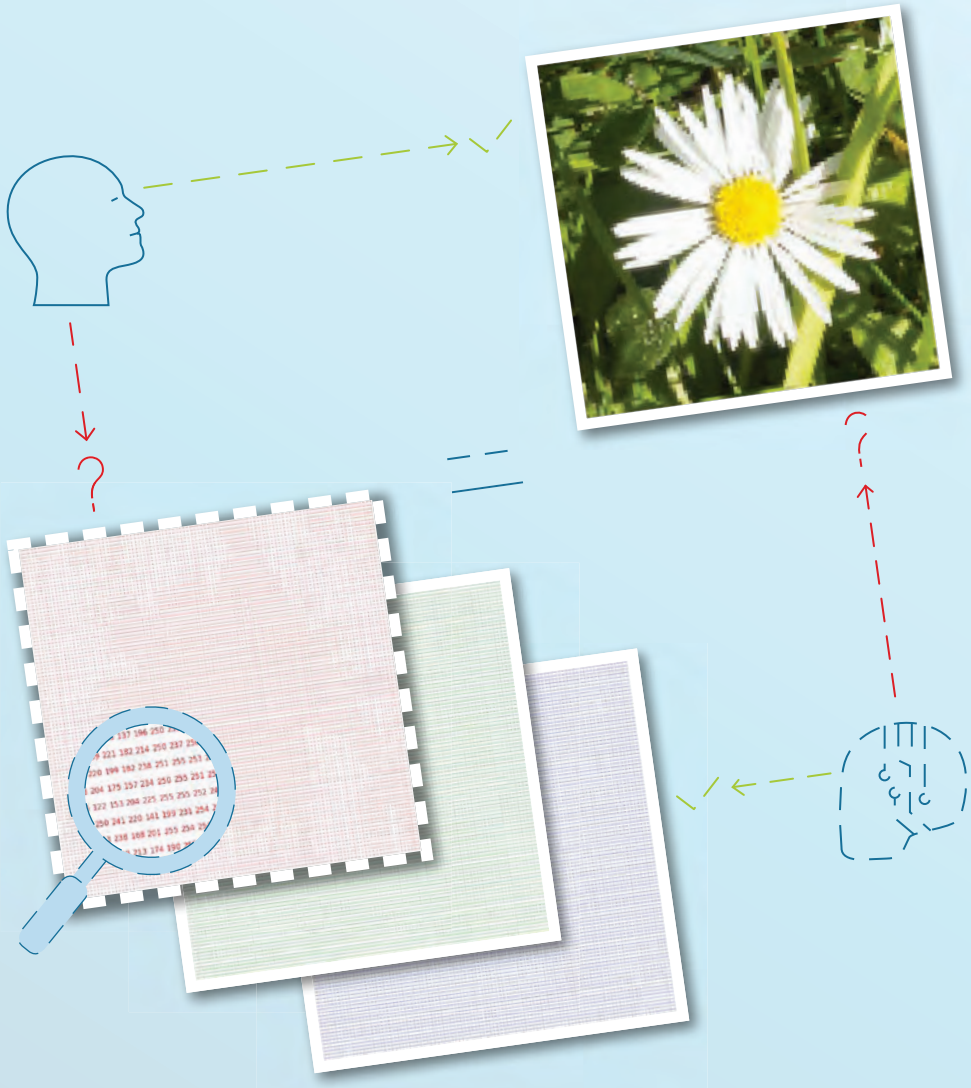


Figure 2: Process model of an algorithm based on machine learning: ongoing monitoring and assessment. The process starts when an algorithm (f) is developed using the training data. Once an algorithm has been identified that meets the desired quality standards, it is put into production. To ensure monitoring and quality control capabilities the production process must make it possible to record the input (x) that enters the algorithm, the output (y) that leaves the algorithm, and the relevant correct value (Y). This information can be used as a basis for monitoring the algorithm in a production environment. To do so, a comparison is carried out to determine the extent to which the output of the algorithm (y) reflects the expected value (Y). The algorithm can continue to be operated without changes in the event of non-critical deviations between these values. If significant deviations are detected, it may be necessary to re-evaluate (i.e. recalibrate) the parameters of the algorithm. If critical deviations are detected, an algorithmic redesign is recommended.

The problem of understanding and comprehending

Humans often find it difficult or impossible to understand methods intuitively if they are described in mathematical or technical terms. This even goes for experts in the field of modelling. Even in the case of relatively simple classification methods that are well understood mathematically (such as logistic regression), almost no one can intuit which result they will return for a given set of input values.

Neural networks for image recognition are a good example of this phenomenon; a human can generally look at a photograph and understand immediately what he or she is looking at, but a human looking at the data structures used as an input for a neural network intended to classify the same photograph is likely to understand almost nothing. This means that, even if a human is familiar with all the digital input values and comprehends all the steps in a neural network, he or she will not necessarily understand the recognition process. If an error occurs, for example, he or she may not be able to determine why recognition has failed and how the problem can be fixed. Humans and machines recognise objects and patterns according to different sets of rules, and it is not always easy to translate between the two.



2.2.5 Algorithmic systems

An algorithmic system generally incorporates multiple algorithms that can work together rather than a single algorithm, and the term "component" is used to describe an executable part of such a system. Different components of an algorithm might be based on different technical implementations. The architectural style known as microservices is a good example. It is important to remember that the individual components of a system of this kind might be subject to different regulatory requirements or protection objectives during their development. In addition, different stakeholders might be responsible for different components of an algorithmic system, for example as suppliers, operators or manufacturers. It should be borne in mind that different requirements or different sets of rules might apply to the individual components, e. g. in respect of data quality, non-discrimination or freedom of contract.

2.3 Software

If an algorithm is formulated in a programming language (formal language) rather than natural language, it is executable in automated form on a computer as a **program** (or **software**). The functioning of software depends not just on the data it processes, but also on the context in which it is executed (cf. concepts such as the "technology stack", which contains all the hardware and software components used for execution) and its parameterisation. Parameters are an "outside-in" method of configuring software. They make it possible to pass information to the software, ranging from simple data (such as display options or path names) through to complex models. More extensive parameterisation options generally go hand in hand with more flexible software use and a more complex development process, making parameters all the more important. For example, software that can be parameterised can be adapted to different contexts with a relatively small amount of effort, and without modifying the source text (i.e. the actual implementation). There are special variants of adaptive systems which over time automatically adapt to their context – such as the individual using these systems or the environment in which they are used.

In order to guarantee or improve the efficiency of high-quality software development processes in spite of increasingly complex framework conditions, and in order to reduce communication problems during these processes, **model-driven development approaches** have been pursued successfully for many years. A generic software component is parameterised on the basis of a complex model, using a language specific to the application context. Mathematical and statistical models represent a special case, and differ from domain-specific languages in that a model is not explicitly specified or programmed; instead, the mathematical or statistical model is (implicitly) taught or trained using data (→ see section 2.2.3 above on machine learning).

2.4 Hardware

Software is executed by hardware, and in particular by **processors**. In recent years, these processors have seen steady gains in performance, while the devices themselves have seen continual reductions in size, meaning that the array of potential applications has become ever wider. Moore's Law (according to which performance should increase a hundredfold every 10 years) is subject to physical constraints, however. When chip components become so small that they are barely bigger than individual atoms, fulfilling Moore's predictions using silicon as a transistor material becomes an increasingly costly and technically challenging task. Researchers are therefore currently investigating alternative materials such as graphene in conjunction with new computing concepts such as photonic quantum computing. The question of whether these will be suitable for everyday use remains open, however. Solutions focusing on parallel computing are more established, and include multi-core and many-core processors or the use of graphics processing units (GPUs). In order to accelerate machine learning using bulk data, application-specific chips (such as tensor processing units, TPUs) that are optimised to handle the highly parallel addition and multiplication of matrices for neural networks have been developed.

The increasingly parallel nature of computing is not without its problems, however; humans find it very difficult to identify any related processor errors, and the calculations performed at the hardware level are **almost impossible to reproduce and comprehend.**

2.5 System architecture

Applications today rarely run on a single computer. Instead, many different software components run on different computers and interact with each other to perform a task. The term **"distributed system"** is used to refer to this method of distributing the work across different hardware nodes. A distributed system is made up of different software and hardware components that interact within a network. The network nodes communicate with each other over wired or wireless links.

A wide range of **protocols and standards** exist for network communication, and are used as a basis for processing data at the network nodes and forwarding these data through the network (i.e. transporting them to other nodes). Specifications outlining the requests that can be submitted to a server are published in an application programming interface (API), for example. As a general rule, steps must be taken to prevent these interfaces being used incorrectly or accessed by attackers.

IT infrastructures that can be reached via the Internet are referred to as the **cloud**, and cloud applications can be accessed by billions of users. Groups of related cloud applications are often referred to as **digital platforms**, and many – such as the "Big Four" or "GAFA" (Google, Apple, Facebook, Amazon or "GAFAM" if Microsoft is also included) – have a high level of name recognition.

In the early days of the Internet of Things, most data were sent directly to the cloud and processed there on large digital platforms. By way of contrast, an increasing number of solutions are currently being developed that involve the processing (or at least pre-processing) of data immediately and as close as possible to the place where they are collected, or in other words "on the edge" of the Internet. This practice of processing data near to where they are collected is referred to as **edge computing**, to distinguish it from situations in which the data are processed in the cloud (cloud computing). Data pre-processing is particularly important, since it allows not only the minimisation of communication effort, but also the creation of more privacy-friendly systems, since any references to individuals that are not required can be removed at this point (close to where the data are collected).

The complex system landscape that has emerged in recent years (incorporating the Internet, edge computing and IoT) entails a high level of interconnection, making it hard to distinguish the individual systems from one another.

The way in which the architecture of distributed systems is designed also has a significant **impact on the business processes** supported by the system, since it acts as a factor in decisions on the technology that is used, the network nodes on which the software runs, the interfaces and protocols used for communications and the other parties involved in these communications. For example, if manufacturers want to use the hardware data collected by their devices for the purpose of long-term efforts to improve those devices, they have the choice of setting up their own communication infrastructure, making use of the user's own infrastructure (where available) or asking the user to make the data available via an interface. The way in which data of this kind are handled in cooperative processes should be transparent and agreed contractually if necessary. Technical parameters may place constraints on the contractual provisions governing the exchange of data.

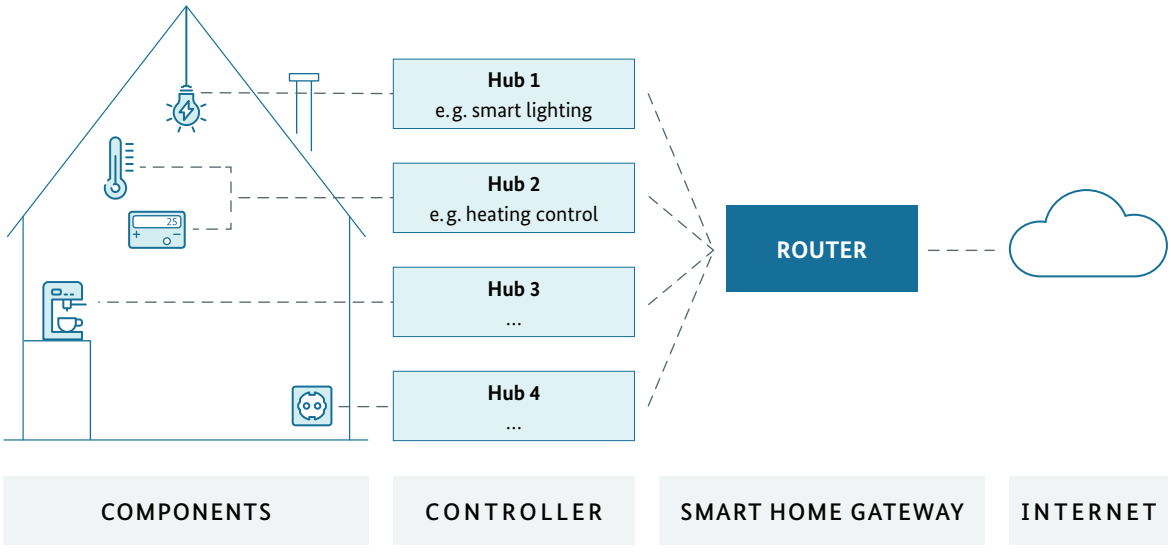


Figure 3: Example of system architecture in the smart home

Blockchain and other distributed ledger technologies

Significant improvements in the field of distributed systems have made it possible to use **distributed ledger technologies (DLT)**. These technologies involve the management of multiple identical copies of a ledger by different partners, instead of the centralised management of a single ledger. New ledger entries are added to all of the copies, and the current accuracy of the database is confirmed by consensus. The underlying architecture of systems of this kind varies from linear approaches to a wide range of graph-based solutions, depending on their intended purpose and the structure of the transactions. A consensus can also be achieved using different methods. These methods are outlined in consensus protocols.

One of the most famous examples of a DLT architecture is the **blockchain** concept, implementations of which include Bitcoin and Ethereum. Blockchains are used to store data as a list of records (“blocks”). The blocks are linked to each other using cryptography, meaning that a transaction stored as a block implicitly confirms the accuracy of previous transactions (i.e. the entire chain), making it extremely difficult for fraudsters to manipulate the data by modifying it or deleting entries. Use of a decentralised consensus protocol eliminates the need for an additional instance that confirms the integrity of transactions.

Part D

Multi-level governance of complex data ecosystems

The high level of complexity and dynamism of data ecosystems means that new challenges must be overcome in terms of regulating, controlling and designing these systems before the ethical and legal framework upon which the Data Ethics Commission has based its work can be implemented in practice; this will require cooperation between different stakeholders and interaction between different governance instruments at many different regulatory levels (multi-level governance). Part D examines **relevant governance instruments and stakeholders**, with further details provided in the following two parts on data and algorithmic systems (in particular regarding the interplay between different instruments and stakeholders).

1. General role of the State

All those who are entitled to exercise ethically justified rights and who are obliged to comply with the corresponding obligations – be they citizens, companies or government agencies – must actually be able to do so in practice. This presents the State with a wide range of tasks. First and foremost, the State is responsible for establishing a **legal framework** within which a data society geared towards the public interest can develop. The speed at which algorithmic systems are developing and infiltrating ever more areas of life poses major challenges for the legislature and the courts that hand down rulings clarifying the legislative provisions. The State must ensure that any regulations adopted in an environment of this kind are sufficiently hard-hitting to steer developments, while at the same time being flexible enough to continue fulfilling their purpose even if the technological parameters change. Statutory provisions must therefore be formulated in a **technology-neutral manner**, and **innovative regulatory models** must be developed.

In addition, the **appropriate infrastructural and technological prerequisites** must be in place – such as enabling technologies, institutions and intermediaries, complemented by the involvement of a broad gamut of civil society actors. The Data Ethics Commission believes that, here too, the State must play a key role in guaranteeing and safeguarding these services of general interest.

The new opportunities opened up by the data society also impose a far-reaching **educational remit** on the State. It is necessary to identify the skills required to take a creative yet reflective approach to the use of digital technologies, and to determine the framework conditions that must be put in place before appropriate training can be offered to a diverse range of target groups. The State's educational remit should be understood in a broad sense, and should incorporate **public outreach work** with the aim of raising awareness in this area.

Furthermore, the State is also generally responsible for encouraging **research and development** (R&D). It is particularly important here to support R&D with regard to ethically sound technologies (e.g. those that uphold the principles of accountability, transparency and anti-discrimination). Extensive research and development programmes are needed to ensure that ethical and legal principles are taken into account, and more funding must be channelled towards these programmes.

Not all of the funding needs to be provided by the State itself or by institutions that are closely aligned with the State, but the State must put in place the **framework** (legal and otherwise) for a data society in which individuals and businesses alike can operate in a self-determined fashion on the basis of ethical values and principles, in which these individuals and businesses are provided with adequate protection, and in which the potential of data and algorithmic systems are harnessed to shape a worthwhile future.

Germany's efforts in the direction of ethically sound and multi-level governance should also include active contributions to **debates at the European and international level**. The global dimension of technological developments means that action by a single nation state or regulations adopted at the national level alone are inadequate. The Data Ethics Commission therefore welcomes the European and international initiatives that have already been launched (by the European Commission and the OECD, for example) with a view to ensuring that our future is shaped on the basis of ethical principles. Safeguarding the **digital sovereignty** of Germany and Europe in the international context is a vitally important task in this regard (→ see Part G for further details).

2. Corporate self-regulation and corporate digital responsibility

Responsibility for mitigating the risks of digitalisation and for leveraging its significant potential should not be placed solely at the feet of the State and its legislators. This **responsibility** should also be shared with the parties that develop, disseminate and use the technologies, even **in the absence of any legal obligation**. Although the State must shoulder most of the responsibility, not least because it is obliged to protect its citizens by guaranteeing the confidentiality and integrity of IT systems and safeguarding other fundamental rights, self-regulation tools are also vitally important, particularly in the context of the digital transformation process.

The term **"corporate digital responsibility" (CDR)** is used at a theoretical and practical level to refer to the idea that companies, as manufacturers and operators of digital technologies, should each assume their own responsibility for the consequences of digitalisation. Like corporate social responsibility (CSR), CDR falls under the broader umbrella of corporate responsibility; in this case, the focus is on voluntary corporate activities in the digital sphere which go beyond what is currently prescribed by law, and which actively shape the digital world to the benefit of society in general, and of customers and employees in particular. To further this aim, in October 2018, the Federal Ministry of Justice and Consumer Protection launched an initiative to clarify the principles and concepts of corporate digital responsibility (*www.bmjv.de/cdr*). According to this initiative, CDR can encompass many topics,[1] including the protection of personal data, inclusion in the digital sphere, transparency (e.g. in relation to algorithms or data protection), the development of digital innovations that help to achieve sustainability objectives, algorithmic use that is geared to the public interest, open data and information security.

The responsible development of digital products and services must be a central priority in all corporate decisions taken at all levels of the company. Ethical questions must not be a matter for legal departments and compliance officers alone. Instead, they must be viewed as a **cross-cutting task** and **integrated into all processes**. All of the parties involved must be aware of their responsibility to consider ethical values such as participation, fairness, equal treatment, self-determination and transparency. The negative social and societal impacts of digitalisation and digital business models on employees, suppliers, clients, society as a whole and the wider environment should thus be minimised, and the new opportunities that digitalisation offers for the achievement of macrosocial goals should be leveraged. When applied correctly, the concept of CDR can lead to improvements in terms of consumer protection, digital participation and the **sustainable development of the digital economy**.

CDR is fundamentally similar to corporate social responsibility (CSR) in that it requires companies to take self-regulatory action on a voluntary basis. Internal strategies such as in-house or industry-specific codes of values are therefore a particularly effective way of implementing CDR. In this respect, the Data Ethics Commission welcomes the proliferation of professional and ethical standards and codes of conduct published by associations and companies in the data-processing industry, with the proviso that these standards and codes must help to clarify exactly what needs to be done; CDR must not be reduced to a metaphorical fig leaf that allows companies to pretend that they are upholding the principles of digital ethics when the truth is very different.

1 Corporate Digital Responsibility Initiative: Shaping the digitalization process responsibly: A joint platform, 2018 (available at: *https://www.bmjv.de/SharedDocs/Downloads/DE/News/Artikel/100818_CDR-Initiative_EN.pdf?__blob=publicationFile&v=3*).

In the Data Ethics Commission's view, the data protection impact assessment that must (under the relevant circumstances) be carried out pursuant to the GDPR while a digital product is still at the development stage should be accompanied by a more comprehensive and general **societal impact assessment** focused on the assumption of foresighted responsibility (including the impact on any employees and customers of a company that are particularly affected by the digital transformation process) which also takes into account the long-term social effects of data-driven business models. It might be a good idea for companies commanding a large market share to set up an advisory panel (along the lines of consumer and customer advisory panels) that could be consulted when drawing up impact assessments of this kind; the panel should be made up of representatives of the groups of people most affected by the relevant business model.

3. Education: boosting digital skills and critical reflection

Digital self-determination presupposes digital skills. The Data Ethics Commission therefore unreservedly welcomes the efforts undertaken by the Federal Government, by consumer protection associations, by legal professional groups and by other bodies to raise public awareness of the importance of the **self-determined use of data and digital technologies** (from smartphone settings through to digital inheritance planning) and to provide straightforward and easy-to-understand information on the available options as well as practical guidance. It also welcomes the steps taken to raise awareness among consumers of the potential inherent to data, and to provide them with much-needed information about their rights and about the real opportunities and risks involved in the economic exploitation of their data. The Data Ethics Commission recommends that all of these efforts should be continued and stepped up.

School pupils should also be made aware of the issues connected with digitalisation as early as possible. Digital skills should be integrated into the **curriculum**, and teachers must be provided with comprehensive training on the subject at regular intervals. This is the only way to ensure that new generations will grow up to become competent "digital natives", able to assess both the opportunities and the risks of new digital applications, to take informed decisions and to assert their rights effectively.

In addition, **lifelong education** on the use of data and digital technologies must be provided to all age groups and social groups. It must be borne in mind that digital skills require not only a basic knowledge of the underlying technology (which, in turn, requires ongoing education in technical and mathematical subjects), but also an adequate familiarity with the economic, legal, ethical and social sciences; this broad spectrum of knowledge is necessary to comprehend, discuss and assess the various opportunities and risks in all their complexity.

Education and training in computer science, data science and software development is of particular relevance in this respect. As well as basic instruction on ethical and legal issues, more in-depth teaching on statistics, methodology and scientific theory is needed. It is particularly important to ensure that questions relating to data ethics and research ethics are embedded in discipline-specific methodological training, and there must be a major push in this area to ensure that ethical and legal considerations are incorporated into early-stage discussions by the parties that develop digital products and services or are involved in decisions on their development.

An essential first step towards achievement of these goals is cooperation between as many different entities as possible, including **government agencies, bodies that are closely aligned with the State and private actors** at federal, State (Bundesland) and municipal levels. The challenges involved in providing the general public with digital skills, maintaining these skills in the long term, and adapting them to each individual's lived experience are so great that they could never be tackled successfully by a single, centralised body. That said, a key role must be played by supervisory authorities (data protection authorities and/or the relevant specialist supervisory authorities), the Foundation for Data Protection, consumer protection associations and training providers. The media and institutions involved in media regulation also have a large part to play in this connection; they must not only provide society with information about the new technologies and cast a critical eye over technical progress, but also establish new forums for debate.

Although government agencies must remain chiefly responsible for imparting digital skills to the general public, this task cannot be realised in full unless the necessary **civil society structures** are put in place, such as digital volunteering, tech accountability journalism and consumer-focused market observation. The Data Ethics Commission therefore recommends that the Federal Government should provide long-term support for the establishment of structures of this kind.

Companies also have a responsibility to provide training to their staff. For example, a company can attain high ethical standards only if its employees (particularly those in management and in product development) have an adequate awareness of potential ethical and legal issues. As far as education and training is concerned, questions relating to data ethics and data law should also be included in a **broad spectrum of academic and professional training routes** and in workplace training. Particular attention should be given to technical and business professions, with a view to ensuring that ethical and legal considerations are incorporated into early-stage discussions by the parties that develop digital products and services or are involved in decisions on their development.

4. Technological developments and ethical design

Efforts to impart more advanced digital skills to the general population must not end up shifting the weight of responsibility away from manufacturers and digital service providers and towards users, not least because users have only limited opportunities to grasp and comprehend all the steps involved in the processing of their data and the underlying business models. Responsibility should be laid first and foremost at the feet of those who are able to exert an influence over the development of products and services. This concept is embodied in the principle of **ethics by design** or **ethics in design**, and appears in the GDPR (with reference to data protection and intrusions into the private sphere) under the heading of data protection by design and by default. Aligning the development of technologies and products (including services and applications) with the ethical values and principles outlined above is also a good way of increasing public confidence in digital products and acceptance of these products.

At the same time, however, the design of every product must be **tailored to the target user groups**. Involving user groups and their needs at an early stage of product development **(participatory product development)** may be helpful in this respect. It is particularly important for products that are targeted at vulnerable and/or less digitally literate user groups to have an **inclusive design**, including privacy-friendly default settings, with a view to protecting the digital self-determination of these user groups. Inclusive design allows manufacturers and operators to meet the constitutional requirement for informational self-determination as enshrined in Article 1 paragraph 1 of the German Basic Law (Human dignity), according to which protection must not be contingent upon individual capabilities and personal circumstances.

The most popular methods and platforms used to develop technologies, the most commonly used libraries and other code components have rarely supported the requirements of ethics by design to date. Components with a "better" design from the perspective of ethics or data protection law are at best a niche interest. There is a need for change in this area so that compliance with ethical principles in general and data protection principles in particular becomes the rule rather than continuing to be the exception. Ethics by design requires the gap between different communities to be bridged, and this has certain implications for the professions affected. The goals of this approach could be furthered not only by information on methods and catalogues, but also by **best-practice concepts, supporting tools, development frameworks** and **(open-source) code components**. Platforms with repositories of these components and usable pools of data (which, in some cases, are a necessary prerequisite for checks) would make it possible to highlight the specific properties required, supply the documentation needed and provide opportunities for exchanging know-how and experience.

Although ethics by design is a crucial governance instrument that allows the process of designing products, processes and services to be aligned with individual and public interests from the outset, it provides no guarantee that the resulting products and services will be ethical. Ethical principles can and should have a positive influence on technological developments, but **ethics is not a task that can be delegated to technology**. Furthermore, decisions about which ethical principles should be implemented and how they should be implemented (for example whether fairness metrics should be applied to algorithmic systems, and if so which metrics) should not be left to developers alone; instead, these decisions should be negotiated on a context-specific basis, if necessary with the involvement of the parties affected.

5. Research

Although data-processing systems with a more ethical design are frequently developed and showcased by researchers, there is a gulf between the world of research and the real world. One of the reasons for this may be the fact that some of these technical solutions (for example those based on cryptographic mechanisms) are counterintuitive in nature and more difficult for many people to understand than conventional methods; a prime example is a digital identification document that changes in appearance every time it is shown, making it impossible to "join the dots" between its holder's observed behaviours. Many people attempt to **understand** these innovative technologies by drawing on **conceptual models** from the surrounding (analogue) world, but these latter provide an insufficient basis for comprehending them or appraising their added value. Despite the advantages offered by these technologies in terms of ethics and data protection law, it is unlikely that their use will become widespread until the public gains a better understanding of them and is more confident in their use.

In many cases, **cross-cutting (and therefore interdisciplinary) cooperation** is an essential starting point for understanding the implications of new developments and designing ethical systems, but cooperation of this kind is not adequately rewarded by the discipline-bound metrics for good science and research. In many areas, interdisciplinary research will be given due recognition only if a shift in mindset occurs (this applies to universities, peer reviews and expert opinions, for example). Research funding should be funnelled towards interdisciplinary cooperation which delivers results that would have been impossible to achieve within the silos of the individual disciplines, and should allow the necessary institutional frameworks and long-term career paths to be established.

In many cases, high-quality and promising technical solutions have already emerged from the research sector, but the demand for these solutions is currently still lacking. There is also a need for methodologies or technologies that **signpost a route** from the current implementation status to an **improved state of technology**. Once again, **funding should be channelled into development and innovation** so that improved solutions can move from the drawing board to reality. Instead of providing support for only a few outstanding success stories, the need for broad-based progress in the field of ethical design must be acknowledged.

6. Standardisation

At the very latest when Lawrence Lessig coined the aphorism "Code is Law",[2] thereby emphasising the relevance of technical reality, it should have been obvious that **technical standardisation** is an essential factor in the implementation of legal and ethical requirements. Bodies responsible for the technical standardisation of communications networks have been established at international level (ISO/IEC, IEEE, IETF, ITU, ETSI or W3C), European level (CEN) and national level (DIN being the prime example in Germany, alongside other specific standards for public bodies). A technical standard by itself has no legal force, and anyone who uses a technical system must also comply with the applicable legislation, even if the provisions of this legislation run counter to the requirements imposed by a global technical standard. Nevertheless, standardisation is hugely influential in terms of what is available on the market; wherever possible, therefore, steps should be taken to avoid adopting standards that infringe the current legislation.

The standardisation process is often criticised for its lack of democratic legitimacy, and it is true that the groups within society that stand to be most affected are often deprived of any opportunity for **representative participation**. For example, non-governmental organisations or other civil society representatives are seldom involved in the standardisation process, and generally speaking even data protection authorities are only rarely involved in the standardisation of technical systems. In a worst-case scenario, this may mean that the operation of a technical system complies with the standards but violates the legislation. Another point of criticism is that a number of international standards that manufacturers or operators are supposed to comply with are not **available free of charge in the public domain**, but must instead be purchased.

Past standardisation efforts in the field of information security served as a major contributing factor to the addition of extra security features and gradual improvements in the level of security, for example of online banking. Yet the Snowden revelations made it clear that a number of intelligence services and government agencies were deliberately attempting to weaken standards by including security loopholes or backdoors as a way of safeguarding access in the future. The role of technical standardisation can be expected to gain in importance over coming years, for example as a result of the GDPR-imposed requirement to take due regard of state-of-the-art technology, or as a consequence of the German IT Security Act (*IT-Sicherheitsgesetz*). The political influence exerted by a number of different countries (not all of which are in Europe) can also be expected to increase.

An **impact assessment** of standards that are currently in existence or are still being debated must go beyond purely technical and economic considerations, and be **expanded** to include ethical and societal factors. The State should ensure that civil society actors, data protection authorities, consumer protection experts or spokespersons for organisations representing the parties affected can play a role in the standardisation process alongside the stakeholders that have dominated it to date.

2 Lawrence Lessig: Code and other Laws of Cyberspace, 1999.

7. Two governance perspectives: the data perspective and the algorithms perspective

In the following two parts, the arguments set out above are applied to data-based algorithmic systems on the basis of two different but complementary approaches. The **general ethical principles and precepts** used as a basis by the Data Ethics Commission (see Part B above) are important in two respects: firstly, they must guide data governance measures, in particular with a view to ensuring that procedures for collecting, accessing and using data are ethically sound; secondly, they must guide the design of algorithm-based systems used to process data (including the oft-cited "artificial intelligence" systems). The perspective that focuses primarily on data (the "data perspective") and the perspective that concentrates mainly on algorithmic systems (the "algorithms perspective") should not be regarded as competing views or two sides of the same coin; instead, they represent two different **ethical discourses, which both complement each other and are contingent upon each other**. These different ethical discourses are typically also reflected in different governance instruments, including in different acts of legislation.

The **data perspective** focuses on the data that are used to train algorithmic systems, as a basis for algorithmically shaped decisions, or for a plethora of other purposes specifically associated with the **context of meaning and the semantics of data** (Part C, section 2.1). In particular, it requires thinking about the origin of these data and the potential impact their processing may have on individuals involved with the context and semantic content of the data. From an ethical and legal perspective, it is important to identify standards for data governance; typically, however, the **rights** that these individuals can assert against others will play an even more significant role. A central distinction in this context is that between personal and non-personal data, since it determines whether the rights granted to data subjects under data protection law apply. Current debates that are pertinent in this connection include those on "data ownership rights" or open data, for example.

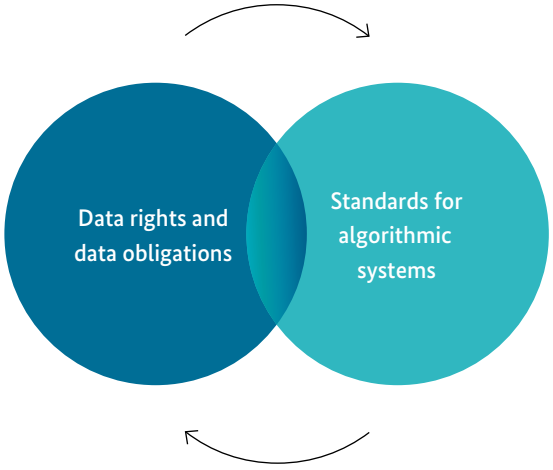


Figure 4:
Data perspective and algorithms perspective

By way of contrast, the **algorithms perspective** focuses on the architecture of data-driven algorithmic systems, their dynamics and the systems' impacts on individuals and society. The ethical and legal discourse in this area typically centres around the relationship between **humans and machines**, with a particular focus on automation and the outsourcing of increasingly complex operational and decision-making processes to autonomous systems enabled by artificial intelligence (AI). The algorithms perspective differs from the data perspective in that the data subjects affected by the system may not necessarily have anything to do with the original training or processing data; even if they do, they are not the focus of attention. The focus is on the **objective requirements** that apply, observance of which may be enforced and failure to comply with which may lead to liability and sanctions. The current debate on "algorithmic oversight" is relevant and important in this respect.

Part E

Data

Data provide access to information, information can lead to knowledge, and knowledge bestows influence and power. In the light of new capabilities of automated data processing and an exponential increase in memory and computing capacity, having access to data can mean an enormous increase in power and opportunities. Controlling important resources is inherently associated with a certain level of responsibility. Thus data, like other resources, may be used only for lawful and ethically acceptable purposes, and, like other resources, the impact of their use on individuals and the general public as a whole must always be assessed. Yet data also exhibit certain characteristics that differentiate them from other resources.

In the following sections, the Data Ethics Commission will therefore take these specific characteristics of data as a starting point, and develop, on the basis of the principles outlined in Part B and without claiming to be exhaustive, general standards of data governance (→ section 1 below) as well as data rights and corresponding data obligations (→ section 2 below). It will then set out specific recommendations for action in relation to standards for the use of personal data (→ section 3 below), improvements to controlled access to personal data (→ section 4 below) and general access to data, in particular non-personal data (→ section 5 below).

1. General standards of data governance

Any attempt to identify specific principles of data governance must start with the differences between data and traditional resources such as oil or goods. The unique characteristics of data include, in particular, the following:

- data are created and processed further in a **distributed and dynamic process**, through the interaction of a number of different players acting in very different roles (e. g. the data subject, the operator of a data-generating system, the developer); this process is, in principle, **never fully complete**;
- data are a **non-rivalrous resource**, i. e. they can be duplicated as often as necessary and used in parallel by multiple different players for multiple different purposes;
- data are **multifunctional and can be used across different sectors**, and the potential and risks inherent to them depend, to an exceptionally large extent, on each data controller's specific goals and opportunities and, in particular, given the importance of effects of scale, the ability to combine them with other data.

1.1 Foresighted responsibility

The special characteristics of data, such as their unusually dynamic nature and the unusually high context dependence of opportunities and risks associated with them, mean that there is a particular need for foresighted responsibility when making decisions about collecting, using or forwarding data. When assessing the potential impacts, including the risk of infringing the rights of third parties, particular consideration should be given to the following points:

- the **volume** of the emerging collections of data, with a particular focus on any cumulative effects, network effects or effects of scale;
- the **technological means** for processing data, with a particular focus on the technological options that are, or will be, available to large corporations and government bodies (especially in relation to the recombination and decryption of data);
- the **purposes** of data processing, with a particular focus on potential changes to the context of data use and the players involved (e. g. as a result of access by government agencies or following a corporate takeover).

In the case of personal data, the principle of foresighted responsibility has found its standardised expression in the maxims of data minimisation and storage limitation that are enshrined in the GDPR. A range of further duties under the GDPR, from the need to carry out a data protection impact assessment to mandatory requirements for controller-to-processor contracts, likewise follow from this principle.

1.2 Respect for the rights of the parties involved

The use of data must always be underpinned by respect for the rights of others. Acts or omissions that are ethically unacceptable or unlawful in general terms, because they violate the **rights of others**, do not become acceptable or lawful simply because they are committed by way of using data (e.g. fraud is a criminal offence regardless of whether it is committed by use of data or otherwise). As data are generated in distributed processes and through the interaction of many different players, parties who have in any way been involved in the process of data generation, for example as the data subject or as the owner of a data-generating device, may – from an ethical and possibly also from a legal perspective – be entitled to **genuinely data-specific rights (data rights)** in relation to these data (→ for further details, see section 2 below). Such data rights must be respected whenever data are used.

Respect for the rights of others implies much more than simply avoiding intrusion into legally protected spheres, such as another party's copyright. What is needed instead from an ethical perspective is in-depth **consideration** for the data-related legitimate interests of parties who are specifically linked to the data and who may therefore have certain rights of co-determination and participation concerning the data. This in-depth consideration may also imply duties to take action, for example by granting another party access to the data in certain ways.

In the case of personal data, the principle of respect for third-party data rights is expressed particularly clearly in the **principles of lawfulness, fairness and purpose limitation** enshrined in the GDPR. The GDPR itself sets out a number of data rights vested in the data subject, e.g. the right to be informed, the right to rectification, the right to restriction of processing, the right to erasure or the right to data portability.

1.3 Data use and data sharing for the public good

Resources that could be used to further key legally protected interests of individuals (e.g. health) or to promote the public good, particularly in pursuit of the UN's 17 Sustainable Development Goals relating to economic, social and ecological aspects, should not be neglected. As a basic principle, there is an **ethical imperative** to use these resources in cases where to do so would increase overall prosperity and where there are no overriding and conflicting interests of other parties (particularly data rights).

One of the special features that make data unique is that they are a non-rivalrous resource. They do not "wear out", even if they are used in parallel by many different players for many different purposes, and they can be duplicated an almost infinite number of times. **Sharing data** can mean that the player who first shares the data is at the very least no worse off, and everyone else involved (however loosely) is better off than they would have been had the data not been shared. An ethically responsible approach to data governance must take this fact into account. Data sharing is also enormously important in terms of safeguarding **fair and efficient competition**.

At the same time, however, conflicts can sometimes arise between the principle of furthering the public good by data use and data sharing on the one hand, and the principles of foresighted responsibility and respect for other parties' data rights, including considerations of appropriate investment protection, on the other. The creation of incentives for **voluntary data sharing** should therefore always be prioritised, and legislative requirements to share data should be the exception.

1.4 Fit-for-purpose data quality

Data, together with their context and semantics, are stored information. Information regularly purports to be the most accurate possible representation of reality as it currently stands, or the most accurate possible prediction of future reality. In situations that do not involve the automated processing of data by algorithmic systems, it is immediately obvious to everyone that incorrect information is not only worthless, but also potentially harmful; as soon as automation comes into play, however, it is all too common for people to fall prey to **false objectivity** and show a foolhardy willingness to rely on the results of calculations that were carried out using incorrect or incomplete data, and are therefore also likely to share these characteristics ("garbage in, garbage out").

In the interests of everyone, therefore, responsible data governance in the data society must also include efforts to achieve a **standard of quality that is appropriate for the intended purpose** (→ Part C, section 2.1.1). The meaning of "appropriate" must always be determined on a **context-specific** basis when used in relation to data quality, however. For example, it is important to remember that data may reflect societal preconceptions, stereotypes and discrimination, which will, in turn, influence the functioning of any algorithmic system trained using these data (→ for further details, see Part F, section 2.6). Data that accurately reflect an existing deficit may therefore be unsuitable for use as a basis for other purposes, even if they are of a high statistical quality.

Another important factor in this connection is that data can be used across different sectors and for different purposes. The **FAIR principle** (*Findable, Accessible, Interoperable, Reusable*) may be relevant in this context, for example as regards data storage and encoding methods. According to this principle, data must be prepared and stored in such a way as to be findable and accessible, and must be coded in an interoperable format and in a way that makes the data reusable in different contexts by as many different players as possible.

In the case of personal data, the desire to achieve a high level of data quality is manifested in the **principle of accuracy** enshrined in the GDPR.

1.5 Risk-adequate level of information security

Data can be freely duplicated, and it is **almost impossible** to recover them once they have gone astray. The wide range of possibilities for **external attack**, many of which are invisible from outside, mean that data are also vulnerable to malicious attempts to falsify or destroy them. A high level of **information security** that is commensurate with the relevant risk potential is therefore, from a technical perspective, directly related to the principles of foresighted responsibility and respect for the rights of the parties involved. Appropriate information security, encompassing a broad spectrum of measures at different levels, is a vital prerequisite for mutual trust on the part of those involved in the data society.

In the case of personal data, the concept of information security is manifested in the **principle of integrity and confidentiality** enshrined in the GDPR.

1.6 Interest-oriented transparency

Since a party that uses and effectively controls data may gain influence and power as a result, this party must, in principle, be able and willing to **account** for its actions. One of the reasons for this is the protection of parties whose data rights might be affected or even violated. An **interest-oriented level of transparency** is required so that these parties (or entities enforcing data rights or data law for the benefit of others) can determine whether and to what extent data rights have, in fact, been affected or violated, and against whom they can lodge claims.

In the case of personal data, transparency – i. e. ensuring that data processing operations are easy **for data subjects to understand** – is a basic principle of the GDPR, and the same is also true for the **principle of accountability**. Many of the provisions of the GDPR, for example those relating to information, documentation and the right to request access, are designed to improve transparency.

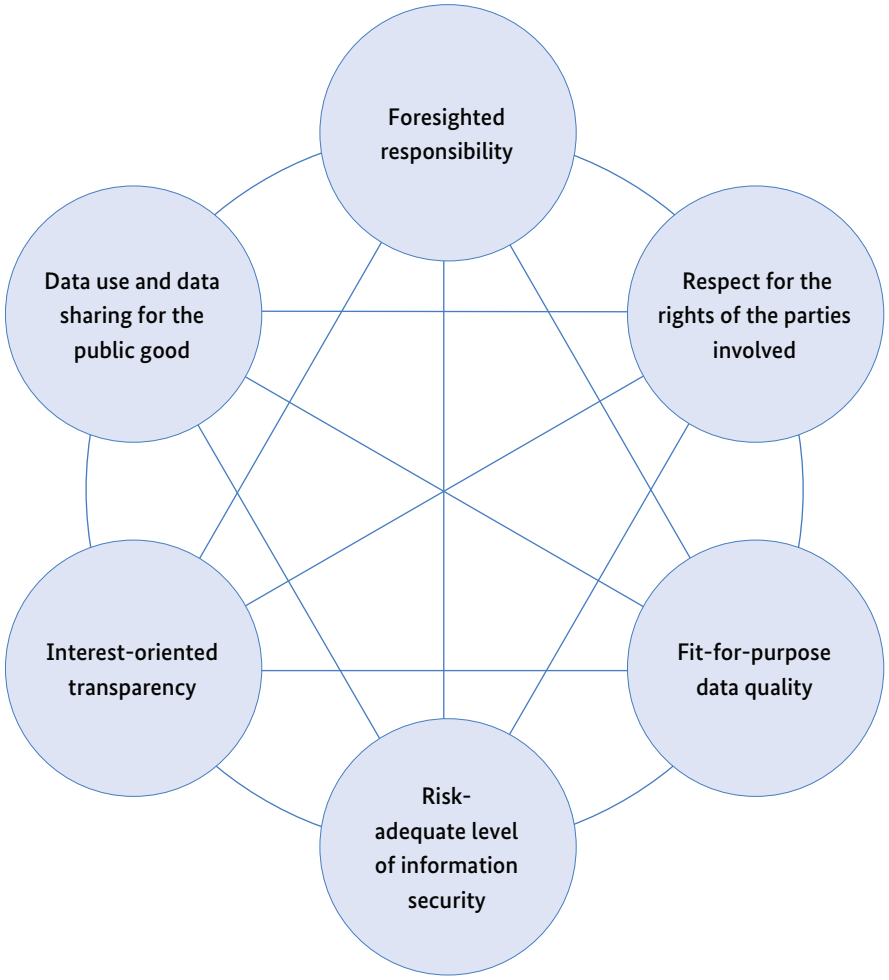


Figure 5: Standards for data governance

2. Data rights and corresponding obligations

According to the ethical principle of digital self-determination, individuals should not merely be perceived as being passive and in need of protection and as facing actual or potential threats, but rather as **self-determined actors in the data society**. Self-determined navigation of the data society by individuals requires that these individuals have certain rights that can be asserted against others. First and foremost among these rights are those which relate to an individual's **personal data**, which derive from the right to informational self-determination that is enshrined as a fundamental freedom, and which are guaranteed by the data protection law currently in force. Digital self-determination also encompasses the self-determined economic exploitation of one's data and the self-determined handling of **non-personal data**, for example the data generated by the operation of one's devices. The Data Ethics Commission takes the view that, in principle, a right to digital self-determination also applies to companies and **legal entities** and – at least to some extent – to groups of persons (collectives). In this context, the Data Ethics Commission believes that it is possible to identify general principles underpinning data rights and obligations that go beyond data protection alone.[1]

2.1 General principles of data rights and obligations

Complex data generation processes (understood in the broader sense, i.e. including various phases of data creation, enhancement and refinement) often involve interactions between different parties that may be pursuing different goals and playing different roles and that contribute, in their respective roles, to the generation of data in the process. A **contribution by a party** (i.e. a natural or legal person) **to the generation of data** may be relevant if any of the following are true:

a) the information stored in the data relates (in terms of meaning) to the party or to an object associated with this party (e.g. belonging to him or her);

b) the data were generated by an activity of that party or by the operation of an object (e.g. a sensor) that belongs to this party; or

c) the data were generated by software or another component (e.g. sensors) created by or invested in by this party.

Where the situation referred to in a), i.e. the situation that a party is the subject of the information stored in the data, relates to natural persons, this is of particular significance since this situation gives rise to the right to informational self-determination and data protection enshrined in constitutional law.

Given the specific characteristics of data and the inextricable link between personal data and personality rights, the Data Ethics Commission believes that a contribution to the generation of data should not give rise to exclusive ownership rights in said data, above and beyond the existing intellectual property rights (→ see sections 3.3.2 and 5.2.4). Instead, a contribution to the generation of data should entitle a party to specific data rights in the form of **co-determination and participation rights**; these rights in turn impose obligations on other actors. From an ethical perspective, this will result in a **dynamic and special relationship** between a party involved in the generation of data and the party controlling the data. The duration of this relationship may vary, as may its intensity. As far as personal data are concerned, the relationship will largely be determined by the applicable data protection law.

From an ethical perspective, the **recognition and design** of data rights, and corresponding data obligations, in dynamic environments depend on the following general factors, which are normally also the factors underlying relevant legal provisions where data rights and obligations have already been substantiated in the law:

a) the scope and nature of the **contribution to data generation** by the party asserting a data right;

1 Model of data rights and data obligations based on Preliminary Drafts no. 2 (February 2019) and no. 3 (October 2019) of the "Principles for a Data Economy" by the European Law Institute (ELI) and the American Law Institute (ALI), made available to the Data Ethics Commission. These preliminary drafts have not yet been adopted by either the ALI or the ELI and do not yet represent the official position of either of these organisations.

b) the **weight of that party's legitimate interest** in being granted said right (in particular the right to require desistance, access, rectification or an economic share);

c) the **weight of any possibly conflicting interests** on the part of the other party or of third parties, taking into account any potential compensation arrangements (e.g. protective measures, remuneration);

d) the **interests of the general public**;

e) the **balance of power** between the party asserting the data right and the other party.

These factors interact with one another in what can be described as a flexible system; if the public interest in data access is particularly high, for example, it may compensate for a relatively insignificant contribution to data generation. Consideration must always be given to the general principles outlined in Part B in order to avoid situations in which crucially important individual interests are undermined by a purported or actual public interest. These factors also determine how certain details (e.g. formats, deadlines, protective measures or financial compensation) should be **fleshed out and put into practice**. This includes the question of whether action should be taken only upon request by the party asserting the data right (e.g. data access claim) or also proactively (e.g. an obligation to publish data).

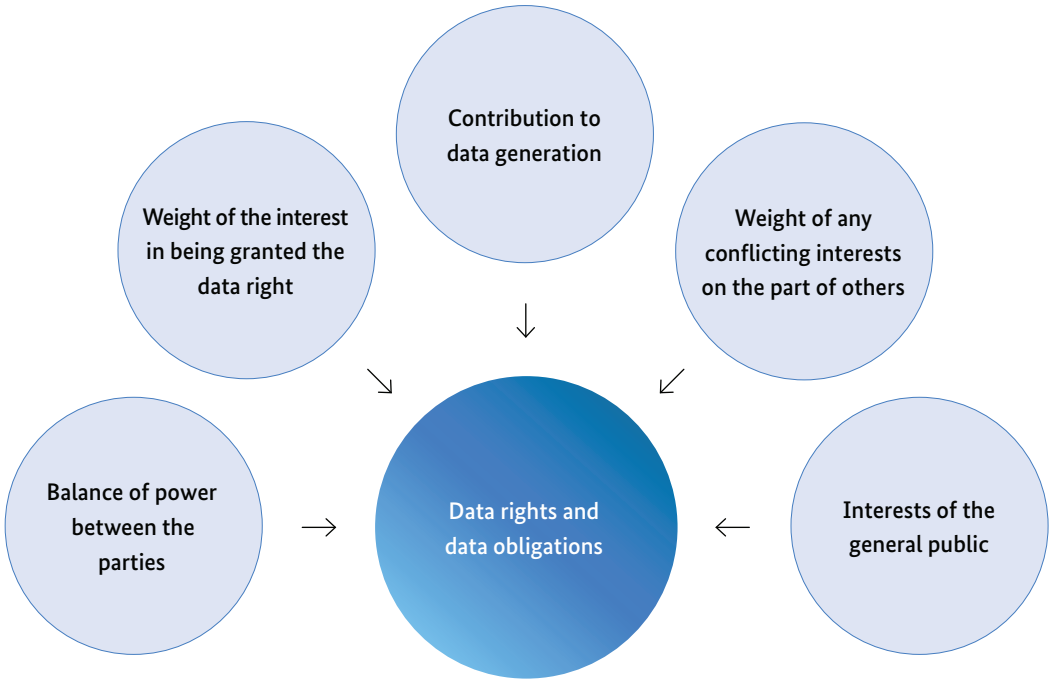


Figure 6: General factors for the shaping of data rights and the corresponding data obligations

The **rights granted to data subjects** by the GDPR are a particularly important manifestation of these principles, aimed specifically at protecting the natural persons to whom the information pertains; they are also to some extent a standardised manifestation given that they hinge on the qualification of data as personal data. The principles formulated here can also be applied to non-personal data, however, and relate not only to individuals, but also to legal entities and collectives.

2.2 Clarification of the general principles with reference to typical scenarios

Data rights may have a number of different **goals**; these include obliging another party to desist from using the data (up to requiring erasure of the data), gaining access to the data (e.g. disclosure, transfer, full portability), arranging for the data to be rectified, or claiming an economic share in the profits derived with the help of the data.

2.2.1 Scenarios involving desistance from use

Situations often occur in which a party requests another party to desist from using data in a certain way. The GDPR even works from the basic assumption that (personal) data should not be used unless there is a legal basis for doing so and a number of other requirements have been met.[2] In a general sense and beyond the scope of the GDPR, if a party has a significant legitimate interest in the controller desisting from data use, the outcome (from an ethical perspective) may be a **right to require said desistance**, potentially even including a right to erasure of the data, where the data processing operation:

a) might cause harm to that party or to a third party; and

b) is inconsistent with the circumstances under which that party contributed to generation of the data, in particular because

   (i) the contribution was made for another purpose, and the party could not reasonably have been expected to contribute to the generation of the data if it had foreseen the present data processing operation; or

   (ii) consent by that party would be invalid for overriding reasons.

Before any such right to require desistance from use can be affirmed, however, the party's legitimate interest in being granted the right must be weighed up against the other factors referred to above (→ in section 2.1). For example, such a right cannot be affirmed in cases where the processing of data is, by way of exception, justified by compelling other interests (e.g. the prosecution of criminal offences).

2 Article 6(1), Article 9(1) GDPR.

With regard to **non-personal data**, requests to desist from the use of data may become relevant, for example, in the context of value creation chains and customer relationships where non-personal data are often of enormous economic significance and a party involved may have a significant legitimate interest to assert such a right (→ section 5.3 below).

Example 1

The non-personal data collected by sensors in modern agricultural machinery (relating to soil quality, weather, etc.) are used by manufacturers as a basis for many of the services they provide (precision farming, predictive maintenance, etc.). If the manufacturers were to forward these data to potential investors or lessors of land, however, the latter would be given information that might prove harmful to an agricultural holding if negotiations over the land were to take place in the future. It can be assumed that the agricultural holding would not have helped to generate the data voluntarily had it known that they would be used for this purpose. When assessing a right to require desistance from an ethical perspective, consideration must be given to the balance of power between the parties in the case at hand, and also to the fact that the agricultural holding made an extremely significant contribution to generation of the data. Third-party rights deemed worthy of protection would include only the manufacturer's interest in maximising their profit and a general interest on the part of investors, lessors, etc. in obtaining accurate information.

From an ethical perspective, a **waiver of a data right** to require desistance is possible only under very limited circumstances. Such a waiver should automatically be ruled out in cases where consent to data use would be invalid for overriding reasons (within the meaning of requirement b) (ii)), for example because it is illegal or inconsistent with public policy; this is because, under our legal system and the fundamental values underpinning it, there exists no such thing as a liberty to do any kind of harm to oneself or to others. In other cases, a waiver may be possible, provided that stringent requirements are met (e.g. there is a separate agreement that is not linked to other services and does not involve the party being placed under pressure) to ensure the voluntary nature of the waiver, meaning that requirement b) (i) would no longer apply.

In Example 1, the agricultural holding could consent to the data being forwarded to third parties, e.g. on the basis of an individual agreement with appropriate remuneration; use of the tractor should not be dependent on the data being forwarded.

For **personal data**, obligations to desist from data use normally follow already from the provisions of data protection law, but the criteria outlined above can be used to determine whether **the substantive limits of consent** have been exceeded (→ section 3.2.1 below) or to guide the balancing of different legitimate interests, for example.

Example 2

Data relating to the activities of a social network user are used for extensive personality profiling; the profile contains the attributes "mentally unstable" and "esoteric tendencies". As a result, the user is shown advertisements by companies that offer personal horoscopes or energy healing services (at significant cost) on an almost daily basis and often immediately after he has posted content that signals stress or anxiety; he often makes purchases as a result. When he set up his user account, he clicked on a checkbox next to the following statement: "I am happy for my data to be evaluated so that my personal preferences and attributes can be identified more accurately and the services offered to me (including by third-party providers) can be personalised to my needs (profiling)." "Consent" of this kind does not make the subsequent data processing operations lawful. There are a number of different arguments for reaching this conclusion: one of them being that processing the data for this purpose may cause significant harm to the user, which would be inconsistent with the circumstances under which he generated the data (because he could not reasonably have been expected to do so had he known that data would be used for this purpose, and because the law does not allow the abuse of mental states of this kind, cf. Section 138 of the [German] Civil Code (Bürgerliches Gesetzbuch, BGB).

There are many circumstances under which an obligation to desist from the use of data cannot be mitigated by consent or a balancing of conflicting interests; in such cases, reference is often made to "red lines" or "**absolute limits**". There is no requirement for these limits to be data-specific, and most are not. For example, it is reasonable to prohibit election manipulation practices that are incompatible with the principle of democracy, regardless of whether said practices involve the use of data. In the view of the Data Ethics Commission, an example for data-specific absolute limits is the total surveillance of individuals.

Example 3

When entering into an employment contract, an employee signs an agreement stating that the location tracking functions on her smartwatch and mobile telephone, as well as a number of apps that collect data (e.g. by tracking sleeping behaviours and emotions), will be kept switched on at all times, even when she is not at work, and that she will hand the devices over to her employer when requested in order for the relevant data to be accessed. It is readily apparent that these arrangements, taken together, are equivalent to total (or almost total) surveillance, which is incompatible with human dignity, self-determination and privacy. This is true even if the employee gave consent to each of these measures, even if she decided of her own accord to enter into a contract with this employer, and even if there were other offers of employment available to her.

Conversely, the criteria that apply to scenarios involving desistance from use may also bear an indirect relevance to situations in which there is an ethical or even legal **obligation to use** data; such an obligation may arise where a party is under a general obligation to protect certain legally protected interests and has, at the same time, access to data that could be used to secure or improve the protection of these interests. In this kind of situation, an obligation to use data arises as the corollary of an obligation to protect certain legally protected interests unless a third party has a conflicting right to require desistance from data use.

Example 4

A hospital is experiencing an outbreak of a multi-resistant pathogen. It wants to analyse the health data of patients who have recently become infected in order to gain a better idea of why certain individuals are more likely to fall prey to the pathogen, as a basis for pinpointing the inpatients that might benefit most from a move to another hospital. Under these circumstances, the hospital has a general obligation to provide new patients with the best possible protection against infection by taking all available and reasonable precautions to this end. This includes the use of health data belonging to patients who have already been infected with the pathogen, provided that said use might protect new patients and there is no obligation emanating from the former group of patients to desist from use of their data.

2.2.2 Scenarios involving access to data

When it comes to scenarios involving a request for access to data, there will be many situations in which the party seeking access to data and the party who effectively controls the data will be able to reach an agreement on the action to be taken. **Voluntary arrangements** of this kind should be welcomed, provided that there are no conflicting and overriding third-party or public interests, and in particular provided that there are no parties with a right to require desistance from use based on the above criteria. Given the enormous potential for value creation inherent to data, however, in-depth discussions are also being held on the circumstances and conditions under which access to data should or even must be granted from an ethical viewpoint.[3]

This may apply in situations in which access to data is required (and perhaps even mandated by law) in order to enable a party to comply with a special **obligation or task** (e.g. prosecution of a criminal offence, public health concern). Any such data access right must be consistent with the rules that apply to this obligation or task; particular attention should be paid to the **principle of proportionality**, and any potential third-party rights to require desistance from use (→ see section 2.2.1 above) must be considered.

There may also be independent requests for access to data, for example within **existing value creation systems**. Such systems typically involve many different parties who contribute to the generation of data in different roles (e.g. as suppliers, manufacturers, retailers or end users), and who are, in principle, familiar with and have agreed to both their own roles and the roles of the other players involved (→ see section 5.3 below for further details). Legitimate interests that can be asserted by a party as a basis for an access request may, in particular, include cases in which the data are required for the following purposes:

3 By way of examples: European Commission: Building a European data economy, COM(2017) 9 final, 10 January 2017, pp. 11 et seqq. (available at: *https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-9-F1-EN-MAIN-PART-1.PDF*); European Commission: Towards a common European data space, COM(2018) 232 final, 25 April 2018, pp. 8 et seqq. (available at: *https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-232-F1-EN-MAIN-PART-1.PDF*).

a) to use an asset in line with its intended purpose within the value creation system (e.g. repair of a connected device by the end user);

b) monitoring and improving the quality of a service provided within the framework of the value creation system (e.g. by a supplier);

c) to ascertain the truth or provide evidence (e.g. in a legal dispute with third parties);

d) to avoid anti-competitive effects (e.g. lock-in effects); or

e) to create new value using the data (e.g. by developing a smart service).

Example 5

A supplier provides the engines for the agricultural machinery referred to in Example 1. It would be extremely useful for the supplier to have access to certain tractor data so that it can verify and constantly improve the quality of its engines. These data are stored in the manufacturer's cloud, however, and the latter is unwilling to allow the supplier to access them. In situations of this kind, it is important to remember that the supplier has made a significant contribution to the generation of the engine data, and that the data are urgently needed to improve the quality of a service provided within the framework of the same value creation system in which the manufacturer is also involved. Consideration must be given not only to the balance of power in the specific case at hand, but also to the fact that all parties involved – including the general public – have an interest in high-quality engines. There may, however, also be relevant economic interests on the manufacturer's side, in particular relating to confidentiality.

Access rights are also being discussed in situations where the party seeking access and the party that effectively controls the data are not yet part of the same value creation system, but where a **new value creation system** could originate in which they are both involved. The outcome of an assessment based on the general criteria will normally be different in situations of this kind, if only because the party seeking access has not typically contributed to the generation of the data, and the justifications that can be cited for granting an access right are rather **public interest considerations** or specific considerations, such as safeguarding **competition** (→ see section 5.5 below for further details).

Example 6

In Example 1, the manufacturer (which holds a dominant position in the tractor market) has been collecting soil and weather data for decades. A start-up recognises the potential of a database for investors using these data, and requests access to them. In this case, consideration must be given to the fact that the start-up itself has not made any contribution to the generation of the data. The existence of a public interest in data access (and the significance of this interest) depends on whether the manufacturer is abusing its market power and on how much the European economy would benefit from the breaking up of a small group of market-dominant companies (presuming that the start-up is based in Europe). In any case, potential harmful effects of data disclosure on trade secrets and other legitimate third-party interests, such as the interests of the manufacturer and the agricultural holdings in Example 1, must be taken into account.

The generally recognised principles of **open government data (OGD)**, which embody the idea that government data should be made available to the private sector, include "open by default" and re-use of data "by anyone for any purpose".[4] There have been calls from many quarters to expand these open data concepts to include data created by and effectively controlled by private entities. The move towards open data, however, also gives rise to complex ethical questions, for example the extent to which a generalised assessment that no longer looks at the individual case is acceptable.

Example 7

A municipality implements a large-scale project to collect mobility data using smartphone signals, with a view to facilitating traffic management (by adjusting the timing of public transport services, for example). Theoretically speaking, the data are "anonymised"; if the data sets are combined with other data sets and some additional knowledge, however, the owner can be identified with a confidence level of 95%. A number of different parties are interested in gaining access to these data; they include a researcher who wants to use them as a basis for identifying the optimal design of urban recreational areas, a start-up that wants to establish an online detective agency via which users can pay to access the mobility profile of their spouse, competitor, etc. and a research institute tasked by a foreign government with investigating the political activities of its citizens. Case-by-case assessments of these three access requests would deliver very different outcomes. It is therefore a difficult question whether the municipality may, or even must, make these data public with a view to the many possible uses of the data that would promote the public good.

The Data Ethics Commission wishes to emphasise, in this context, the importance of the (potential) rights of individual parties who have contributed to data generation, in particular the rights of data subjects, to require desistance from data use. It follows not only that all possible and reasonable protective measures (including anonymisation techniques, to be improved on an ongoing basis) should be taken after weighing up the potential for harm and the expected benefit for the public good, but also that – depending on the potential for harm – the granting of blanket access may be out of the question (→ see section 5.4 below for further details).

2.2.3 Scenarios involving rectification

Not all data are of a high quality. Problems that are particularly likely to arise include an unsuitable context, **inaccurate** encoding or **incomplete** data in the sense that any deductions obtained using the data are also **incorrect**. In circumstances of this kind, a party involved in the generation of data may have an ethically justified right to require rectification of the underlying data or of the deductions obtained using the data. The threshold for a right of this kind to be granted is relatively low, since in principle there is neither a protected individual interest nor a public interest in the processing of inaccurate or incomplete data. As a general rule, only the following requirements must be met:

a) the processing of inaccurate or incomplete data must be potentially harmful to a party (in particular the party to whom the information relates); and

b) the rectification must not be disproportionate, taking into account the severity and likelihood of harm on the one hand and the effort involved in rectifying the data on the other.

4 See Recital 16 of Directive (EU) 2019/1024 on open data and the re-use of public sector information (PSI Directive); Principles 1 and 3 of the G8 Open Data Charter signed at the G8 Summit on 18 June 2013; and Principle 1 of the International Open Data Charter signed in September 2015 at the Open Government Partnership Summit.

Example 8
A very high error rate has been detected in the engine data stored by the manufacturer in Example 5. This is problematic for the company that supplies these engines, not only because it deprives the company of the possibility to fulfil its quality assurance remit, but also because these engine-related data are pooled with engine-related data from other engine suppliers as a basis for evaluations, and poor performance metrics for the engines from the relevant supplier might reduce the latter's chances of securing orders from other manufacturers. In this case, the processing of inaccurate data causes harm to the supplier, and there are no indications that the effort involved in rectification would be disproportionate.

If the amount of effort involved in rectifying the data is excessive but the potential for harm is significant, a right to require desistance from use will frequently arise (→ see section 2.2.1 above).

2.2.4 Scenarios involving an economic share

Cases where a party uses data to create value after other parties have contributed to the generation of said data are an everyday occurrence, and a good thing in principle. Provided that no one is entitled to a right to require desistance from use (→ see section 2.2.1 above), such use of the data must normally be tolerated by the parties who contributed to their generation. Given the strong affinity which the data rights and obligations set out in this section have with **considerations of public good**, there are potent arguments against recognising a general right to remuneration for all parties who have contributed to the generation of data. Instead, such parties must content themselves with existing mechanisms of collective economic participation, in particular through the taxation of value creation.

In cases where there is no valid contract to back up a claim for remuneration, financial compensation should at most be considered as a mitigating measure, for example if the exercising of a data right without compensation appears disproportionate in the specific case at hand (→ see section 2.1 above, factor c). From an ethical perspective, and in the view of the Data Ethics Commission, a party who has contributed to the generation of data should be entitled to **independent remuneration** for their use by others only in very **exceptional cases**. Cases of this kind might arise if:

a) the party's contribution to the generation of data required an unusual amount of **effort** or was **particularly unique**, and it would hardly be possible (from an economic viewpoint) to replace it with contributions by other players; and

b) an exceptionally **large amount of value** has been created using the data; and

c) the circumstances under which the contribution to data generation was made mean that it would have been **impossible or unreasonable** for the party to engage in negotiations on any **remuneration**.

The amount of any remuneration paid in such exceptional cases must be adequate; in particular, basic incentives of using data to create value must not be removed. It must also be remembered that the party creating the value has typically incurred financial risks.

2.3 Collective aspects of data rights and data obligations

An answer must be found to the issue of whether (and if so to what extent) the above arguments concerning the right to require desistance from use, the right to access data, the right to rectification and the right to an economic share in profits derived with the help of the data can also be applied to **collectives** in the sense of defined groups of persons (e. g. indigenous peoples with regard to the use of their genetic data), i. e. whether collectives may be entitled to certain data rights in connection with the use of "their" data. For example,

thought must be given to the question of whether – ethically speaking – a population (of a nation state, or of the EU) which has generated data should have a right to an economic share in profits, such as in the form of taxes or transfer payments. The Data Ethics Commission believes that this question can, in principle, be answered in the affirmative.

Example 9

An Internet giant earns billions from the data generated when individuals all around the world use its services. Yet even though this megalith of a company generates 10-digit sums year on year using data from EU-based individuals, it pays virtually no taxes in the EU. The question arises whether the company should be obliged on ethical grounds to allow the general public in the EU to share (through taxation) in the value it creates. The issue raises fundamental questions about distributive and participatory justice, and about what a just economic system looks like. However, aspects such as market power and the unique nature of contributions (e.g. if audio data in a certain language is used to develop new voice-controlled services) may also have to be taken into account.

The **relational nature of many data types** makes it particularly important to include groups and collectives in any debate. This relational nature is apparent from the way that many digital services require users to disclose data about their contacts or "friends", for example. As far as data rights and the corresponding obligations are concerned, the "friends" may have the right to require desistance from use of the data and the right to gain access to the data, etc.; at the same time, their potential interests must always be taken into account when weighing up whether a data right should be granted (→ see section 2.1 above). However, there are also cases in which a party contributes to the generation of data, and these data then indirectly provide **information on other parties** – even if the latter played no role (not even in the broadest sense of the word) in their generation. This is particularly relevant in the sphere of genetic data, but also applies to other data types. There is still another, closely related group of cases, where individualised data (even in aggregated form) may have implications with potentially negative **third-party effects** that extend beyond the individual who supplied the data.

Example 10

A health insurance company offers reduced premiums as an incentive to sign up for health tracking schemes. While those who agree to disclose their data will benefit from lower premiums, those who refuse to do so may end up paying more.

Issues relating to the **representativeness of data** used to train algorithmic systems can also be interpreted as problems of relationality: the lack of any relationship between the parties who supply the training data and the parties to whom the trained systems are applied may result in systematic bias and potential discrimination (→ see Part F, section 2.6 for further details).

To overcome this hurdle, individualistic approaches to data rights in ethics, law and technology design must be expanded to include **relational concepts of data rights** (cf. also the debate on group privacy). Under certain circumstances, it may therefore be possible – at least when viewed through the lens of ethics – for one group member's contribution to data generation to be attributed to the other group members as well, potentially entitling these latter, in spite of the fact that they themselves made no individual contribution, to certain rights of their own (the right to request desistance from use or the right to gain access, for example).

3. Standards for the use of personal data

3.1 Personal data and data relating to legal entities

Any information relating to an **identified or identifiable natural person** is regarded as personal data. An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identify of that natural person (Article 2(1) GDPR).

Even though the remainder of this section focuses on personal data in the legal sense of the term, the Data Ethics Commission wishes to stress that the **protection of companies and legal entities** is a valid concern that should not be relegated completely to the sidelines. The potential hazards confronting legal entities have been exacerbated yet further by the networking of all machines, the exchange of data between factory components, and the storage of all production data generated by Industry 4.0 plants in "digital twins". If individual sets of data (generated through the operation of devices, for example) are pooled together, the result may be an almost seamless overview of a company's internal operating procedures, which may – in the absence of appropriate protective mechanisms – easily fall into the hands of the wrong parties outside the company (competitors, negotiating partners, authorities, prospective buyers, etc.). The Data Ethics Commission believes that the risk posed not only to the digital self-determination of companies and legal entities but also to the **digital sovereignty of Germany and Europe** (since data flows predominantly involve third countries) is concerning from an ethical viewpoint, and that steps must be taken to mitigate against it.

A key legislative starting point for protecting enterprise data is the **protection of trade secrets**, in particular the [German] Act on the Protection of Trade Secrets (*Gesetz zum Schutz von Geschäftsgeheimnissen*, GeschGehG). When interpreting and applying this Act, efforts must be made to guarantee the comprehensive protection of sensitive business data, given the central importance of the latter in building a fair and competitive economic system as the basis for economic and social well-being. In many respects, however, Directive (EU) 2016/943 (the provisions of which were transposed into the Act on the Protection of Trade Secrets) is not adequately tailored to the reality of IoT and Industry 4.0. The Data Ethics Commission therefore calls on the Federal Government to **step up data-related protection of German and European companies**.

The recommendations for action relating to personal data put forward by the Data Ethics Commission in the remainder of this section, for example in relation to a risk-adequate interpretation of the applicable legal framework (→ section 3.2.2 below) or privacy-friendly design of products and services (→ section 3.6 below) also apply to the protection of data relating to companies and legal entities (in a modified or attenuated form where appropriate).

3.2 Digital self-determination: a challenge to be tackled by the legal system as a whole

3.2.1 Cooperative relationship between the applicable legal regimes

Our economy and society are heavily reliant on the use of personal data in a huge variety of different contexts, and yet there is always a degree of tension between this use of personal data and the fundamental rights of individuals. The constitutional right to informational self-determination (as part of the general right of personality) is essentially part of the protection of human dignity. **Data protection law**, in particular the GDPR, clarifies these benchmarks and has binding force on public and private bodies.

The GDPR is one of the great achievements of the EU legislator, and currently functions as a source of inspiration for other countries. It is important to temper our expectations of this piece of legislation, however; the GDPR is focused on data protection rather than on comprehensive promotion of individual welfare and the public good in the data economy. Taken in isolation, it is not a suitable tool for averting all the harm that an individual may suffer as a result of his or her personal data being processed, and cannot therefore be regarded as protecting his or her integrity in all respects. All of the different mechanisms provided by the legal system as a whole must be used to safeguard these legally protected interests, particularly those that are **not specifically addressed by the provisions of data protection law** (e.g. economic interests, the right to life and health, physical integrity and reputation). This applies even in situations where personal data are at play.

The **concept of consent** that is enshrined in **data protection law** is a vitally important mechanism for safeguarding informational self-determination in the digital and analogue spheres. Yet the concept of a right to self-determination that is not subject to substantive limitations and that includes the freedom to inflict any kind of harm on oneself or third parties would be an alien element in our legal system, and is ethically indefensible. The law should limit or even prohibit an individual's free and informed consent – as an expression of his or her general freedom of action, which is protected as a fundamental right – only in narrowly defined exceptional circumstances. However, consent under data protection law should be subject to substantive limitations, by way of analogy to the limitations to freedom of contract or to consent when it comes to intrusions on bodily integrity.

In the view of the Data Ethics Commission, it has become clear that the average individual is **systematically overwhelmed** by the number and complexity of the decisions that he or she is required to take in connection with consent under data protection law, and by the difficulty involved in estimating all the potential impacts of data processing. The Data Ethics Commission believes that inadequate use of consent by providers of digital services is one of several reasons for a general **loss of trust** in the digital society. As things stand, individuals can often no longer rely on the fact that the State and the legal system have put in place the framework conditions necessary for them to navigate the world in safety and (relatively speaking) free from care, without needing to worry about the possibility of suffering serious harm from other parties. For business-to-consumer transactions, contract law, and more specifically unfair contract terms, control has provided the basis for 'rational indifference' on the part of consumers and for far-reaching protection even in low-value cases. The same result should be achieved by way of applying the **fairness test to declarations of consent**.[5] In applying the fairness test, general values and principles underlying the legal system as a whole must be taken into account.

3.2.2 Risk-adequate interpretation of the applicable legal framework

The Data Ethics Commission wishes to stress that the existing legal framework must be interpreted and applied in such a way as to mitigate to the maximum the new hazards we are facing in connection with the widespread collection, use and analysis of personal data.

Notwithstanding the need to comply with the requirements of data protection law, data processing operations are also subject to a number of **absolute limits**. Wherever possible, any uses of data that go beyond these limits should be prevented by interpreting and applying the law in force[6] in a manner consistent with fundamental rights. In the view of the Data Ethics Commission, this is relevant, for example, for:

5 Cf. also Recital 42 of the GDPR.

6 This relates in particular to the fairness test applied to general terms and conditions of business (Sections 307 et seqq. of the [German] Civil Code (*Bürgerliches Gesetzbuch*, BGB)), the principles of public morals (Section 138 of the Civil Code), wilful immoral damage (Section 826 of the Civil Code) and contractual and quasi-contractual protection and fiduciary duties (Section 241 paragraph 2 of the Civil Code).

- **incursions into personal privacy and integrity** that are incompatible with fundamental rights and that result from profiling and/or scoring (e.g. certain methods of determining personality traits, emotions or expected behaviours);

- **total surveillance** that is incompatible with human dignity, *inter alia* through a "comprehensive surveillance footprint" or "super scoring";

- **immoral exploitation** of situations of urgent need or of medical conditions;

- **election manipulation practices** that run counter to the principle of democracy.

The legislation currently in force already categorises ethically reprehensible attempts to mislead or manipulate consumers in a commercial context – which should include business practices aimed at persuading the party to disclose his or her personal data – as **misleading or aggressive commercial practices** under the [German] Unfair Competition Act (*Gesetz gegen den unlauteren Wettbewerb*, UWG), regardless of whether the provisions of data protection law have been infringed; any such attempts will therefore trigger the appropriate legal consequences (e.g. rescission on grounds of fraud or threat, injunctive relief and compensation). The Data Ethics Commission wishes to cite the following as potential examples of such practices:

- **addictive designs**, i.e. technologies which exert undue influence on a user (in particular by means of mechanisms that promote addictive behaviour) and which are therefore liable to have a substantially adverse impact on his or her freedom to decide whether to use them (and stop using them);

- **dark patterns**, i.e. technologies (mainly user interfaces) that are designed in such a way as to deceive a user about certain facts and/or manipulate him or her into taking a certain decision (which may have financial implications).

Absolute limits must also be imposed on data processing in order to protect individuals against being placed at an **undue financial disadvantage**, and the existing legislation contains various provisions that can be used to enforce this protection.[7] In the view of the Data Ethics Commission, examples of unfair contract terms and violations of contractual or pre-contractual duties of a fiduciary nature include the following:

- preventing access to data that have been generated by a device and that are required for normal **use** of said device, including for the performance of repairs by an independent workshop, or making it unreasonably difficult to access these data (e.g. access only granted in accordance with Article 12 GDPR, i.e. only within one month or even three months);

- preventing access to the data needed to operate a **pre-owned** networked device, or making it unreasonably difficult to access these data (e.g. for an individual who has bought a house equipped with smart home technology);

- making it harder for individuals to switch provider by means of **data lock-in** (i.e. refusing to hand over data analyses for which the user has already paid from an economic perspective, and which are not protected trade secrets);

- processing user generated data by a manufacturer or another member of the supply chain and for a purpose that runs completely counter to the user's **economic interests** (e.g. price differentiation with the aim of extracting the maximum from each individual that he or she is willing to pay).

7 Cf. the instruments referred to in Footnote 6.

Social media monitoring

Social media monitoring is the systematic **oversight** of social media content on a particular topic. It has evolved into a data utilisation tool that takes advantage of the fact that social networks not only expand users' communication options but also allow their digital behaviour to be constantly monitored.

Companies frequently deploy data generated by social network users, e. g. for the purpose of market research or marketing. Although public-sector bodies have so far been slower to make use of the opportunities afforded by social media monitoring, it is by no means an unheard-of practice; for example, the tax authorities use web crawlers to trawl through content that is publicly available on the Internet as a way of pinpointing business sellers that are not paying VAT.

Algorithmic systems can be used to make information collated from social media monitoring **usable and exploitable** for more far-reaching and intrusive purposes (in particular the creation of personal profiles for commercial purposes). Provided that the weighing up of interests pursuant to Article 6(1)(f) GDPR supports such a use or exploitation or there is another legal basis for processing, this may be entirely consistent with the law. Pursuant to Recital 51 GDPR, the fact that the data subject has disclosed the data himself or herself does not, in itself, justify the further use and exploitation of the data.

The Data Ethics Commission takes the view that monitoring activities can, at any rate, be deemed to have crossed the boundary between lawful and unlawful when publicly available information is monitored and the scope of this monitoring could not have been gauged by the data subject when the information was disclosed (for example – generally speaking – statements made by minors without due consideration), or alternatively when the information is highly sensitive (for example suicidal ideation statements). Even if applicants for a job have willingly made data public, these data should not be used during the recruitment process if they represent too great an intrusion into personal integrity or if they are not clearly related to the applicant's job history (e. g. statements about his or her sexual orientation). The same applies to any other systematic evaluation of data originating from an individual's private life (e. g. tracking data).

Particularly when the modes of use and exploitation are more far-reaching and intrusive, a weighing up of interests may result in limits being placed on their admissibility (e. g. businesses that target advertisements on the basis of sexual orientation or exploit individuals known to be in an emotionally vulnerable state). Certain providers (in particular providers of social networking sites) are technically capable of carrying out in-depth evaluations of the communications that are exchanged via the central platforms they operate; even if general access to the content is prevented using end-to-end encryption, metadata provide them with the means to obtain highly instructive analytical findings. A legislative ban on the evaluation of communications between individuals or within closed groups should also be imposed on private providers in keeping with the principle of telecommunications secrecy. The Data Ethics Commission therefore recommends that the Federal Government should not delay in its efforts to secure the introduction of such a ban during the forthcoming negotiations on adoption of the ePrivacy Regulation.

3.2.3. The need to clarify and tighten up the applicable legal framework

As things currently stand, a level of protection of legally protected interests that is in line with constitutional requirements can be achieved, for many questions arising in a digital society, only through case-by-case interpretation of general legal concepts and blanket clauses by supervisory authorities and courts. The Data Ethics Commission believes that this situation is untenable. General legal concepts and blanket clauses offer the advantage of being flexible and keeping future options open, yet the authorities and courts often take years or even decades to develop established case law for new phenomena (digital phenomena in particular), which in the meantime results in a **structural enforcement gap** with regard to the law in force and in a **lack of legal certainty**. Given the extent to which this issue affects fundamental rights and the uncertainty as to whether and when solutions will emerge that meet constitutional requirements, the Data Ethics Commission believes that prompt action to establish a clear and binding regulatory framework falls squarely within the remit of the democratically legitimised legislator.

In view of the hazards posed to individuals by **personality-sensitive profiling** (sometimes resulting in **scoring**), the Data Ethics Commission believes that there is an urgent need to take effective action to tighten up the current legal framework in this particularly critical area, in order to effectively counter the risks of individuals being manipulated or suffering discrimination.

Profiling

"Profiling" is defined in **Article 4(4) GDPR** as any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

Profiling ultimately involves making **deductions** (drawing conclusions) on the basis of input data, in particular using certain statistical inference methods (→ Part C, section 2.2.2). These deductions may relate to the actual or purported "properties" of an individual (e.g. "mental stability", "reliability", "social acceptability") and/or take the form of predictions if they relate to an individual's future behaviour (e.g. a particular consumption pattern).

In addition to profiling, attempts are frequently made to assign users to a predefined **stereotype category** on the basis of their observed behaviour when interacting with digital systems, using "matching algorithms"; for example, someone who books a holiday might be classified as a sports fan, a culture enthusiast, a family man or woman, a keen hiker, a sales representative or a gourmet. The stereotype that is instantiated for an individual user is used to store typical preferences, goals and personality traits, which will be used in subsequent algorithmic processing operations.

Sometimes it is not the profiles themselves that are stored; instead, **ad-hoc deductions** (in particular behavioural predictions) are generated dynamically and in real time using raw data (e.g. "is now ready to purchase shoes").

Given that profiling makes it possible to personalise a wide range of digital products and services to a degree that many users perceive as convenient and helpful, a categorical ban on it would overshoot the mark. However, the Data Ethics Commission recommends that the Federal Government should speak out – during the forthcoming evaluation of the GDPR, for example – in favour of **expanding the GDPR to include specific rules on profiling** that go beyond the existing provisions of Article 22 GDPR on the permissibility of automated decision-making; alternatively, the Federal Government could lobby for a separate EU legislative act that would effectively counter the risks that profiling poses to the fundamental rights of individuals. If an adequately hard-hitting European solution proves unworkable in the foreseeable future, legislative rules should be put in place at national level (within the scope of what is permitted by EU law) to regulate profiling procedures that pose a potential risk to fundamental rights.

The Data Ethics Commission believes that there is a particularly urgent need for provisions (horizontal and/or sectoral) on profiling concerning the following matters, as far as solutions do not already follow from correct interpretation of the GDPR:

a) imposition of **absolute limits**, i. e. the prohibiting by law of certain **critical applications** (e. g. when selecting from a pool of job applicants, the use of profiles that have been generated on the basis of data originating from their private lives), of profiling procedures that involve **highly sensitive personal data**, for example in connection with emotion detection software and biometric data, and of data processing operations that entail an **unacceptable potential for harm** to the data subjects or society;

b) imposition of **admissibility requirements** for critical profiling procedures, including quality requirements in relation to the meaningfulness and accuracy of the profiles generated (→ see Part F, section 4.2.1 for further details), and a risk-adequate system of opt-ins and opt-outs (the latter being appropriate only if the level of risk is very low);

c) clarification of the **principle of proportionality**, *inter alia* as regards the requirements that apply to the nature and scope of the data used for profiling, the permitted level of detail in the conclusions drawn for profiling purposes, and in particular the purposes for which profiling may admissibly be used;

d) imposition of specific **labelling, disclosure and information obligations**, *inter alia* as regards the existence and purpose of algorithmic systems that may be used to carry out **ad-hoc deductions**, and any critical deductions that have already been carried out (instead of providing information only on automated decisions taken at a later stage in the process);

e) provision of feasible options for data subjects to **exert an influence** over the profiles that have been created about them, including the option to erase/rectify/verify them; this also includes the right to a "digital new start" involving the erasure of existing profiles (e. g. upon reaching the age of majority), as recently suggested by an EU High-Level Expert Group.[8]

8 High-Level Expert Group on Artificial Intelligence: Policy and Investment Recommendations for Trustworthy AI, 26 June 2019, pp. 14, 40 (available at: *https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence*).

Voice assistants

Voice assistants promise a great deal in terms of convenience and easier access to digital technologies (particularly for people with disabilities), yet they also harbour risks as far as self-determination by data subjects is concerned.

Voice assistants record ambient noise, often without the user having activated any related function. If these recordings include speech by the user or third parties, they are regarded as **biometric data** for the purpose of the GDPR. Speech recordings are analysed in real time so that a response can be given to spoken commands, and automated processes often log certain data types in a log file. The unique timbre of an individual's voice and his or her speech patterns can be analysed as a basis for **uniquely identifying** the individual or deciphering **speech emotions**. Profiling of this kind represents a particularly deep and invasive intrusion into the core area of personality rights, and entails the risk of further exacerbating the structural imbalances between the demand and supply side of the market. Enormous **potential for misuse** is also present given the possibility of recombining or digitally reconstructing the spoken word (deep fakes).

In reality, individual users often have only a vague idea of how data processing is carried out, and indeed of whether it is carried out at all. Particularly if a user is relatively inexperienced in technical matters, he or she may easily be persuaded to disclose additional sensitive personal data upon hearing an authentically **human-sounding voice**. In many cases, voice assistants are not limited simply to recording what is going on in their immediate vicinity, but instead – when networked with other virtual assistants and smart home products – act as the control centre and "technological heart" of modern homes.

The Data Ethics Commission believes that the creation of comprehensive profiles, based on the use of voice assistants and the integration of a wide range of software and hardware components, poses a critical risk. The ease, convenience and apparent benefits of connecting voice assistants to other devices may ultimately lead users into a "plug-and-play trap". In the view of the Data Ethics Commission, a range of measures should be taken to mitigate against the risks associated with voice assistants. These include not only bans on particularly critical profiling procedures and applications, but also the following:

a) binding technical requirements that implement the principles of data protection by design and by default (→ see also section 3.6 below), especially **the processing of speech files on an exclusively local basis** (as well as the option to erase these files locally), and restrictions stating that data may be forwarded to operators or third parties only in the form of commands that have already been translated into machine language (e. g. an order that has been placed);

b) binding technical requirements that include an option to **switch off** the microphone and Internet connection and a way of telling (i. e. a **visual indication**) whether the microphone is on or off (→ see also section 3.6 below);

c) **transparency obligations** which are designed in a manner appropriate to the medium (→ see Part F, section 4.1), i. e. which ensure that the most important information is also provided **acoustically**, either when a pertinent situation arises or at regular intervals.

In addition to special legislative measures of this kind aimed at protecting users, the Federal Government should examine the extent to which it would be possible to lobby for a new or expanded legislative framework to ensure appropriate data governance, preferably at European level but otherwise at national level; this framework should be entirely separate from the goals of data protection law (i. e. outside the scope of the GDPR). The Data Ethics Commission wishes to issue the following special recommendations in this connection (→ see section 3.2.2 above for further examples in each case):

a) blacklisting of data-specific **unfair contract terms** (Sections 308 and 309 of the [German] Civil Code (*Bürgerliches Gesetzbuch*, BGB)) and data-specific contractual and pre-contractual **duties of a fiduciary nature** (Section 241 paragraph 2 of the Civil Code);

b) specification of data-specific **torts** under the umbrella of the existing tort of intentional infliction of harm contrary to public policy (e. g. as a new Section 826a of the Civil Code);

c) blacklisting of data-specific misleading and aggressive **commercial practices**, such as addictive designs and dark patterns, by expanding the blacklist that already exists in the [German] Unfair Competition Act (*Gesetz gegen den unlauteren Wettbewerb*, UWG); the full harmonisation approach of the EU's Unfair Commercial Practices Directive means that this change would need to be initiated at EU level, however.

When profiling is carried out by **government agencies**, the potential for cumulative infringements of fundamental rights or for aggregated surveillance must be taken into account, as must potential side effects or "collateral damage". The Data Ethics Commission believes that there is particular potential for abuse if individual subsystems are connected, resulting in the pooling of data and analytical findings from very different areas and sectors, which significantly steps up the intensity of surveillance. Intelligent pattern recognition techniques (in particular facial recognition) make it easier to link up personal information across a variety of surveillance systems and to merge profiles; in view of this fact, the Data Ethics Commission recommends firstly that pattern recognition techniques of this kind should come into play only when their use is an **absolutely vital prerequisite** for the fulfilment of state obligations, and secondly that clear **legal limits** – beyond the separation rule concerning intelligence activities – must be imposed **on the exchange of information** and patterns between authorities. This may also encompass new legal provisions banning particular types of use and exploitation, particularly as regards the sharing of data between government agencies engaged in preventive and repressive measures.

3.2.4 Uniform market-related supervisory activities

The task of supervising compliance with data protection law by players in the German economy is shared between federal and *Land* authorities. Discrepancies can be observed in terms of the interpretation of data protection law and in the approach to enforcement; this raises certain challenges for the parties affected. Although the European Data Protection Board (EDPB) has been introduced by the EU Member States with the aim of ensuring uniform application of the GDPR, and this institution also has the power to adopt binding decisions in individual cases, the coexistence of different data protection authorities in the various German *Länder* within the framework of the federal system has, to date, prevented the emergence of any such **binding and uniform approach** at national level.

In the event that it proves impossible to strengthen and formalise cooperation between the German data protection authorities, thereby safeguarding the uniform and consistent application of data protection law, consideration should be given to the establishment of a new **data protection authority** at federal level for market-related data activities. Concentrating supervisory powers within a single body would make it possible to build up the specialist expertise required to enforce data protection law in an environment characterised by highly dynamic technological developments. The single authority – either acting alone or in close cooperation with other authorities – would also need to be able to safeguard the enforcement of **other data-related areas of law** that have close functional ties to data protection legislation (e. g. general private law and unfair commercial practices law). The establishment of a single body able to wield market supervisory powers in the field of data protection might also make Germany's voice louder within the European Data Protection Board, since all of the Member States are already represented on the EDPB by a data protection authority with national jurisdiction. Finally, the centralisation of official competencies should go hand in hand with the designation of a single court responsible for judicial control over market-related supervisory authorities in the field of data protection, so that this court can also build up the relevant expertise and set forth a consistent body of case law.

Various models are conceivable from the perspective of organisational law. Based on its powers to regulate economic law, the Federal Government could transfer supervisory competences for data protection in the economy (i. e. the private sector) to the Federal Commissioner for Data Protection and Freedom of Information, and provide the latter with the relevant resources. By setting up a number of different satellite offices, the Commissioner could ensure the nation-wide presence of data protection bodies, similar to the Federal Office for Migration and Refugees or the Bundesbank. Alternatively, the *Länder* could establish a joint facility on the basis of an interstate treaty, by way of analogy to similar projects in the broadcasting sector, for example, or the Central Offices of the *Länder* for Safety Engineering and Health Protection. The joint facility responsible for supervisory activities in the field of data protection would need to be an independent body, and this principle should be enshrined in the interstate treaty. Irrespective of the decisions taken in this connection, the authorities should be provided with better **human and material resources** to allow them to "punch at their weight".

For reasons of constitutional law, the **data protection authorities at *Land* level** should retain jurisdiction **for the public sector**.

3.3 Personal data as an asset

3.3.1 Commercialisation of personal data

The economic significance of personal data is hard to overestimate. It is generally acknowledged that the protection of personality rights as fundamental rights also encompasses the individual's right to decide whether certain **aspects of his or her personality** should be made **available for a fee** (e.g. the right to one's own image), or in other words whether they should be exploited for economic purposes.[9] In the same way that there is not a complete ban on the exploitation of data by individuals, however, there are no rules categorically stating that personal data may not be exploited for economic purposes on the initiative of third parties. Some people compare the situation to the trade in human organs, but this comparison is flawed in several respects: unlike human organs, data are a non-rivalrous resource, and so the mere fact that personal data are processed by someone else does not in and of itself necessarily cause harm to the data subject – harm is caused only by the processing of data in specific contexts or for specific purposes.

Interpreting the right to informational self-determination as a natural corollary of human dignity makes it clear that the **limits** imposed on the economic exploitation of personal data should generally coincide with the general limits placed on the processing of personal data (→ see sections 3.2.1 and 3.2.2 above), including the substantive limitations on consent. Against this backdrop, the economic exploitation of personal data is neither subject to more stringent rules in general, nor privileged in any way. Economic aspects frequently come into play when general data protection rules are applied, however (for example, consent may no longer be freely given if the data subject is exposed to economic pressure).

3.3.2. Data ownership and the issue of financial compensation

As things stand, the Data Ethics Commission does not believe that there are **adequate grounds** for introducing additional ownership-like rights of exploitation that would allow data subjects to request an economic share in the profits derived with the help of data (often referred to under the concepts of "**data ownership**" or "data producer right").[10] Both data protection law and general private law already provide the individual with a range of legal rights that are effective vis-à-vis third parties, and on the basis of these rights individuals could theoretically make their toleration of data activities dependent on payment of an appropriate fee. If the individual fails to negotiate a fee of this kind, this can be attributed to circumstances (e.g. lack of negotiating power and/or poorly functioning competition) that have nothing to do with the absence of any additional ownership-like right of exploitation.

In theory, the imbalance in negotiating power could be counter-balanced through the introduction of **collective societies** that collectively exercise ownership-like rights to exploit data. Extending the concept of personal data to include an ownership-like economic component would, however, potentially be **at odds with data protection**, in particular as regards the voluntary nature of consent, the ability to withdraw consent at any time and the right to request erasure. It would also create **questionable financial incentives** by encouraging the generation of a maximum of personal data, and would put pressure on individuals (in particular on vulnerable groups such as minors and low earners) to disclose as much data as possible. If industry passes the costs of any such remuneration on to the customers, **privacy-conscious individuals** might also be forced to shoulder a comparatively **greater burden** in financial terms.

9 See e.g. Section 22 of the [German] Act on the Protection of Copyright in Works of Art and Photographs (*Gesetz betreffend das Urheberrecht an Werken der bildenden Künste und der Photographie*, KunstUrhG).

10 By way of examples: European Commission: Building a European data economy, 10 January 2017, COM(2017) 9 final (available at: *https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-9-F1-EN-MAIN-PART-1.PDF*); Arbeitsgruppe "Digitaler Neustart" der Konferenz der Justizministerinnen und Justizminister der Länder [Working Group "Digital New Start" of the Conference of Ministers of Justice of the Länder]: Report of 15 May 2017, pp. 29 et seqq. (available at: *https://www.justiz.nrw.de/JM/schwerpunkte/digitaler_neustart/zt_bericht_arbeitsgruppe/bericht_ag_dig_neustart.pdf*).

The above arguments do not hold water to the same extent when it comes to anonymised data. However, given the huge number of individuals that contribute to the generation and processing of data, the level of **complexity** of a fair remuneration system and the 24/7 monitoring that would be required to measure data flows would be out of all proportion to any potential gains in terms of justice. **Data quality** might also be negatively affected, since incentives would be created to generate data "artificially" (e.g. through the creation of fake profiles), ultimately producing a distorted picture of reality. The Data Ethics Commission therefore counsels against **introducing rights of exploitation** designed as exclusive rights, **either for anonymised data or for other data types**.

3.3.3. Data as counter-performance

A large number of digital content and service types (e.g. search engines, social networks, messenger services, online games) are offered to end users for no monetary consideration. They are financed in other ways, in particular through payments received from third parties in exchange for personalised advertising and other personalised information services targeted at users, or for user profiles and user scores. Personal data are therefore often referred to in shorthand terms as "counter-performance" for digital content or services, for example in the original draft of Article 3(1) of the Digital Content Directive (although the term was removed at a later point in the legislative procedure).[11] The extent to which the economic model described above is, in fact, compatible with the **prohibition under Article 7(4) GDPR of "tying" or "bundling" consent with the provision of a service**[12] must ultimately be clarified by the European Court of Justice.

The Data Ethics Commission argues that **data should not be referred to as "counter-performance" provided in exchange for a service**, even though the term sums up the issue in a nutshell and has helped to raise awareness among the general public. Firstly, personal data form an integral part of an individual's personality, and are protected under constitutional law. Secondly, their classification as a counter-performance might have unintended consequences. For example, it might be abused as an argument in favour of largely excluding data-related standard contract terms from unfairness control, or as a justification for triggering contractual sanctions against consumers who withdraw consent or exercise their right to erasure, etc.

In this connection, the German legislator should not – when implementing Directive (EU) 2019/770 on certain aspects concerning contracts for the supply of digital content and digital services – use the leeway available to Member States in any way that might prevent the individual from seeking legal remedies under data protection law. In particular, if an individual withdraws his or her consent to the processing of data, the provider may have a right to terminate provision of its service with immediate effect; however, it should not be possible for the provider to request **payment for services already provided**, and there should be no retrospective and **automatic reversion to a pay option**.

Pay options are increasingly being discussed as a way of avoiding the "tying" or "bundling" of consent with the provision of a service. Yet even the smallest of financial burdens represents a disadvantage, in particular for vulnerable population groups, and may dissuade data subjects and encourage them to disclose excessive amounts of personal data. It is also to be feared that the financial burden on privacy-conscious individuals would be disproportionate. **Commercial users** that have previously been able to use certain digital content or services for free (e.g. a company's page on a social networking site) should therefore be the **preferred source of funding**.

11 European Commission: Proposal for a Directive of the European Parliament and of the Council on certain aspects concerning contracts for the supply of digital content, 9 December 2015, COM(2015) 634 final (available at: *https://ec.europa.eu/transparency/regdoc/rep/1/2015/EN/1-2015-634-EN-F1-1.PDF*).

12 European Data Protection Supervisor: Opinion 4/2017 on the Proposal for a Directive on certain aspects concerning contracts for the supply of digital content, 14 March 2017, p. 15 (available at: *https://edps.europa.eu/sites/edp/files/publication/17-03-14_opinion_digital_content_en.pdf*).

Pay options may, however, increase consumer awareness of the financial value of their own data, and also create transparency. For these reasons, the Data Ethics Commission believes that **offering pay options as an alternative** may be an ethically acceptable way to ensure that consent given by users is genuinely free. At the same time, however, the price must not be abusive and exceed market prices; from the consumer's perspective, it must represent a realistic alternative to the disclosure of personal data. From an ethical viewpoint safeguards must be put in place to protect privacy-conscious users from having to "cross-subsidise" other users; equally, the needs of socially vulnerable groups must be taken into consideration, for example through government transfers.

3.3.4 Data as the basis for personalised risk assessments

Price-related predictions obtained using algorithmic systems for the purpose of **personalised risk assessment** (e.g. on a one-off basis when approving a loan or on an ongoing basis in the case of black box schemes operated by insurance companies) are characterised by a higher level of granularity. This is ultimately a sector-specific use case for a certain profiling technique and the associated scoring procedures (→ see section 3.2.3 above and Part F, section 4.2.2 below for further details of profiling in general). The processing of additional personal data for the purpose of personalised risk assessments regularly requires consent from the data subjects. Individuals who hope to gain economic advantages as a result are particularly likely to grant such consent, yet the granting of consent by one individual may have significant impacts on others, and give rise to chain reactions that are problematic from an ethical viewpoint (unravelling effects). This may put data subjects under disproportionate pressure, and jeopardise the voluntary nature of consent.

Example 11

Insured parties who are healthy are particularly likely to consent to the processing of their data by a health insurance company. As a result, others come under pressure to also grant consent in order to avoid arousing any suspicions regarding their state of health.

In cases where individual behaviour can influence the parameters, models of this kind can also have a significant **influence on how people lead their lives**. Another ethical consideration that is particularly relevant in the insurance sector is that the goal of increasingly granular risk assessments runs counter to the **basic principle of collective risk sharing** by the community of all insured persons. Taken to its extreme (i.e. if the insurer has access to "comprehensive" information and adjusts the price to the individual risk), the whole concept of insurance would be reduced to absurdity.

The Data Ethics Commission therefore believes that personalised risk assessments must comply with the following ethical requirements in particular:

a) data processing must not intrude into **the core of an individual's private life**; it must be restricted to areas where the individual is already in contact with the exterior world and must therefore expect conclusions to be drawn on the basis of his or her behaviour. This principle dictates that it would be ethically acceptable for a car insurance company (for example) to record the miles driven or traffic offences committed by a driver, but not purely private behaviour inside his or her vehicle, even if this behaviour might be relevant from a risk perspective (e.g. how often he or she yawns, whether he or she chats to passengers), or even the driver's state of health (e.g. heart problems) or other lifestyle factors (e.g. purchasing behaviour in relation to coffee or alcohol);

b) a **clear causal relationship** must exist between the data being processed and the risk to be determined, and any linking of data must avoid **discriminatory** repercussions (→ see Part F, section 2.6 below for further details);

c) the data must not allow conclusions to be drawn directly that have **implications for relatives or other third parties**;

d) full **transparency** is required as regards the specific parameters and their weighting, and the impacts on pricing or other conditions; the individual must also be provided with clear and comprehensible explanations of how to improve these conditions (→ see Part F, section 2.7);

e) in order to keep unwanted chain reactions in check, the difference between the "optimal" conditions and the conditions that apply if consent is refused must not exceed a certain ceiling (e.g. **maximum price difference**).

In the view of the Data Ethics Commission, most of the problems arising in connection with **price differentiation in the narrow sense** and measures of a similar ilk relate to the regulation of algorithmic systems (→ see Part F for further details). At the same time, however, price differentiation morphs into a data use problem as soon as consumers are led to believe that they can access prices that are lower overall by disclosing as much personal data as possible or by exhibiting certain behaviours tailored to the relevant criteria (e.g. making online purchases using a computer manufactured by a certain company), or conversely if it is suggested that consumers who refuse to consent to the processing of their data for the purpose of personalised pricing will always pay **higher prices on average**. The Data Ethics Commission believes that the latter would also pose an ethically questionable risk to the voluntary nature of consent.

3.3.5 Data as reputational capital

When coupled with **personalised economic conditions** (personalised prices, personalised ranking and personalised products and services), personal data, profiles and scores serve as reputational capital. Personalised behavioural rewards aimed at increasing **customer loyalty** (e.g. the granting of discounts depending on the quantity purchased in the previous month) incentivise consumers to consent to the processing of their personal data, and may be apt to influence the way they lead their lives. No evidence that the ethical limits outlined above (→ section 3.3.4) are currently being disregarded in the German economy in connection with customer loyalty programmes has come to the attention of the Data Ethics Commission, but developments should continue to be monitored.

True reputational data that are also visible to external third parties (e.g. "stars" indicating that someone with a profile on an online platform is a good person to do business with) are gaining ever more economic and non-material significance. To a certain extent, reputational data of this kind are covered by the new **Regulation (EU) 2019/1150 on promoting fairness and transparency** for business users of online intermediation services.[13] The regulatory approach chosen by the lawmakers who drafted this Regulation – based for the most part on transparency requirements and self-regulation – was cautious, and the Data Ethics Commission welcomes this approach in principle. However, it is worth noting that certain sectors are heavily dependent on true reputational data, and that this factor in particular might lead to significant lock-in effects that may jeopardise competition and cause problems if individuals are unable to take their data with them when switching to a different online intermediary platform.

13 Cf. Article 9 of this Regulation on data access and many general provisions, e.g. on general terms and conditions of business and ranking.

Example 12
A micro entrepreneur who offers taxi services via an online platform has been ranked highly by many of his former passengers, and now wishes to switch platform and take these rankings with him.

The Data Ethics Commission is aware of the problems that would arise if a general obligation to recognise ranking profiles built up on a different platform were to be enshrined in law. However, it recommends that the Federal Government should examine the conditions under which commercial users with profiles of this kind might nevertheless be granted a **right to portability**, with a view to lobbying for broader regulation at European level.[14]

By way of contrast, the rise in significance of **social reputation data** (number of "likes", "followers" or "friends") is part of a wider trend in our society, and – with the limited exception of "influencers" – can no longer be viewed predominantly through the lens of personal data as an economic asset, but must instead be discussed in relation to its systemic societal implications.

3.3.6 Data as tradeable items

A significant number of companies are already deriving financial gain (and, in some cases, earning a great deal of money) by compiling personal data, profiles and scores or personalised statistical evaluations (carried out using aggregated raw data) and then reselling them to third parties, or by enriching existing profiles with estimated data and then placing them on the market. In the following section, business models of this kind will be referred to as **"data trading"**.

The GDPR does not currently contain any provisions relating specifically to data trading; instead, business models of this kind are categorised merely as normal data processing operations that are subject to the general provisions of the GDPR. In many cases, closer examination of the applicable provisions leads to the inescapable conclusion that certain types of data trading infringe the provisions of the GDPR, and are therefore contrary to the law. Generally speaking, however, the field of data trading is characterised by a **significant enforcement gap**. The Data Ethics Commission therefore believes that urgent action should be taken by the data protection authorities in relation to this sector, and that the European Data Protection Board (EDPB), or alternatively the Conference of Independent Data Protection Authorities of the Federal Government and the *Länder (Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder)*, should develop – in keeping with the GDPR's risk-based approach – clearly delimitable categories for different types of lawful data trading. Greater clarity is needed regarding the instances of data trading where the data subject must grant consent to the forwarding of data, the instances where the data subject has a right only to object to the processing of data, and the instances where compelling reasons rule out even the right to object.

Having regard to the general principles governing data processing (Article 5 GDPR), the forwarding of data to third parties should be permitted only within closely prescribed limits in situations that are not covered by the existing provisions of data protection law. The Data Ethics Commission therefore recommends that the Federal Government should speak out at European level – in connection with the forthcoming evaluation of the GDPR, for example – in favour of **expanding the scope of the GDPR to include specific provisions on data trading**. The following **ethical considerations**, some of which are already enshrined in the GDPR, should be taken into account when drafting future legal provisions of this kind:

14 Cf. for example Articles 6 and 7 of the draft "Model Rules on Online Intermediary Platforms" by the European Law Institute, which were made available to the Data Ethics Commission.

a) The individual's right to informational self-determination should be the starting point for any balancing exercise, meaning that data trading in principle requires prior **consent** by the data subject, with due regard for the **substantive limitations** on consent (→ sections 3.2.1 and 3.2.2 above).

b) If data are processed on a legal basis other than consent (which is likely to occur only in isolated cases), the individual must have a straightforward opportunity to exercise his or her **right to object** in advance (e. g. by unchecking a checkbox immediately before the data are collected), and must not be forced to communicate his or her objection via separate communication channels.

c) Data trading models that deprive data subjects of any **choices** whatsoever should only rarely be considered, and only if and to the extent that the data need to be forwarded in order to further public interests that manifestly outweigh the countervailing interests. Comprehensive legislative clarification of this category is required.

d) The GDPR contains detailed provisions on the transfer of data to processors and on the forwarding of data to third countries. Given the content and rationale of the GDPR, it would be illogical to assume that the requirements that apply to transfers of data to third parties within the EU should be any less stringent than those that apply to transfers outside the EU, and certain other points can also be inferred from the general provisions, e. g. that these requirements should be regarded as "appropriate safeguards". Nevertheless, the Data Ethics Commission recommends that urgent action be taken to clarify (explicitly and by law) the obligations that apply when transferring data to third parties, e. g. control obligations, as well as the circumstances under which parties may be held liable.

e) Controllers should be obliged to document and disclose the specific source of the data they have collected or generated by the use of algorithmic systems, as well as the identity of the individual recipients of the data; the information must be provided in a standardised and machine-readable format, which allows e. g. automated data management using a privacy management tool/personal information management system (→ see section 4.3 below for further details). This would take due account of the fact that **data subjects** have largely been **left in the dark** as regards the existence of data traders, which means that a simple list of the different categories of sources or recipients would be of little use to them.

f) Given the large number of data traders in the market, data subjects will be able to **exercise their rights effectively** only if central mechanisms are established that facilitate this process or assume responsibility for it (e. g. data protection authorities, → see section 3.2.4 above, or privacy management tools/personal information management systems, see section 4.3 for further details).

g) Given that dispersion effects give rise to higher risks and the potential for loss of control, data traders should be subject to a **certification obligation under data protection law** that includes regular audits by the certification bodies. The Data Ethics Commission recommends that specific certification criteria should be adopted as appropriate by the independent data protection authorities of the Federal Government and the *Länder*, and that these criteria should take due account of the risks and recommendations it has outlined.

3.4 Data and digital inheritance

Modern communication technologies and data processing capacities make it possible to record every last detail of an individual's private activities for decades on end, and to evaluate these recordings using automated systems. Handing the data collected about a deceased individual over to his or her heirs or another third party adds a **whole new dimension of privacy risk**, both for the deceased person and, in particular, for the individuals with whom he or she communicated during his or her lifetime. These data are often compared to diaries and personal correspondence, but this comparison is flawed because many channels of digital communication (messenger services, chats, e-mails, etc.) serve as a functional replacement for the ephemeral spoken word rather than for letters.

3.4.1 Precedence of living wills

The Data Ethics Commission believes that, in the best-case scenario, a data subject should make intentional and informed dispositions during his or her lifetime. In many cases, however, people neglect to make any such dispositions for the sole reason that they are unaware of the legal and practical options or put off by the level of uncertainty. Against this backdrop, the Data Ethics Commission believes that there are justified grounds for **obliging service providers** to alert users to the option of making dispositions that provide for ongoing incapacity to provide consent (e.g. due to dementia) or for death, and to provide the technical means for making said dispositions, with the minimum of barriers (i.e. with the fewest possible changes of medium). Corresponding provisions could be added to the **[German] Telemedia Act (*Telemediengesetz*, TMG)**.[15]

In the view of the Data Ethics Commission, the situation following a data subject's death is merely an extreme example that should serve as a prompt for further reflection on the general design of digital modes of communication. The Data Ethics Commission therefore recommends that the Federal Government should examine the possibility of making it obligatory for messenger services to offer a **default option of erasing messages** after a certain period of time; if a user chose this option, a message would automatically be erased after expiry of the relevant period unless it had been manually archived by the recipient or the sender.

3.4.2 The role of intermediaries

Growing awareness of the topic of digital inheritance has allowed new business models to flourish, and a large number of companies are now offering services in this field (ranging from the central storage of account data and passwords through to comprehensive digital inheritance management). These services may provide useful guidance, but they are also associated with certain hazards, including inadequate provision for cases in which a company goes bankrupt or is otherwise liquidated, and shortcomings in information security (up to and including genuine fraud). The Data Ethics Commission believes that **quality assurance**, new **regulations** (characterised by a cautious approach) and **public awareness-raising** about the potential advantages and risks are required in order to protect citizens.

15 For a previous discussion of this topic, see Mario Martini, Juristenzeitung (JZ), 2012, p. 1154.

In addition, it recommends that the Federal Government, as part of its remit to provide services of general interest to the public, should set up a body that is (at the very least) subject to **state supervision** and that provides affordable basic digital inheritance protection and planning services to citizens; these services must reflect the latest developments in the field of information security technology. When a German citizen writes a will, he or she can choose to store it privately or with a notary or district court, and similar options (private or private-sector solutions or a government-run service) should also be available for an individual's digital inheritance.

3.4.3 Post-mortem data protection

The Data Ethics Commission does not recommend a wholesale rejection of the principles set forth by the German Federal Court of Justice[16] regarding the **transfer of estates to heirs**, since the potential advantages would be far outweighed by the effects (either undesirable and/or excessive) of a different default solution, e.g. a trust model imposed by law or a distinction between user account content that is regarded as an asset and content from the same user account that is regarded as highly personal. Conversely, inheritance law should not apply at all if the nature of a user account (e.g. an online account with an Alcoholics Anonymous group) renders all of the data within it financially worthless but highly sensitive. In cases where the **principle of telecommunications confidentiality** applies, *inter alia* to protect the deceased's communication partners, the legislator will, in any case, still have to reconcile this with the right to inheritance (which is enshrined as a fundamental right), for example through a corresponding reference in the part of the Civil Code devoted to inheritance law.

The principle set forth by the Federal Court of Justice – that an estate should be transferred to the deceased's heirs – is linked to the existence of a contractual relationship. If there is no contractual relationship, or if a transfer to the heirs cannot take place owing to the highly sensitive nature of the data, the heirs will have no right of legal recourse. Since **post-mortem data protection** is not provided by the GDPR, there are also no means of legal recourse for relatives under the current state of data protection law. Ethical concerns are raised by the fact that controllers have almost unlimited power to dispose of a deceased's personal data as a result, and the Data Ethics Commission therefore recommends that the Federal Government should follow in the footsteps of several other EU Member States and make use of the option provided by Recital 27 of the GDPR, by enacting provisions on **post-mortem data protection**. Even after the death of a data subject, the latter's relatives should be able to exercise his or her fundamental rights, such as the right to erasure and the right to rectification of incorrect data. At the same time, suitable measures should be taken to ensure compliance with dispositions made by the deceased during his or her lifetime, even if these dispositions are only implied (e.g. through a deliberate choice to publish a "life story").

16 Judgment by the German Federal Court of Justice of 12 July 2018, ref. III ZR 183/17.

3.5 Special groups of data subjects

3.5.1 Employees

The fact that employers collect employees' location data and performance data, which is a widespread phenomenon in certain modern workplaces, poses a significant risk to these employees' **right to informational self-determination and general rights of personality**; the same is true of the creation of biometric profiles which is a necessary precursor to certain forms of collaboration. Questions to be considered include not only the legal basis for data processing and for the granting of co-determination rights to employee representation bodies, but also obligations to provide employees with information (e.g. on the hazards posed by multi-sensor fusion) and, depending on the context, with opportunities to object, issues regarding data retention procedures, terms of data retention and the extent to which employees' data may be disclosed to third parties, the right to rectification of incorrect or obsolete data (in personal profiles, for example) and appropriate erasure procedures. Further points for consideration include framework conditions for (limited) control and surveillance of employees, restrictions on the tracking of employees' locations and a ban on comprehensive location profiles, restrictions on any obligation to share social media accounts or to allow an employer to access data in the context of "bring your own device" models, framework conditions for the use of biometric systems, and restrictions on psychological investigation methods.

The Data Ethics Commission recommends that the Federal Government should invite the social partners to work towards a common position on the legislative provisions that should be adopted with a view to **stepping up the protection of employee data**, based on examples of best practices from existing collective agreements. The concerns of individuals in non-standard forms of employment should also be taken into account during this process, and collective agreements and works council agreements should continue to play a significant part in employee data protection. Yet the foundational principles of employee data protection should not be regulated solely by collective agreements and works council agreements, firstly because not all employees are covered by these latter, and secondly because of the importance of these principles from a fundamental rights perspective. It is also worth noting that the legal uncertainty currently reigning over the scope of the GDPR provisions is having a negative impact on investment security.

With reference to the wider field of legal bases for the processing of employee data, the Data Ethics Commission believes that the traditional construct of **consent** under data protection law is not suitable in all contexts, since it is difficult to put in place the framework conditions necessary for consent to be given voluntarily in all employment situations, and impossible to find an appropriate balance in all cases between the employer's needs and the option for employees to revoke consent and request the erasure of data at any time. Employee data protection measures should therefore focus on **legal grounds of justification** that are specifically tailored to the employment context, and that guarantee a high level of protection and an appropriate weighing up of interests against fundamental rights. The outcomes may look very similar to consent in certain respects, while taking into account the power structures that typically exist in an employment context.

When deciding whether **interest groups should be granted co-determination rights**[17] in relation to the processing of data within companies, due regard must be given to the **asymmetry of knowledge** that exists between employers and employees as regards the operating principles and details of these data processing operations. There is a need for models that go further than the existing mechanisms by allowing interest groups to access external expertise, while at the same time ensuring not only the appropriate involvement of the company data protection officer, but also the protection of trade secrets. Given the constant advancement of data-processing systems within companies (software updates, self-learning elements, etc.), there should be a shift away from consent as a single, one-off event and towards **ongoing oversight of processes** by interest groups.

Progress in the field of employee data protection should not neglect the stages of **applying** for a job and **entering into an employment relationship**. For example, care must be taken to ensure that the provisions of applicable law that prohibit employers from asking certain questions during the application procedure or when recruiting an individual (e.g. asking whether a woman is pregnant) are not circumvented through the use of "human resources" algorithms or through a request to grant the employer access to social media accounts.

Steps must also be taken to ensure that persons in **non-standard forms of employment** are not excluded from progress in the field of employee data protection. The upsurge in these forms of employment in the platform economy means that many people no longer have access to the traditional employee rights and rights of co-determination. The imbalance of power that arises between the client or the platform operator on the one hand and the contractors or the platform workers on the other is often significant and may have implications in terms of data protection and informational self-determination. Appropriate legislative provisions should be adopted (ideally at EU level) and the institutional framework developed further (e.g. through an interest group) to mitigate against this risk.

3.5.2 Patients

In view of the benefits that could be gained from **digitalising healthcare**, as a basic principle the Data Ethics Commission recommends **swift expansion of digital infrastructures** in this sector and the introduction of **procedures for reviewing and assessing digital healthcare services**. Both the range and the quality of digitalised healthcare services should be improved to allow patients to exercise their rights to informational self-determination and become more health literate.[18]

Even as things stand today, the provision of healthcare services involves the processing of huge volumes of personal data. The data involved are typically health data and genetic data, or in other words special categories of personal data within the meaning of Article 9 GDPR. When designing a future health landscape that will be primarily digital in nature, comprehensive account must be taken of the need to provide **special protection for these data** at the same time as boosting the **right to self-determination** of patients and those with health insurance policies, *inter alia* in the field of research (→ see section 4.1 below).

17 For examples of current legislative provisions, see e.g. Section 87 (1) (6) of the [German] Works Constitution Act (*Betriebsverfassungsgesetz*, BetrVG) (in relation to works councils), or Section 75 (3) (17) of the [German] Federal Staff Representation Act (*Bundespersonalvertretungsgesetz*, BPersVG) (in relation to staff councils).

18 German Ethics Council (*Deutscher Ethikrat*), Big Data and Health, Opinion, 30 November 2017 (available at: *https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/englisch/opinion-big-data-and-health-summary.pdf*).

In this connection, the Data Ethics Commission emphasises the urgent need to introduce and roll out an **electronic health record** with a view to improving the quality, transparency and cost-effectiveness of medical care.[19] Given the vital role that an electronic health record would play in digitalising the healthcare sector, the Data Ethics Commission wishes to make it clear that greater attention should be paid to both information security and patient autonomy while implementing this system; the existing cryptosecurity concept (based on the decentralised management of keys (PINs) for insured parties) should continue to apply, for example. It should also be possible to use the electronic health record even if a patient is incapable of granting consent, based on the provisions concerning legal representation that otherwise apply and regardless of the type of health insurance policy held by the patient.

Digital health services and products that are not collectively financed **(consumer-funded health market)** are becoming ever more important, not least because the digital healthcare services offered by the statutory health insurance funds have been few and far between to date. It is important not to underestimate the relevance of these services – which include not only fitness, health and wellness apps, but in particular digital **self-monitoring** apps and the associated wearables – in the context of a digitalised healthcare sector. Yet these apps are often of questionable (and poorly verified) quality, meaning that the data they collect are of limited usefulness; this carries a risk to the health of the affected patients and users, which can, in some cases, be significant. It should furthermore not be assumed that patients are able to assess the quality of these products and services independently, in particular their compliance with the principles of data protection and information security; equally, access to digital healthcare services should not be dependent on individual financial wherewithal. With this in mind, the Data Ethics Commission welcomes the plans for the Federal Institute for Drugs and Medical Devices *(Bundesinstitut für Arzneimittel und Medizinprodukte)* to introduce a procedure for examining and assessing apps of this kind.

3.5.3 Minors

The Data Ethics Commission welcomes the efforts which have been undertaken – and which include both the adoption of legislation and voluntary self-regulation – to develop special **protective mechanisms** allowing minors to exercise their right to digital self-determination. The primary goal of these mechanisms should be to step up the level of data protection and the degree of protection against profiling, manipulation through dark patterns and addictive designs, etc.; their secondary goal should be to provide greater protection against content that is not age-appropriate (that glorifies violence, for example).

At the same time, however, the Data Ethics Commission wishes to make it clear that all such protective mechanisms will prove futile unless a reliable **identity management system** is in place, ensuring that the age of minors is detected and that they are treated appropriately. Relying on users to be honest about their age is without question the wrong approach. When viewed through the lens of ethics, however, it would also be problematic to ask providers to ascertain a user's age themselves by collecting personal data, some of which may be highly sensitive (e.g. facial recognition, with data transferred to the provider's cloud); at the same time, placing the entire burden on whoever holds parental authority may easily result in a situation where the latter feels that too much is being asked of him or her. The Data Ethics Commission therefore recommends that the Federal Government should promote the emergence of **family-friendly technologies** that allow minors to exercise their right to self-determined development while, at the same time, reliably guaranteeing their protection.

19 See in this respect the Data Ethics Commission's previous recommendation on participatory development of an electronic health record, dated 28 November 2018 (available at: *www.datenethikkommission.de*).

The Data Ethics Commission recommends that the Federal Government should lobby at European level for measures to enforce compliance with the principles of **data protection by design and by default** as enshrined in the GDPR, particularly in the case of mobile end devices, in order to protect the right to informational self-determination of minors and protect their privacy. The German and European data protection authorities, the competition authorities, the media regulators and the technical regulatory authorities should take action within their relevant remits and spheres of responsibility to force the manufacturers of operating systems for mobile end devices and the providers of digital services to adhere to all of the legislative requirements that apply to the age groups in question and to block services that are not age-appropriate. The parties responsible for procuring relevant systems with a view to their use in schools and kindergartens should also incorporate these requirements into the tendering procedures. A more detailed discussion of the need to force manufacturers to comply with the principle of data protection by design and by default can be found below (→ section 3.6.1).

As far as further action in this area is concerned, consideration should also be given to the introduction of an EU-wide obligation that forces manufacturers of child-friendly mobile end devices to program them from the outset as devices that are specifically intended for children, and to ensure that "jail breaking" or "rooting" is impossible (or possible only with a key). The devices programmed in this way should enforce compliance with all of the legislative provisions aimed at protecting children, and block services that are not age-appropriate. **If the relevant settings are enabled** on the device/operating system upon activation, minors should not be able to change these settings without their parents' consent. A solution of this kind would also offer clear advantages over parental control apps, firstly because these apps often pose data protection and information security problems in their own right, and secondly because they raise ethical questions in terms of the opportunities they afford for the total surveillance of private life.

3.5.4 Other vulnerable and care-dependent persons

In many cases, data belonging to vulnerable individuals are processed for the benefit of these individuals, e.g. in the care sector. Digital technologies can make it much safer for older people to remain in the environment to which they are accustomed, for example, and they may also help to alleviate some of the negative impacts of the skills shortage in the care sector and ensure better healthcare provision. In particular, **digital assistance systems** – when used correctly – can serve as a bridging technology, and adjust adaptively to the varying needs of different people.

The right to life, the right to bodily integrity and also the right to informational self-determination are fundamental rights that must be reconciled with each other in accordance with the principle of practical concordance. Particular consideration must be given to two questions in particular: whether **risks are posed to life or health**, and the extent to which the right to informational **self-determination** is encroached upon.

The Data Ethics Commission believes that **standards and guidelines** on surveillance by professionals in the care sector should be developed by the Conference of Independent Data Protection Authorities of the Federal Government and the *Länder*. In particular, these standards and guidelines should specify the legal provisions upon which the professionals can base their action in particular situations, and the cases in which (especially if **consent** has not been granted by the data subject or his or her caregiver) surveillance is either prohibited or possible on the basis of Article 6(1)(f) or (d) GDPR. They should also outline arrangements for the provision of information, whereby the Data Ethics Commission takes the position that differentiated information on digital surveillance options should be provided prior to their use in an institutional setting (e.g. care home, kindergarten or school), and that consent must also be obtained on a differentiated basis in cases where there is no legal basis for data processing. Standards and guidelines of this kind would also be an appropriate way both to provide more legal certainty for care home operators and care staff and to reduce liability risks. Section 1901a of the Civil Code should be amended accordingly to clarify the fact that **living wills** can also include dispositions in which the relevant data subject grants prior consent to the processing of data.

As a basic principle, a particularly high level of protection should also be accorded to people in their own homes, since they are likely to regard the space within their four walls as a safe haven of privacy. Once again, new technologies have opened up new and expanding **options for the surveillance of private individuals by other private individuals** (e.g. the surveillance of romantic partners, children or persons with disabilities), which range all the way through to the ethically alarming prospect of total private surveillance. Given that awareness of this topic is lacking in many quarters, the Data Ethics Commission recommends that **awareness-raising campaigns** in this area should be initiated both by the Federal Government and by the governments of the *Länder*, since the latter often hold jurisdiction in this field. Although it recommends that the Federal Government should continue monitoring developments, it does not believe that legislative measures (e.g. new criminal offences) are required at present.

3.6 Data protection by technical design

Citizens, companies, government agencies or other parties that are entitled to assert ethically justified data rights and that are obliged to comply with the corresponding data obligations must be in a position to do so in the first place. The necessary **technical framework** must be put in place, and enabling technologies must play a prominent role in this respect. Yet enabling technologies of this kind must not lead to a situation in which responsibility for the protection of fundamental rights and freedoms is offloaded onto individual users. Instead, the State must, as a matter of principle, adopt the **regulations** that are required to provide reliable protection for these fundamental rights and freedoms, without the need for action on the part of individuals.

3.6.1 Privacy-friendly design of products and services

As its heading suggests, Article 25 GDPR makes it mandatory for controllers to comply with the principles of **"data protection by design and by default"**. Designers of new technologies must therefore take due account of concerns relating to data protection (based on the interpretation of the term applied in Article 5 GDPR), while following a risk-adequate approach. The technical and organisational measures that must be implemented to this end may be required prior to processing (i.e. when the controller determines the means by which the data will be processed) as well as during the processing operation itself.

Data protection by design and by default

Data protection by design imposes conditions on the selection of technical and organisational measures (relating to the state of the art, implementation costs, processing and the risk posed to the rights and freedoms of natural persons, for example). **Data protection by default** imposes no such conditions, since this principle must be adhered to without exceptions. In practice, however, it is often the case that excessive amounts of personal data (e. g. identifiers) are processed, that inadequate restrictions are placed on processing, that retention periods are too long and that an inappropriately high number of people are able to access the data.

The field of **"privacy engineering"** has therefore emerged under the banner of additional protection-related goals such as non-linkability, transparency and intervenability; the standard data protection model (SDM) used by the German data protection authorities now incorporates these goals as "warranty objectives".[1] Like the IT Baseline Protection Catalogues published by the Federal Office for Information Security (*Bundesamt für Sicherheit in der Informationstechnik*, BSI), the SDM defines modules that can be used by controllers and designers of new technologies as a basis for choosing technical and organisational measures that are appropriate to their protection needs. Although only the first few modules are currently available, others are planned. The fact that many developers use the IT Baseline Protection Catalogues and the ISO 2700x series of standards as reference works means that these developers are familiar with the fundamental concept and able to take better account of the legal requirements when designing and implementing technical systems.

The choice between **centralisation and decentralisation** is another question that must be clarified on a case-by-case basis when designing technical systems. As a general rule, centralised systems allow operators to exercise a higher level of control and influence. This might be a good thing, for example if the underlying aim is to incorporate features that contribute to data protection or information security. Yet it can also be a bad thing, since the potential for misuse – either by malicious third parties wanting to steal data or sabotage data processing, or by the operators themselves exploiting the large volumes of data they have amassed for purposes other than those notified to the data subjects – is greater if data are stored centrally and the processing of these data is also controlled centrally. When designed appropriately, however, decentralised systems can help to decrease or prevent data linkability, and reduce disruptions to overall system availability.

1 Technology Working Group of the Conference of Independent Data Protection Authorities of the Federal Government and the *Länder*: Das Standard-Datenschutzmodell – Eine Methode zur Datenschutzberatung und -prüfung auf der Basis einheitlicher Gewährleistungsziele V.1.1 – Erprobungsfassung [The standard data protection model – a method for data protection consulting and assessment on the basis of uniform warranty objectives, V.1.1 – test version], 2018 (available at *https://www.datenschutzzentrum.de/sdm/*).

The design specifications of data protection law have a high level of practical relevance in relation to **end devices**. Some end devices are designed to be worn on the body (wearables, e. g. a smartwatch or smart textiles) or at least carried close to the body (e. g. a smartphone), while others are designed to be mobile by other means (e. g. a networked car) or immobile (e. g. smart home facilities). When designing software systems for end devices of this kind, the amount of time that should be spent reflecting on the ethical questions they raise depends on the likelihood that they will be used in close proximity to the body or in private and intimate spheres (e. g. bathrooms and bedrooms), on the probability that their use will affect particularly vulnerable persons (e. g.

children and young people, care-dependent persons, persons with disabilities), and on the extent to which they encroach upon an individual's personality. The high level of responsibility (or autonomy) granted to or demanded from the users who assemble, configure and operate these devices represents a particular challenge when attempting to design technologies that foster self-determination.

The Data Ethics Commission recommends that the Federal Government should step up its support for R&D efforts on **technical standards** for end devices. It also urges the Federal Government to lobby at European level for the introduction of **technical requirements** aimed at safeguarding self-determination and product safety in the private sphere, with particular reference to **end devices for consumers**. The Data Ethics Commission takes the view that the following principles should, as a minimum, be enshrined in any end device requirements that are adopted:

- Products must be protected against **cyber attacks and improper use** of data; the measures taken must be commensurate with the need for protection and comply with the state of the art, and suitable guarantees must be provided in particular for sensitive data (e.g. health data). A high level of cyber resilience must be achieved, and this is a joint task incumbent upon the State, industry and each individual.

- Users must be able at all times to identify the **functions that are currently enabled**; in particular, they must be able to see whether the camera, microphone, GPS or other sensors are switched on, whether the device is connected to the Internet, and whether their data are being transferred outside a closed local area.

- It must be easy to turn off **data transfers**, including transfers outside the local area, and data that are stored locally after this function is switched off must not be transferred without the user's consent when it is next switched on (and the same must also be true for individual applications, e.g. on smartphones or smart TVs).

- If **basic device functions** are technically possible **without data transfers of this kind**, the functions must remain available when data transfers are turned off (e.g. a smart fridge must continue to keep its contents cool).

- Devices should be supplied with **"user onboarding"** software; onboarding should take place automatically when the devices are first put into operation, and it should be possible to repeat the onboarding process as often as necessary, even for second users. The information provided to users should cover not only the mode of operation, but also the collection and further processing of user data.

- If end devices have a direct connection to the Internet (e.g. routers) and are secured using a password, it should not be possible to put them into operation without changing the factory password beforehand. On the system side, **passwords** should be allowed only if they comply with the state of the art.

Comprehensibility and transparency

Data protection by design also encompasses the comprehensibility and transparency of systems, including the applications, scripts, sources and elements for each point in time during the development procedure and the process itself. The Data Ethics Commission welcomes the ongoing efforts to develop best-practice models for good terms and conditions of business and "one-pagers" for consumers. As part of a multi-level approach, consumers should initially be provided with simple and "boiled-down" information on the most important data processing operations; if necessary, they should then be informed in detail about the general terms and conditions of business and data protection measures. On its own, however, this approach will not solve the underlying problem, which is that the information provided often fails to do its job, either because it is inadequate and/or because it exceeds the consumer's capabilities.

So that consumers can make informed purchase decisions, standardised, machine-readable and readily understandable graphical symbols **(icons)** should be introduced at European level, following broad consultations with industry and civil society. These icons should convey the key digital characteristics of products (including digital products such as apps) and services; "Basic functions available only with Internet connection", "Internet connection required for enhanced functions", "User data transfers" and "User tracking" are examples of possible characteristics. The icons could also be **colour coded**, which would be particularly useful in the case of product characteristics that apply to a greater or lesser degree. The Data Ethics Commission recommends that the Federal Government should lobby the European Commission to develop standardised icons of this kind, in keeping with Article 12(8) GDPR.

Increased transparency for consumers could also be achieved by supporting the development of certified **electronic shopping assistants**, which would identify a product in a brick-and-mortar or online shop and then serve up product information to the consumer in a format that he or she is likely to understand.

The way in which products, services and applications are designed has a huge influence on the extent to which controllers and processors are able to comply with the data protection obligations incumbent upon them, and yet manufacturers that are not directly responsible for processing personal data fall outside the scope of the GDPR. Controllers that cannot or do not want to use solutions they have developed themselves must therefore insist on "baked-in" data protection.[20] With this in mind, the Data Ethics Commission recommends that the Federal Government should either take steps itself or support action by other parties with the aim of **forcing manufacturers to shoulder a greater share of the responsibility**. Suitable measures might include the following:

- direct imposition by the legislator of **product design and product safety requirements**;
- new and **effective legal remedies** along the distribution chain that can be used to shift the burden of responsibility for inadequate data protection by design and by default onto manufacturers[21] (whereby a certain amount of progress has been made in the new Directive (EU) 2019/771 on certain aspects concerning contracts for the sale of goods in terms of shifting the burden of responsibility from consumers onto retailers and along the distribution chain);

20 Cf. Recital 78 of the GDPR.

21 Christiane Wendehorst: Verbraucherrelevante Problemstellungen zu Besitz- und Eigentumsverhältnissen beim Internet der Dinge, Teil 2: Wissenschaftliches Rechtsgutachten [Consumer-oriented problems relating to possession and ownership structures in the Internet of Things, Part 2: Scientific legal opinion], Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen [Studies and opinions on behalf of the Advisory Council for Consumer Affairs], December 2016, p. 120 (available at: *http://www.svr-verbraucherfragen.de/wp-content/uploads/Wendehorst-Gutachten.pdf*).

- **calls for tenders** and guidelines for **public procurement measures** that are designed in such a way as to require evidence of all-round compliance with the GDPR, including the principles of data protection by design and by default;

- **incentives** that encourage compliance with particularly high standards of data protection by design and by default, for example requirements to this effect in government funding programmes.

3.6.2 Privacy-friendly product development

The importance of data protection by technical design must also be taken into account at the product development and enhancement stages. This applies, in particular, to the **development of algorithmic systems**, since these latter typically require data in bulk, for example to use as training data (→ see Part C, section 2.2 for further details).

Privacy-friendly training of algorithmic systems

Various options are available for complying with the principles of data protection enshrined in Article 5 GDPR while training algorithmic systems. In January 2018, for example, Datatilsynet (the Norwegian data protection authority) proposed privacy-friendly means and methods for the training of algorithmic systems:[1]

8. use of **data minimisation procedures** in relation to training data, e. g. through the use of synthetic data (using generative adversarial networks, for example), through federated learning or through the use of data-minimising variants such as those proposed for neural networks;
9. use of **encryption procedures** such as differential privacy, homomorphic encryption or other procedures that allow the retrieval of information without granting full access to the database;
10. use of **procedures that promote transparency** to achieve a higher level of comprehensibility and traceability.

The Data Ethics Commission believes that **research is still needed** in all of these areas, and this also applies to options for the privacy-friendly testing of algorithmic systems.

1 Datatilsynet: Artificial intelligence and privacy, Report, January 2018, pp. 27 et seq. (available at: *https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf*).

Summary of the most important recommendations for action

Standards for the use of personal data

1

The Data Ethics Commission recommends that **measures be taken against ethically indefensible uses of data**. Examples of these uses include total surveillance, profiling that poses a threat to personal integrity, the targeted exploitation of vulnerabilities, addictive designs and dark patterns, methods of influencing political elections that are incompatible with the principle of democracy, vendor lock-in and systematic consumer detriment, and many practices that involve trading in personal data.

2

Data protection law as well as other branches of the legal system (including general private law and unfair commercial practices law) already provide for a range of instruments that can be used to prevent such ethically indefensible uses of data. However, in spite of the widespread impact and enormous potential for harm, too little has been done to date in terms of harnessing the power of these instruments, particularly against the market giants. The various factors contributing to this **enforcement gap** must be tackled systematically.

3

As well as steps to make front-line players (e.g. supervisory authorities) more aware of the existing options, there is an urgent need for the **legislative framework in force to be fleshed out more clearly and strengthened in certain areas**. Examples of recommended measures include the blacklisting of data-specific unfair contract terms, the fleshing out of data-specific contractual duties of a fiduciary nature, new data-specific torts, the blacklisting of certain data-specific unfair commercial practices and the introduction of a much more detailed legislative framework for profiling, scoring and data trading.

4

In order to allow supervisory authorities to take action more effectively, these authorities need significantly better human and material resources. Attempts should be made to strengthen and formalise cooperation between the different data protection authorities in Germany, thereby ensuring the uniform and coherent application of data protection law. If these attempts fail, consideration should be given to the **centralisation of market-related supervisory activities** within a federal-level authority that is granted a broad mandate and that cooperates closely with other specialist supervisory authorities. The authorities at *Land* level should remain responsible for supervisory activities relating to the public sector, however.

5

The Data Ethics Commission believes that **"data ownership"** (i. e. exclusive rights in data modelled on the ownership of tangible assets or on intellectual property) would not solve any of the problems we are currently facing, but would create new problems instead, and **recommends refraining from their recognition**. It also advises against granting to data subjects copyright-like rights of economic exploitation in respect of their personal data (which might then be managed by collective societies).

6

The Data Ethics Commission also argues that **data should not be referred to as "counter-performance"** provided in exchange for a service, even though the term sums up the issue in a nutshell and has helped to raise awareness among the general public. Regardless of the position that data protection authorities and the European Court of Justice will ultimately take with regard to the prohibition under the GDPR of "tying" or "bundling" consent with the provision of a service, the Data Ethics Commission believes that consumers must be offered **reasonable alternatives** to releasing their data for commercial use (e. g. appropriately designed **pay options**).

7

Stringent requirements and limitations should be imposed on the use of data for **personalised risk assessment** (e. g. the "black box" premiums in certain insurance schemes). In particular, the processing of data may not intrude on intimate areas of private life, there must be a clear causal relationship between the data and the risk, and the difference between individual prices charged on the basis of personalised and non-personalised risk assessments should not exceed certain percentages (to be determined). There should also be stringent requirements in respect of transparency, non-discrimination and the protection of third parties.

8

The Data Ethics Commission advises the Federal Government not to consider the issues falling under the heading of **"digital inheritance"** as having been settled by the Federal Court of Justice's 2018 ruling. The ephemeral spoken word is being replaced in many situations by digital communications that are recorded more or less in their entirety, and the possibility that these records will be handed over to a deceased's heirs adds a whole new dimension of privacy risk. A range of mitigating measures should be taken, including the imposition of new obligations on service providers, quality assurance standards for digital estate planning services and national regulations on post-mortem data protection.

9

The Data Ethics Commission recommends that the Federal Government should invite the social partners to work towards a common position on the legislative provisions that should be adopted with a view to **stepping up the protection of employee data**, based on examples of best practices from existing collective agreements. The concerns of individuals in non-standard forms of employment should also be taken into account during this process.

10

In view of the benefits that could be gained from **digitalising healthcare**, the Data Ethics Commission recommends swift expansion of digital infrastructures in this sector. The expansion of both the range and the quality of digitalised healthcare services should include measures to better allow patients to exercise their rights to informational self-determination. Measures that could be taken in this respect include the introduction and roll-out of an electronic health record, building on a participatory process that involves the relevant stakeholders, and the further development of procedures for reviewing and assessing digital medical apps in the insurer-funded and consumer-funded health markets.

11

The Data Ethics Commission calls for action against the significant enforcement gap that exists with regard to statutory **protection of children and young people** in the digital sphere. Particular attention should be paid to the development and mandatory provision of technologies (including effective identity management) and default settings that not only guarantee reliable protection of children and young people but that are also family-friendly, i. e. that neither demand too much of parents or guardians nor allow or even encourage excessive surveillance in the home environment.

12

Standards and guidelines on the handling of the personal data of **vulnerable and care-dependent persons** should be introduced to provide greater legal certainty for professionals in the care sector. At the same time, consideration should be given to clarifying in the relevant legal provisions on living wills that these may also include dispositions with regard to the future processing of personal data as far as such processing will require the care-dependent person's consent (e. g. for dementia patients who will not be in a position to provide legally valid consent).

13

The Data Ethics Commission believes that a number of binding requirements should be introduced to ensure the **privacy-friendly design of products and services**, so that the principles of privacy by design and privacy by default (which the GDPR imposes on controllers) will already be put into practice upstream, by manufacturers and service providers themselves. Such requirements would be particularly important with regard to consumer equipment. In this context, standardised icons should also be introduced so that consumers are able to take informed purchase decisions.

14

Action must also be taken at a number of different levels to provide manufacturers with adequate **incentives to implement features of privacy-friendly design**. This includes effective legal remedies that can be pursued against parties along the entire distribution chain to ensure that also manufacturers can be held accountable for inadequate application of the principles of privacy by design and privacy by default. Consideration should also be given, in particular, to requirements built into tender specifications, procurement guidelines for public bodies and conditions for funding programmes. The same applies to **privacy-friendly product development**, including the training of algorithmic systems.

15

While debates on data protection tend (quite rightly) to centre around natural persons, it is important not to ignore the fact that **companies and legal persons must also be granted protection**. The almost limitless ability to pool together individual pieces of data can be used as a means of obtaining a comprehensive picture of a company's internal operating procedures, and this information can be passed on to competitors, negotiating partners, parties interested in a takeover bid and so on. This poses a variety of threats – *inter alia* to the digital sovereignty of both Germany and Europe – in view of the significant volumes of data that flow to third countries. Many of the Data Ethics Commission's recommendations for action therefore also apply on a *mutatis mutandis* basis to the data of legal persons. The Data Ethics Commission believes that action must be taken by the Federal Government to **step up the level of data-related protection afforded to companies**.

4. Improving controlled access to personal data

All types of data (both personal and non-personal) represent a **key resource** within the data economy and serve as a vital ingredient in many applications that foster the public good. The breakneck speed of development of digital technologies – some of which benefit each and every one of us enormously – can be attributed in part to the ability to evaluate data generated by billions of users. Although data protection must always remain the central priority for applications involving personal data, more and more people are asking whether general improvements in the area of controlled access to personal data might be ethically tenable or even desirable, in keeping with the principle of data use and data sharing for the public good (→ section 1.3 above) and within the framework prescribed by data protection law.

4.1 Enabling research that uses personal data

4.1.1 Preliminary considerations

Research serves as the basis for almost all our technical achievements, and the current onslaught of digitalisation means that data-based research is becoming **increasingly important**. Its significance has already been recognised by the GDPR, backed up in certain cases by national law (i.e. the [German] Federal Data Protection Act (*Bundesdatenschutzgesetz*, BDSG) and the data protection acts of the *Länder*). The Data Ethics Commission wishes to emphasise the fact that data processing operations involving genetic, biometric and other **health data** are of enormous value in terms of furthering research goals, promoting preventive methods and developing new diagnostic and therapeutic approaches. The use of artificial intelligence holds the promise of significant progress in certain areas, but – depending on the problem being tackled – may rely on large pools of data. The issue of releasing health data for research purposes (referred to as **"data donation"**) is a recurrent topic of debate. This term "data donation" is **misleading**, however, because data that have been donated – unlike organs or money – can be reused as often as necessary and in parallel, even by the data donor himself or herself.

Provided that the research can, for the most part, be described as a public-good activity in terms of the way that it uses data (e.g. for providing healthcare services, developing sustainable mobility concepts or improving living conditions in the broader sense), the Data Ethics Commission recommends that full use should be made of the existing **privileges under data protection law**, and that research should be viewed as a particularly valuable good when weighing it up against competing interests.[22] It additionally recommends that the *Länder* should exercise the regulatory powers they already hold (for example in the area of higher education law or within the framework of data protection law) in such a way as to foster innovation and in keeping with the aforementioned notion of special privileges for research. A broad interpretation should be placed on the term "scientific research" in this context, *inter alia* with reference to consistent past decisions by the Federal Constitutional Court, and it should be irrelevant whether the research in question is being carried out by government-funded or private institutions.

The Data Ethics Commission wishes to point out that – challenging though the task may be – an **appropriate balance** must be sought between the researchers' fundamental rights and the data subjects' right to informational self-determination. When carrying out the weighing up of interests required by law, special priority should be accorded to the **protection of sensitive data** and the associated rights of data subjects such as patients and insured parties. For example, the duty of confidentiality imposed on certain individuals (such as doctors) who are subject to a code of professional secrecy (cf. Section 203 of the [German] Criminal Code (*Strafgesetzbuch*, StGB)) may also apply to the work of research institutions if these latter use data collected or stored by the individuals in question. The procedural precautions imposed by law with a view to protecting the right to informational self-determination would then need to be observed.

22 Cf. Conference of the Independent Data Protection Authorities of the Federal Government and the *Länder*: Orientierungshilfe der Aufsichtsbehörden für Anbieter von Telemedien [Guidance by the supervisory authorities for telemedia providers], March 2019, p. 14 (available at: *https://www.datenschutzkonferenz-online.de/media/oh/20190405_oh_tmg.pdf*).

4.1.2 Legal clarity and certainty

Although the law as it currently stands permits and promotes data-based research, **questions of interpretation** arise in relation to certain details, and these questions require further clarification by the supervisory authorities and courts. For example, it has yet to be definitively clarified whether the **further processing** of data that have already been lawfully collected for one purpose (e.g. healthcare provision) can – on the basis of Article 5(1)(b) GDPR and, in the light of Recital 50, with "appropriate safeguards" within the meaning of Article 89 GDPR – automatically be deemed lawful if they are processed for research purposes, or whether the requirement for a separate legal basis pursuant to Article 6(1)–(3) or Article 9 GDPR applies just as it did when the data were first collected (for example, Section 27 of the Federal Data Protection Act states that health-related data can be processed only if express consent has been provided or if the research interests "substantially outweigh" the data subject's interests). It has also been suggested in certain quarters that the right to process the data further can only be invoked by the party that collected the data in the first place; similar uncertainty reigns over the scope of the term "research" as regards **product development and enhancement**.

Even though a legal framework exists for data-based research in Germany, *inter alia* in relation to health-related data and other special categories of data, the finer details of this regulatory framework lack uniformity, if only because the country's federal structure means that both the Federal Government and the *Länder* hold constitutionally enshrined legislative powers. From a research perspective, the resulting **legal uncertainty** is exacerbated yet further by an ongoing lack of reliable guidance, in particular as regards the criteria that must be met in order for consent to be deemed valid and in order for the data subject's interests to be "substantially outweighed" by research interests within the meaning of Section 27 of the Federal Data Protection Act. This legal uncertainty could prove a stumbling block for data-based research in Germany. The Data Ethics Commission believes that **recommendations for action and interpretative criteria** should therefore be developed – perhaps by the Conference of Independent Data Protection Authorities of the Federal Government and the *Länder*, with the involvement of relevant stakeholders from politics, the healthcare industry and civil society – so that the relevant rules can be applied in a feasible and legally compliant way (for further information on pseudonymisation and anonymisation standards, → see section 4.2 below).

With a view to **further harmonisation** aimed at overcoming regulatory discrepancies in the field of research (different regulatory approaches by the Member States, division of regulatory scope between the Federal Data Protection Act and the data protection acts of the *Länder*, special regulations for specific subjects), the Data Ethics Commission recommends that the Federal Government should:

a) push for synchronisation of the research-specific **legal bases** in the Federal Data Protection Act, in the data protection acts of the *Länder* and in subject-specific acts;

b) drive forward projects at **European level** aimed at greater harmonisation of the regulatory frameworks put in place by the Member States in respect of research data protection; and

c) lobby for a **duty of notification** incumbent upon Member States when adopting national laws in this area, and for the establishment of a European **clearing house** for cross-border research projects.

4.1.3 Consent processes for sensitive data

Voluntary, informed and explicit consent by the data subject is a critically important means of protecting individuals (test subjects) participating in research projects, particularly in the case of clinical research and research involving health data and other particularly sensitive categories of data, because it provides the test subject with an opportunity to exercise his or her right to **informational self-determination**. Since it necessitates the provision of easy-to-understand information about the research project, it also ensures that the test subject will not discover at a later date that his or her **values or preferences** prevent him or her from participating in the study. As a protective instrument enshrined in law, it improves the transparency of research and therefore increases people's level of confidence in it. Not least among its benefits is the fact that it also promotes the integrity of research and researchers.

Yet researchers who act as controllers face considerable challenges when it comes to obtaining informed consent, particularly when the project involves sensitive data. For example, if researchers want to embark on a new project using health data already available in a database, the data subjects must be contacted so that consent can be obtained again (unless the data subjects originally consented to the reuse of their data in future, or provided – to use the term preferred within ethics discourse – **broad consent**). Researchers wishing to use health data collected in the course of routine medical care for research purposes must first contact patients and ask them to grant informed consent, which is a task fraught with huge practical obstacles. With this in mind, the Data Ethics Commission recommends that appropriate **model procedures for the obtaining of consent** should be designed and developed with a view to making it easier to process data for research purposes.

With explicit reference to the link that exists between consent and a data subject's fundamental rights, the Data Ethics Commission also calls for the development of **innovative consent models** in the research sector. **Dynamic consent** models that involve tailoring declarations of consent to the individual context are already being trialled, for example. In this connection, it must be ensured that the consenting party remains able to control his or her data even after granting consent; in order to ensure that this is the case, the Data Ethics Commission recommends that more emphasis be placed on the development and design of privacy management tools (PMT) and personal information management systems (PIMS) (→ see section 4.3 below) for the research sector, such as **digital consent assistants** or data agents. Consent assistants of this kind may make it significantly easier for data subjects to keep track of the data processing operations to which they have granted consent, even after these operations have commenced; equally, they may make it possible to go back and ask data subjects for consent again if circumstances change, and to provide data subjects with a straightforward way of revoking their consent.

Calls are being heard increasingly often – particularly in connection with research using health data – for **blanket consent** models that involve a data subject granting consent to a wide range of data uses in the field of research, without reference to a specific course of treatment or other event. Although the research sector can advance compelling reasons for models of this kind, there are a number of concerns and obstacles that must be overcome before they are adopted (in particular the need for consent to be informed and for it to be linked to a specific purpose). They would make it impossible to take a consenting party's preferences and values into account on a differentiated basis, even if far-reaching legal safeguards were provided against misuse of his or her data and encroachments upon his or her privacy.

Against this backdrop, the Data Ethics Commission recommends further discussion of the innovative model known as **"meta consent"**.[23] After being appropriately informed – and without being in a situation where consent is specifically required – the data subject decides on the type of research projects and research contexts for which he or she wishes to grant consent and the type of consent involved (specific or broad). Consent may be limited on the basis of considerations such as the following:

- research context (e.g. private or public research, commercial or non-commercial research, national, European or international research);
- data sources (e.g. electronic health record, human tissue, health data, lifestyle data from wearables);
- type of research (e.g. preventive research, research into cancers or neurodegenerative disorders, any kind of health research).

If researchers later wish to use the data for a specific research project, the data subject is **informed** in advance and given the opportunity to **object** to this use of his or her data.

Each real-life implementation of this model should be under the **oversight** of a data trust scheme, an ethics commission or another responsible body tasked with ensuring that the consenting party's preferences are, in fact, taken into account. It should also be possible for the data subject to amend the terms of his or her meta consent at any time, and the technical and regulatory framework required to do so must be in place.

Example 13

Example 13 A data subject specifies that the data from his electronic health record may be used for public and commercial research. He also specifies that his blood and tissue samples may be used for public and commercial research into degenerative diseases. He consents to the processing of data from his electronic health record provided that the data are not transferred out of Europe. A company from Spain would like to use data from his electronic health record as well as data from his tissue samples for dementia research. The data subject is informed of their intention to do so, and told that he has four weeks to object to his data being used in this way.

23 Thomas Ploug / Søren Holm: Bioethics, 2016 (30:9), pp. 721 et seqq.

When deliberating on and designing a model of this kind, care must be taken to ensure that any constraints placed on the **freedom of research** and the research privilege for secondary use of data are equivalent in scope to the restrictions imposed under the current legal system. Preference should be given to meta consent models that emphasise the ability of data subjects to express their **values and preferences** regarding the use of their health data for research purposes; this would also increase public confidence in health data governance.

Another ethical question that must be considered is that of accountability – not only in relation to the use of data, but also in relation to their non-use, since this may block potential progress in vital areas and result in discrimination against certain groups as a result of their **exclusion from progress**. For example, methodological reasons mean that clinical studies involving older people suffering from several different chronic diseases and taking several different kinds of medication at the same time must necessarily be very limited in scope. If high-quality procedures can be used to evaluate their health data, however, key findings might be obtained on the interactions between these different medications and their actions under everyday conditions; these findings could then be used as a productive basis for more extensive research and the treatment of these patients going forwards.

With the above in mind, and given the significance of the European healthcare sector from both a medical and economic perspective, the Data Ethics Commission recommends proactive **support for a "learning healthcare system"** in which healthcare provision is continuously improved by making systematic and quality-oriented use of the health data generated on a day-to-day basis, in keeping with the principles of evidence-based medicine. A learning healthcare system imposes high requirements in terms of multi-level governance and requires a cross-disciplinary approach to healthcare provision that puts the insured party or patient front and centre.

4.1.4 Legal protection against discrimination

At the same time, however, the Data Ethics Commission wishes to emphasise that all parties involved in developing and designing new health-related research projects must take due account of the significant **potential for discrimination** that is opened up through the availability of sensitive data (e.g. when a data subject looks for a job or takes out an insurance policy). Technical progress has made it possible to sequence and decode the human genome, and data scientists are now able to analyse biometric and behavioural data collected in the course of daily life; this means that it is also possible to profile an individual's risk of falling ill in the future, typically based on the likelihood that he or she will suffer from this or that disease – and when genetic data come into play, his or her relatives may also be affected.

With this in mind, the Federal Government should examine the possibility of **including new grounds for action under the [German] General Act on Equal Treatment (*Allgemeine Gleichbehandlungsgesetz*, AGG)**, as well as specific **bans** on using information about a person's health (by way of analogy to the corresponding provisions on genetic data in the [German] Genetic Diagnostics Act (*Gendiagnostikgesetz*, GenDG)).

4.2 Anonymisation, pseudonymisation and synthetic data

Operations that involve accessing personal data must always comply with the applicable provisions of data protection law, and abide by the rules on data processing laid out in these provisions – from the purpose limitation principle right through to appropriate protective measures. Under certain circumstances, therefore, it may be vitally important for businesses or other users to know for certain that their operations either fall outside the scope of data protection law or are compliant with data protection law. The Data Ethics Commission believes that there is a lack of **legal certainty** in a number of different areas, for example concerning the anonymisation and pseudonymisation of data, the identification and consideration of a link between individuals and (allegedly anonymised) data sets, and synthetic data.

Anonymised and pseudonymised data

Anonymisation involves processing a set of personal data in such a way that any link to the data subject is broken irrevocably. A distinction is made between randomisation and generalisation; both are different ways of approaching the task of anonymisation, and they can be used individually or in combination. **Randomisation** involves modifying data in such a way that the anonymised data can no longer be matched up with the data subject. This can be achieved by falsifying individual data sets, for example. Appropriately designed randomisation methods ensure that the statistical properties of the original data set are retained, for example by swapping values rather than changing them. **Generalisation** involves aggregating pieces of [less] detailed information, such as age categories instead of dates of birth, names of regions instead of postcodes, or periods of time instead of time stamps that are accurate to the nearest second.

Three main strategies are used to identify natural persons in a data set:

a) **singling out:** a method of pinpointing data sets relating to specific individuals from a larger pool of data, for example by using unique characteristics that make it possible to identify these individuals;

b) **linkability:** a method that involves linking up at least two data sets that relate to the same individual or group of individuals on the basis of matching values that appear in both data sets, such as identifiers, spatial coordinates or times. Even a small amount of data available on an individual can be augmented using this linking strategy, allowing him or her to be identified;

c) **inference:** a method that involves deriving the highly probable value of a characteristic from the values of a number of other characteristics, again allowing the data relating to an individual to be augmented and increasing the likelihood that he or she will be identified.

Anonymised data sets make it impossible to recreate the links that once existed between the data and the individuals to whom the data relate, or to create such links for the first time, given the technological means that are reasonably likely to be used and that are available or being developed at the time of the processing (cf. Recital 26 GDPR); an attacker wishing to identify one or all of the data subjects (through de-anonymisation) would find the task impossible.

Modifications to a set of data – in particular the artificial addition of fuzziness (also referred to as noise or blurring, depending on the context) – ensure that it is impossible to pull out data that belong to a specific individual, that linkable data are not used and that inferences cannot be drawn; these modifications typically also place constraints on the utility of the data. If the user is aware of the evaluations that will later be carried out using the data set, the anonymisation procedures can be optimised with this in mind, for example by retaining the necessary level of detail for the relevant characteristics wherever possible. The same applies to comparisons of different data sets (interoperability); if the user knows which comparisons will be carried out, appropriate anonymisation methods can be designed by categorising the data into identical groups as required, and taking into account the increase in risk that may occur as a result of incorporating information from other data sets.

Pseudonymisation involves processing data in such a way that they can no longer be assigned to a specific data subject without additional information, which may take the form of mapping tables or cryptographic hash methods, for example. Pseudonymisation differs from anonymisation in that a reference to a person (in the legal sense of the term) is retained. The controller must prevent (unauthorised) access to the additional information whenever the pseudonymised data are processed in future, since otherwise it would be possible to map the data to the data subjects. The GDPR refers to pseudonymisation several times as a technical and organisational measure for reducing the risk to the rights and freedoms of natural persons.

Both anonymisation and pseudonymisation involve processing a set of data that is already available, and must be distinguished from **pseudonyms**, which are deployed on the user side. Users may choose their own pseudonymised identifiers (e. g. user names for online services or e-mail addresses), or use identifiers provided automatically by a technological system, for example the online ID function of an electronic ID card or attribute-based authorisation certificates designed with data protection concerns in mind. In the vast majority of cases, the use of pseudonyms provides little in the way of protection against identification of the data subject, particularly if they are used across contexts and communication partners, which allows the data in a user-specific profile to be linked to other data and augmented. Conversely, constantly changing "transaction pseudonyms" are restricted to a specific context, making it much harder to identify the individual in question.

Internet-based procedures aimed at **concealing the link** between a data subject and the data relating to that data subject cannot generally be regarded as anonymisation in the strict sense of the term, but may nevertheless provide some level of protection against identification and observation. Simple web proxies make it possible to surf the Internet using the identifier (i. e. the IP address) of an intermediary server; multiple users (whose identifiers are known to the proxy server) may therefore have the same identifier as far as the destination web servers are concerned, provided that they avoid identifying themselves through the use of cookies, etc. Further steps to prevent identification can be taken by arranging multiple intermediary servers one behind another, for example in mix networks such as Tor or in mix cascades such as JonDo. Once again, noise can be added by sending artificially created "dummy traffic", as an additional obstacle in the path of anyone attempting to observe the human users.

4.2.1 Procedures, standards and presumption rules

It is often not possible to **anonymise** data – i.e. completely break the link between data and the data subject to whom they belong in such a way that it cannot be recreated – without losing any of the data's utility. At the same time, however, perfect anonymisation is often not required, firstly because many goals can (upon closer examination) be achieved using data with a somewhat lower level of utility, and secondly because the GDPR already contains exemptions for data processing operations that serve the public good (e.g. in the research sector), meaning that even personal data can be processed without obtaining consent from the data subjects. Nevertheless, efforts aimed at developing effective **anonymisation technologies and procedures** should be stepped up with a view to allowing data to be processed wholly outside the scope of the GDPR.

Ultimately, legal certainty can be achieved only by developing **standardised technologies and procedures**, which must always take due account of the whirlwind pace of technological development. The Data Ethics Commission therefore recommends that the Federal Government should lobby – in particular at EU level – for easy-to-use **anonymisation standards** that would benefit both data subjects and users, and for **pseudonymisation measures** that are commensurate with the level of risk faced by data subjects in their private lives (as featured on the agenda of the Federal Government's Digital Summit).

In particular, anonymisation standards should be combined with clear rules imposing a rebuttable legal **presumption**, which would provide legal certainty for users, who could rely on their data processing operations falling outside the scope of the GDPR where the standard has been met. In this context, it is important to remember that restrictions may need to be imposed in these presumption rules, for example on the period of validity (by way of analogy to cryptographic procedures),[24] or on the authorised methods of processing (for example stating that data may not be published or made accessible to an unspecified number of people). As long as there is no legal basis for rebuttable presumption rules, the Federal Government should support the development of technical **best practices** and industry-specific **codes of conduct**, with a view to building up experience in these fields.

In certain fields, the standardisation of anonymisation and pseudonymisation procedures may also impose rules on the way in which the link between a data subject and the data relating to him or her should be broken, making it possible to compare different data sets and improving **interoperability**. At least in areas where improved interoperability is a sought-after outcome, the Data Ethics Commission recommends that context-specific rules should be developed for preferred groupings (e.g. value ranges of age categories, postcodes or IP addresses). A similar approach is already followed by Germany's statistical offices when handling data, for example.

24 Federal Office for Information Security: Technical Guidelines BSI TR-02102 Cryptographic Mechanisms: Recommendations and Key Lengths, last updated in February 2019 (available at: *https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/TechGuidelines/TG02102/BSI-TR-02102-1.pdf?__blob=publicationFile&v=9*).

Anonymisation and pseudonymisation procedures are carried out on repositories of data that are known to (or at least suspected to) contain personal data. These differ from repositories of data that are not thought to contain personal data, but could be used as a means or at least a starting point (either on their own or in combination) for creating a link between purportedly anonymous data and the data subject to whom these data belong. Once again, the Data Ethics Commission recommends the development and binding implementation of **standardised methods for checking whether data subjects can be identified from a set of data**; these methods must allow the user to conclude with a reasonable degree of certitude that the data are either personal or non-personal.

4.2.2 Ban on de-anonymisation

Presumption rules should also be accompanied by appropriate **bans on de-anonymisation**, and any infringement of these bans (i. e. cases where it proves possible to identify a data subject using formerly anonymous data, for example as a result of technological developments) should be subject to a **penalty**. The bans would need to be designed in such a way as to avoid placing roadblocks in the way of research into the detection and removal of links between data and data subjects in repositories of data, since any options for de-anonymisation that are available must be investigated further with a view to developing appropriate anonymisation standards and verifying their effectiveness. In addition, the introduction of bans on de-anonymisation and penalties for their infringement must not be misused as a pretext for downgrading the standards that apply to anonymisation or diluting the meaning of the term "personal data" as used in the GDPR, since companies involved in vitally important efforts to drive forward the technology of anonymisation using technical means would otherwise be placed at a competitive disadvantage. The same applies to the reversal of pseudonymisation in the absence of justified reasons (a list of which should be drawn up).

4.2.3 Synthetic data

A distinction should be made between genuine data and **synthetic data**, i. e. data that are generated artificially rather than being collected directly in the real world. Synthetic data boast several advantages over real-world data;[25] firstly, they can be produced in any quantity, which is particularly important when dealing with simulations for which real-world data cannot be generated. Secondly, steps can be taken when synthetic data are created to ensure that the entire range of values is mapped as comprehensively as possible, e. g. in order to test how a technical system would behave when confronted with unusual data combinations. Thirdly, the quality of synthetic data can be measured, and if necessary it can be guaranteed in individual cases that the properties of a set of real-world reference data are retained; alternatively, distortions occurring in sets of real-world data can be pinpointed and removed in order to avoid discrimination. If the set of synthetic data contains no references to persons, it is anonymous and does not fall within the scope of the GDPR.

The Data Ethics Commission recommends that the Federal Government should support **research in the field of synthetic data** on a number of different issues, including the question of whether, to what extent and in which contexts synthetic data might replace real-world data in processing operations, and how closely the synthetic data should resemble the real-world data in terms of their properties. The Data Ethics Commission recommends further investigations into the creation and use of synthetic data, with a particular emphasis on topics including data quality and the avoidance of bias and discrimination.

25 Jörg Drechsler / Nicola Jentzsch: Synthetische Daten: Innovationspotenzial und gesellschaftliche Herausforderungen [Synthetic data: potential for innovation and societal challenges], Stiftung Neue Verantwortung, May 2018 (available at: *https://www.stiftung-nv.de/sites/default/files/synthetische_daten.pdf*).

4.3 Controlled data access through data management and data trust schemes

4.3.1 Privacy management tools (PMT) and personal information management systems (PIMS)

In an ever more complex environment, one of the major challenges faced by individuals in exercising their data rights is a **lack of oversight over personal data** – data subjects typically have no records documenting the times when they have granted consent, for example. Sharing of data by the original data controller can also result in the "scattering" of data, with a significant decrease in transparency and a corresponding increase in data protection risks for the data subjects (→ see section 3.3.6 above regarding the problem of data trading). There are currently not enough standards and software tools that data subjects can use to track and control, on an ongoing basis, who has been granted access and to whom data have been transferred, which would be necessary for them to exercise their data rights effectively.

An increasing number of technical and institutional measures are being proposed in response to this problem. **Privacy management tools (PMT)** range from applications that make consent management easier for users (dashboards, etc.) through to AI tools that automatically implement individual user preferences ("data agents"). Where the focus is not so much on the provision of technical applications but rather on the service end, it is more common to use the term **personal information management systems (PIMS)**. Such services range from single sign-on services, local data safes and online storage systems through to offers (both comprehensive and less so) for third-party management of user data (data trust models). When designed as data trust models, PIMS may support digital self-determination by shouldering some of the responsibility for exercising the data subject's rights under data protection law, such as granting and withdrawing consent and exercising the right to information, the right to rectify data, the right to erase data, the right to data portability and the right to object. The Data Ethics Commission recommends that the Federal Government should promote innovation and standardisation in relation to software tools and services of this kind.

4.3.2 Need for regulation of PMT/PIMS

The above notwithstanding, privacy management tools/personal information management systems may pose **risks** if they fail to comply with certain requirements, some of which go beyond the scope of the GDPR. If these tools or systems fail to be properly designed, for example, there is a risk that data subjects will not be empowered to exercise true self-determination, but will instead unwittingly find themselves on a path of **external determination**. In particular, privacy management tools/personal information management systems that are designed in such a way that data subjects "write a blank cheque" by handing over the majority of decisions to the operators of these tools/systems, or that result in data subjects taking decisions contrary to their own interests under the influence of these tools/systems, would ultimately be inconsistent with the ethical value of self-determination. Privacy management tools/personal information management systems must be available as aids for data subjects, but they must not usurp the power of these latter to take self-determined decisions, and they must certainly not manipulate them using dark patterns et al. (→ see section 3.2.2. above).

Given the significant risks that these systems and tools may pose to fundamental rights and the lack of options for data subjects to carry out quality assurance measures themselves, the Data Ethics Commission recommends that the Federal Government should develop **quality standards for privacy management tools/personal information management systems and introduce a certification and monitoring system**. The latter should apply in particular to systems that act on behalf of or in place of a data subject, or that – as a result of their technical design – play a major role in steering and channelling the data subject's decisions. In cases where data are stored directly by the operators of these tools/systems (i. e. if they are not stored on a decentralised basis and simply managed, which is also possible), provision must also be made for the company's insolvency or liquidation.

Privacy management tools/personal information management systems can operate reliably only if cooperation on the part of all relevant controllers is guaranteed. The only possibility to achieve the wide-ranging coverage required is by imposing a **legal obligation** that applies (under appropriate conditions) to controllers within the meaning of the GDPR, with a view to ensuring that any access to personal data can be monitored by the tool/system and that any information that is relevant in terms of data protection reaches the tool/system so that the tool/system can effectively protect the data subject's interests in relation to all of his or her personal data. A **sector-specific approach** – for social networks, for example – might be a realistic option to start with.

In the view of the Data Ethics Commission, systems of this kind could either be operated on a non-profit basis and without any involvement of commercially motivated actors – such as by charitable **foundations** and similar independent bodies – or organised as **private-sector enterprises** provided that the operator derives profits from managing rather than from using the data. In either case, the fiduciary duties that are owed to the data subject must be precisely defined in legislation, the involvement of parties with conflicting interests must be ruled out, and appropriate opportunities for oversight must be built into the system as a whole (such as to minimise bias and discrimination). If the private-sector option is chosen, it will also be necessary to ensure that the operator's commercial motivations do not undermine the role it plays as custodian of the data subject's interests, and that operators that have access to personal data are based in the European Union.

The Data Ethics Commission recommends that the Federal Government should lobby for appropriate **amendments to the GDPR** in the form of a clearer and legally secure framework for privacy management tools/personal information management systems. Steps should also be taken (in addition to action on legal matters relating to mandates, etc.) to prevent excessive centralised storage of personal data, since arrangements of this kind increase the level of risk for data subjects in the event of cyber attacks or similar incidents. Machine-interpretable formats and communication protocols must be standardised for the automated execution of services.

4.3.3 PMT/PIMS as a potential interface with the data economy

Provided that the appropriate regulations are adopted, privacy management tools/personal information management systems could also serve a dual function. On the one hand, these tools/systems might help individuals to exercise their right to informational self-determination effectively and to verify compliance with any limitations on use that have been imposed; on the other hand, however, they could also be used to release data from the confines of "data silos" and allow them to be used within the European data economy (in particular by exercising the right to data portability granted by Article 20 GDPR). The main idea underlying privacy management tools/personal information management systems is to improve an individual's control over his or her personal data, which does not in and of itself promote third-party data access. An indirect data access function might, however, be compatible with the **principle underpinning data trust schemes** if third parties were allowed to access the data only to pursue certain purposes approved by the data subject (→ in connection with research, for example; see section 4.1.3 above), or if the economic exploitation of the data served the data subject's interests and took place with his or her express consent (→ see section 3.3 above for a discussion of the problems raised by treating personal data as an economic asset).

The Data Ethics Commission believes that – if it is decided that privacy management tools/personal information management systems should play a dual role and also serve as a platform for legally secure data access by companies – it must be ensured that these qualified dual-function tools/systems do not ultimately subvert the goal of protecting data subjects' rights. Strict compliance with the principles of privacy and ethics by design must be enforced; in particular, the objective pursued must not be the broadest possible exploitation (and "scattering") of data. The Data Ethics Commission wishes to emphasise the fact that privacy management tools/personal information management systems must continue to serve as dedicated custodians of data subjects' interests, and that **conflicts of interest must be ruled out**.

4.4 Data access through data portability

4.4.1 Promotion of data portability

The **right to data portability** granted by Article 20 GDPR is a tool that a data subject can use to determine whether companies should gain access to his or her personal data which another company has already collected, and if so, which companies. It includes the right to receive the data provided in a "structured, commonly used and machine-readable format" or to have them transmitted directly to another controller. The right to data portability has two main implications:

a) It prevents unwanted lock-in effects if data subjects switch providers, thereby protecting both the individual data subjects' right to economic self-determination and free competition.

b) Even if data subjects do not switch providers, it allows them to ask the controller to make the data available either to them or to other companies. This provides the other companies with an option for gaining access to data that might otherwise not have been available to them, bearing in mind that they need a separate legal basis for data processing under data protection law (e.g. consent or a contract).[26]

Despite the fact that providing data in a "structured, commonly used and machine-readable format" is a basic prerequisite that must be met in order for data subjects to exercise the right to data portability effectively, to date this requirement has been subject to an enormous range of varying interpretations in practice. The Data Ethics Commission therefore recommends that the Federal Government and the data protection authorities – in implementation of Recital 68 of the GDPR – should support the development of **industry-specific codes of conduct and standards** at European level so that the right to data portability can be realised uniformly and effectively in practice, to the benefit of all parties involved.

In the absence of new intermediaries (→ see section 4.3 above), the stimulus to exercise the right to data portability often stems from a company that has gained a new customer. Companies that offer a convenient and automated process for data subjects to exercise their right to data portability are likely to be particularly successful (e.g. a provider of a map service that allows data to be ported from a mobility service provider at the click of a button). There are also grounds for assuming – in view of the potential for network effects and effects of scale – that the companies likely to benefit most from the right to data portability, at least in the medium term, will be those that already hold a dominant position in the market and have accumulated huge amounts of data. The Data Ethics Commission therefore recommends that the Federal Government should **observe developments closely** and, in so far as it judges necessary, lobby at European level for measures that specifically encourage and facilitate the porting of data from market-dominant and data-rich companies to other market participants, including start-ups.

26 For an example of the debates on the requirement for a separate legal basis of this kind under data protection law, see Article 29 Data Protection Working Party: Guidelines on the right to data portability, WP 242, rev. 01, last revised and adopted on 5 April 2017, p. 7 (available at: *http://ec.europa.eu/newsroom/document.cfm?doc_id=44099*).

4.4.2 Should the scope of the right to data portability be extended?

Debates are ongoing on whether the scope of the right to data portability should be extended in various ways, in particular by expanding it to cover data other than the (raw) data provided to a controller (e.g. certain forms of processed or derived data), or by widening it to include a right to dynamic real-time portability (e.g. real-time streaming of data flows). As things currently stand, and further to the above recommendation, the Data Ethics Commission proposes that the Federal Government should not lobby for amendments to the GDPR aimed at extending the scope of the current right to portability; given that the GDPR has been in force for such a short period of time, a "wait and see" approach should instead be adopted until more clarity has been gained on its practical application, supervisory practice by the data protection authorities and interpretation by the courts.

4.4.3 From portability to interoperability and interconnectivity

Network effects (e.g. in the case of messenger services) mean that data portability alone will not be sufficient to mitigate the risks posed by existing and future data and service oligopolies, or to lower the barriers to market entry for new competitors to the extent that they represent a serious challenge to the market-dominant providers. The Data Ethics Commission therefore recommends that the Federal Government should push for the introduction of **sector-specific interoperability obligations**, of the sort that have previously been imposed for postal services and mobile telephony, for example. At the same time, measures must be taken to ensure that interoperability features comply with data protection principles, such as privacy-friendly default settings; examples include an option to use different and changing identifiers instead of a single universal identifier, a reduction in the use of central components to collect large volumes of data, and other suitable examples of interoperable technical interaction at different levels.

Asymmetric interoperability obligations could be imposed on **powerful companies and new market entrants** respectively (for example, a market-dominant provider of messenger services might be obliged to allow customers of smaller providers to send messages directly to its own customers and to allow its own customers to send messages directly to the customers of smaller providers); at the same time, however, it must be ensured that interoperability requirements are not abused for the purpose of increasing yet further the flow of personal data towards data-rich and powerful companies. If this risk can reliably be averted, it would be useful to impose certain **interconnectivity obligations, e.g. for short messaging services and social networks**, with a view to counteracting the concentration effects of these networks and promoting the aims of data portability more effectively (i.e. healthier competition and easier access for new market entrants to a data-intensive economy). A model of this kind is also a prerequisite for building up or strengthening certain basic services of an information society in Europe, thereby promoting the digital sovereignty of both Germany and Europe.

4.5 Crowdsensing for the public good

Crowdsensing has also been hailed as a way of opening up new data resources for the data society and data economy; in order to do so, it deploys users' technical devices in the form of "sensors" that collect data (in a certain locality, for example) and forward them to a higher-level instance that analyses the collected data. The Data Ethics Commission acknowledges the potential inherent to this technology, especially if it is **put to use for the public good**. For example, crowdsensing can be used in a smart city for real-time analysis of traffic conditions, the state of repair of infrastructure, air quality and so on. At the same time, however, the Data Ethics Commission believes that achieving an ethically appropriate design will be a significant challenge. An analysis carried out using crowdsensing techniques will typically have an extremely high level of granularity, meaning that the data involved may fall under the category of "sensitive" not only from the perspective of the individuals that generated them, but under certain circumstances also from the perspective of anyone in their vicinity. Efforts must therefore be stepped up to introduce **standards for anonymisation and pseudonymisation** (→ see section 4.2 above) with a view to preventing not only situations in which data can be traced back to (non-consenting) users or potentially to other persons affected, but also other forms of misuse. Crowdsensing-related data transfers may also overstrain the resources of users' devices and raise security issues.

Consideration must be given to these points even if users participate voluntarily and intentionally in crowdsensing programmes ("participatory sensing"), and thought must therefore be given to the **substantive limitations to consent** that exist in this connection (→ see section 3.2 above). Even when data are used for purposes that serve the public good, it must always be ensured that the requirements outlined in legislation – in particular data protection law and consumer protection law – are complied with in full. In this case, it should also be remembered that decisions and measures taken by government agencies must not be based solely or customarily on data collected using participatory sensing techniques, since these data are necessarily **incomplete** owing to the voluntary nature of participation, and it is likely that they will also exhibit **bias**.

The Data Ethics Commission believes that any discussion of whether crowdsensed personal data should be collected, forwarded and compiled without the user's knowledge ("opportunistic sensing") ignores the potential for such measures to violate the fundamental principles of data protection. It believes that decisions must be taken on a case-by-case basis as to whether a **legal obligation** can justifiably be imposed to force data subjects to make available technical devices so that the data from these devices can be collected and forwarded automatically, if and to the extent that the analysis of these data promotes vital public interests.

Summary of the most important recommendations for action

Improving controlled access to personal data

16

The Data Ethics Commission identifies enormous potential in the use of data for research purposes that serve a public interest (e.g. to improve healthcare provision). Data protection law as it currently stands acknowledges this potential, in principle, by granting far-reaching privileges for the processing of personal data for research purposes. Uncertainty persists, however, in particular as regards the scope of the so-called research privilege for secondary use of data, and the scope of what counts as "research" in the context of product development. The Data Ethics Commission believes that appropriate **clarifications in the law** are necessary to rectify this situation.

17

The fragmentation of research-specific data protection law, both within Germany itself and among the EU Member States, represents a potential obstacle to data-driven research. The Data Ethics Commission therefore recommends that **research-specific regulations should be harmonised**, both between federal and *Land* level and between the different legal systems within the EU. Introducing a notification requirement for research-specific national law could also bring some improvement, as could the establishment of a European clearing house for cross-border research projects.

18

In the case of research involving particularly sensitive categories of personal data (e.g. health data), **guidelines** should be produced with information for researchers on how to obtain consent in a legally compliant manner, and **innovative consent models should be promoted and explicitly recognised by the law**. Potential options include the development and roll-out of digital consent assistants or the recognition of so-called meta consent, alongside further endeavours to clarify the scope of the research privilege for secondary use of data.

19

The Data Ethics Commission supports, in principle, the move towards a **"learning healthcare system"**, in which healthcare provision is continuously improved by making systematic and quality-oriented use of the health data generated on a day-to-day basis, in keeping with the principles of evidence-based medicine. If further progress is made in this direction, however, greater efforts must be made at the same time to protect data subjects against the significant potential for discrimination that exists when sensitive categories of data are used; this might involve **prohibiting the exploitation of such data** beyond the defined range of purposes.

20

The development of procedures and standards for data **anonymisation** and **pseudonymisation** is central to any efforts to improve controlled access to (formerly) personal data. A legal presumption that, if compliance with the standard has been achieved, data no longer qualify as personal, or that "appropriate safeguards" have been provided in respect of the data subject's rights, would improve legal certainty by a long way. These measures should be accompanied by rules that – on pain of criminal penalty – prohibit the de-anonymisation of anonymised data (e.g. because new technology becomes available that would allow the re-identification of data subjects) or the reversal of pseudonymisation, both in the absence of narrowly defined grounds for doing so. Also research in the field of **synthetic data** shows enormous promise, and more funding should be funnelled into this area.

21

Fundamentally speaking, the Data Ethics Commission believes that **innovative data management and data trust schemes** hold great potential, provided that these systems are designed to be robust, suited to real-life applications and compliant with data protection law. A broad spectrum of models falls under this heading, ranging from dashboards that perform a purely technical function (**privacy management tools**, PMT) right through to comprehensive data and consent management services (**personal information management services**, PIMS). The underlying aim is to empower individuals to take control over their personal data, while not overburdening them with decisions that are beyond their capabilities. The Data Ethics Commission recommends that research and development in the field of data management and data trust schemes should be identified as a funding priority, but also wishes to make it clear that adequate protection of the rights and legitimate interests of all parties involved will require additional **regulatory measures at EU level**. These regulatory measures would need to secure central functions without which operators cannot be active, since their scope for action would otherwise be very limited. On the other hand, it is also necessary to protect individuals against parties that they assume to be acting in their interests, but that, in reality, are prioritising their own financial aims or the interests of others. In the event that a feasible method of protection can be found, data trust schemes could serve as vitally important mediators between data protection interests and data economy interests.

22

As far as the right to **data portability** enshrined in Article 20 GDPR is concerned, the Data Ethics Commission recommends that industry-specific codes of conduct and standards on data formats should be adopted. Given that the underlying purpose of Article 20 GDPR is not only to make it more straightforward to change provider, but also to allow other providers to access data more easily, it is important to evaluate carefully the market impact of the existing right to portability and to analyse potential mechanisms by which it can be prevented that a small number of providers increase yet further their market power. Until the findings of this evaluation are available, expansion of the scope of this right (for example to cover data other than data provided by the data subject, or real-time porting of data) would seem premature and not advisable.

23

In certain sectors, for example messenger services and social networks, **interoperability or interconnectivity obligations** might help to reduce the market entry barriers for new providers. Such obligations should be designed on an asymmetric basis, i.e. the stringency of the regulation should increase in step with the company's market share. Interoperability and interconnectivity obligations would also be a prerequisite for building up or strengthening, within and for Europe, certain basic services of an information society.

5. Debates around access to non-personal data

The data economy will play a key role in the future competitiveness of German and European companies; the growing penetration of the Internet of Things (IoT) and the Internet of Services (IoS) means that data which are collected automatically by sensors and which can potentially serve as a basis for developing new business models and innovations are acquiring ever-greater industrial significance. **Germany is at the cutting edge of developments** as far as many IoT/IoS-related technologies are concerned (e.g. sensor technology, mechanical engineering and embedded systems), and also plays a leading role in the broader field of industrial production and the digital services that cater for this sector; given the increasingly cut-throat nature of international competition, it must build on this head start in order to safeguard the country's future prosperity. A differentiated and robust research landscape, a diversified economic structure and a reputation as a global leader in key technological segments such as Industry 4.0 put Germany in the perfect position to leverage the potential associated with the data economy as a basis for creating future value.

5.1 Appropriate data access as a macroeconomic asset

The Data Ethics Commission believes that providing appropriate access to data for German and European companies and decreasing the current level of dependency on a small number of data oligarchs would go a long way towards building a market-oriented data economy that serves the public good, and towards strengthening the digital sovereignty of both Germany and Europe. In this connection, data access in the narrower sense firstly relates to the extent to which the data required for a particular business model or other project can be **used on a *de jure* and *de facto* basis**. In order to benefit from **access to data** in this narrower sense, however, stakeholders must have a sufficient **degree of data-awareness** and have the **data skills** that are necessary to make use of the data. Also, access to data proves to be disproportionately advantageous to stakeholders that have already built up the **largest reserves of data** and that have the best **data infrastructures** at hand. The Data Ethics Commission therefore wishes to stress that the factors referred to should always receive due attention when discussing whether and how to improve access, in keeping with the **ASISA principle** *(Awareness – Skills – Infrastructures – Stocks – Access)*.

The discussions in this section focus on non-personal data. **Genuinely non-personal data** hold enormous potential for science, the economy and society, and yet this potential is often underestimated. Most scientific data can be categorised as non-personal; these include data originating from the technical sciences (e.g. engineering and materials science), data from the fields of physics (e.g. data from particle accelerators), biology (e.g. data from the plant and animal kingdoms), geology and chemistry, environmental data, weather data and ocean data right through to economic data (e.g. data from the financial markets). If they can be analysed (using big data methods, for example) and used (to develop AI applications, for example), these non-personal data hold enormous value for science, the economy and society; focused support must therefore be provided to researchers using these data, and systematic efforts must be undertaken to make data access an easier task.

The broad nature of the GDPR's definition of "personal data" means that it can safely be assumed that a substantial proportion of data repositories are mixed in nature (i.e. contain both non-personal data and data that are or could become personal); at the same time, the processing of personal data is a vital prerequisite for certain activities that fall under the heading of the data economy and that provide benefits for both individuals and the general public. Any discussion of data access that concentrates solely on non-personal data would therefore appear counter-productive. A more appropriate approach would be to work towards **general data access arrangements** that are **superseded by data protection law** only in cases where personal data are processed (meaning that activities falling under the heading of the data economy would need to comply with the provisions of the GDPR). Equally, it should not be forgotten that the GDPR already allows the economic exploitation of personal data in many circumstances; in addition to consent, for example (Article 6(1)(a) GDPR), there are five additional justifying grounds (Article 6(1)(b)–(f)), some of which are explicitly tailored to economic interests and needs.

5.2 Creation of the necessary framework conditions

5.2.1 Awareness raising and data skills

The use of data to create value presupposes that operators (whether they belong to the private sector or serve a public interest) are adequately well-informed about the relevant options and risks, and also have the data skills required (which may involve drawing on technical, economic, ethical and legal knowledge; (→ see Part D, section 3. above). In certain areas of the German **economy**, companies have still not tapped into the potential that exists to make more productive use of their data flows and repositories (in some cases for the benefit of the public). The Data Ethics Commission welcomes the steps that have been taken to raise awareness and build digital skills by various stakeholders (e. g. chambers of industry and commerce, associations or vocational institutions). A value-based approach to improving data skills across the board is, however, required, for example in the form of **initial and continuing training courses**. A further objective of these courses must always be to raise awareness of the risks posed to individuals and society from the viewpoint of data protection law and ethics.

Government bodies have been slow to recognise the import and implications of the huge volumes of data they have already generated (for statistical purposes, for example), and the advantages and risks entailed by models in which they share data with businesses (government-to-business (G2B) data sharing) or in which the businesses share operating data with them (business-to-government (B2G) data sharing). The current reticence on the part of public authorities to utilise these opportunities means that a large-scale shift in mindset is required, modelled on forerunners in the field of e-governance such as the Scandinavian countries or Estonia. The Data Ethics Commission also recommends that the Federal Government should support work in this area by the relevant research institutions.

5.2.2 Building the infrastructures needed for a data-based economy

Although Germany continues to occupy a leading position in the field of science and technology research, the tech companies providing vital data and analysis infrastructures for the new digital economy primarily hail from the USA (and increasingly from China). This means that a great deal of European data – consumer data, enterprise data and research data – is stored outside Europe and analysed in third countries using software belonging to non-European companies. This makes it crucially important for Germany to develop a data-based economy using **home-grown infrastructures**.

The Data Ethics Commission recommends that the Federal Government should support the following **measures at European level**, which have been initiated by the European Commission:

a) establishment and expansion of the Support Centre for data sharing;

b) development of model contracts for the data economy;

c) support for forums and consortiums tasked with developing open standards for legally compliant data exchanges, in particular formats and programming interfaces (APIs) that are tailored to data exchanges and that increase the traceability of data flows;

d) promotion of European platforms for legally compliant data exchanges; and

e) establishment of a European Open Science Cloud (EOSC).

Key precursors to the achievement of digital sovereignty by Germany include **access control** for sensitive data and the option to carry out appropriate **audits** on critical data analysis software, which would require manufacturers to disclose their source code and design criteria, for example. Given the ethically problematic nature of these analyses, they should, wherever possible, be carried out within the geographical purview of the German legal system.

The Data Ethics Commission **expressly welcomes a number of initiatives** by the Federal Government and other stakeholders aimed at creating secure international data spaces (spearheaded by Germany) for different application domains, allowing companies and organisations of all sizes and from all sectors of industry to retain sovereignty over their data and exchange data securely with each other.

The Data Ethics Commission also recommends the setting up of an **ombudsman's office** at federal level to provide assistance and support in relation to the negotiation of problematic data access agreements and dispute settlement. The competent data protection authorities should be consulted on cases involving personal data, and decision-making power must ultimately rest with the aforementioned authorities in order to avoid conflicting decisions.

Establishment of data infrastructures

The **Federal Government's initiatives** aimed at establishing data infrastructures include the following:

f) Efforts by the German Research Foundation *(Deutsche Forschungsgemeinschaft)* to establish a national research data infrastructure, the aim of which is to implement a science-driven process that systematically opens up data repositories and provides long-term data storage, backup and accessibility across the boundaries of different disciplines and *Länder*.

g) The open International Data Spaces Consortium (IDS, formerly Industrial Data Space) promoted by the Federal Ministry of Education and Research, the aim of which is to provide the companies and organisations taking part with a standardised interface to a platform for exchanging data, based on a federal architecture concept.

h) An initiative to develop a comprehensive network of big data and AI centres, with nodes distributed throughout Germany, as part of a national and generally accessible ecosystem. The aim is for this network not only to provide access to a large amount and variety of data on a 24/7 basis, but at the same time for it to offer easy-to-use tools along the entire data value creation chain (preparation, analysis, visualisation, exploitation) and develop them further on the basis of user feedback.

In addition to these technical platforms, other interesting developments include platforms developed by the Federal Government in collaboration with associations with a view to promoting coordinated research and development and the standardisation and practical implementation of data-hungry applications in the form of socially and economically innovative future projects, such as Industry 4.0, Smart Service World and Learning Systems.

At European level, the **European Commission** is implementing similar projects (e.g. the future-oriented FIWARE project), and is currently developing a freely available toolbox of open-source software components that can be used to configure innovative Internet services in a short space of time. The Big Data Value Public-Private Partnership (organised by the European Commission and the Big Data Value Association, BDVA) has developed an interoperable data-driven ecosystem at European level as a launchpad for new business models using big data, which has already delivered an impressive number of flagship projects. Lastly, the European Institute of Innovation and Technology (EIT Digital) has fostered the emergence of a Europe-wide technical and economic ecosystem involving 180 companies and research institutions.

5.2.3 Sustainable and strategic economic policy

As far as the data economy is concerned, the biggest challenges facing Europe include the lack of **sustainable** funding that is so often a problem for research projects, and a paucity of **venture capital** (the latter being required to make ideas that have already been developed market-ready and inject capital at the appropriate points so that start-ups can reach a competitive size). One of the reasons why the USA has been so successful in the field of digital products and services is because of the country's many "angels" willing not only to invest billions into high-risk projects, but, in many cases, to forfeit those investments. Another trend worth noting is that innovative companies are often **bought out** by foreign companies or forced by international investors to move their headquarters to other countries outside Europe.

Thinking outside the box of the "European path" explicitly endorsed by the Data Ethics Commission (→ see Part G, below), German start-ups must be given access to a larger **pool of funding** and better **tax incentives** so that Germany can continue to attract the brightest and the best and remain at the cutting edge.

Sectors such as education, public administration and medicine are characterised by a high level of public interest and the existence of mandatory values (as expressed through the legal system and professional ethics). At the same time, there is enormous potential to achieve efficiency gains through digitalisation and AI in these sectors, and global platforms have not yet gained a stranglehold over the market to the same extent as in other areas. The Data Ethics Commission therefore recommends that public funding should be channelled into these three areas in particular, and that it should be used to incentivise the **development of platforms** in Germany that reflect our values and are also internationally scalable.

5.2.4 Improved industrial property protection

Also from the perspective of the data economy, the Data Ethics Commission does not see any benefit in introducing **new exclusive rights** to data (often discussed using the terms "data ownership" or "data producer right", → see section 3.3.2 above). Rights of this kind, which would need to be incorporated into (and aligned with) the existing provisions of data protection law or intellectual property law or the rules on the right of personality, trade secrets, ownership rights to storage media, etc., would do nothing but increase the (already significant) level of complexity and legal uncertainty, without any clear indication that rights of this kind would be necessary or even particularly helpful in making data more marketable.

The Data Ethics Commission does, nevertheless, believe that the calls made by industry and government bodies to afford **limited third-party effects** to contractual agreements (e.g. to restrictions on data utilisation and onward transfer of data by a recipient) are justified. Under the legal situation as it currently stands, third-party effects of this kind are at most afforded in extreme situations (unless protection under intellectual property rights law applies, including the "sui generis" protection of databases). Consideration should be given to extending the scope of recognition of third-party effects, along the lines of the model provided by Article 4(4) of the Trade Secrets Directive (Directive (EU) 2016/943);[27] according to this approach, the acquisition, use or disclosure of data would also be considered unlawful whenever a person, at the time of the acquisition, use or disclosure, knew or ought, under the circumstances, to have known that the data had been obtained directly or indirectly from another person who was using or disclosing the data unlawfully. This approach would further the interests of the data economy and also fit seamlessly into the existing (primarily contract-focused) model.

27 This solution is endorsed in Preliminary Drafts no. 2 (February 2019) and no. 3 (October 2019) of the ALI-ELI Principles for a Data Economy (see footnote 1 above), for example.

5.2.5 Data partnerships

The Data Ethics Commission believes that cautious development of the current legal framework would also be appropriate in the field of **anti-trust law**. The breakneck pace of developments in the data economy poses fresh challenges for this field of law; in return, the provisions of anti-trust law pose fresh challenges for digital companies. The Data Ethics Commission recommends that the Federal Government should pay particularly close attention to the opportunities and risks entailed by **data partnerships**; consideration should be given to the introduction of a mandatory but confidential procedure for notifying data partnerships to the anti-trust authorities and to the supervisory authorities under data protection law (in the case of personal data). Mention should also be made of the proposals presented by the Commission of Experts on Competition Law 4.0 under the headings of "data exchange" and "data pooling".

5.3 Data access in existing value creation systems

5.3.1 Context

Fair and efficient data access plays a significant role in modern value creation systems. The area of law most suitable for regulating fairly and efficiently the ability of various stakeholders to access data in a commercial context is **contract law**, since this is the branch of the legal system where the autonomy of private entities ("private autonomy") is expressed most explicitly. At the same time, there is a general presumption that freely negotiated agreements – except in cases of market failure – achieve an efficient allocation of resources and thus increase the general level of prosperity.

Unfair and inefficient contractual arrangements may arise, however, as a result of imbalances of power and information asymmetry; for example, certain issues relating to data access are typically underestimated during the negotiation process, with the result that they are skimmed over or omitted entirely. Given the dynamic nature of data-specific interests and the correspondingly dynamic assessment of data rights and data obligations (→ see section 2.1 above), it is often difficult for parties to determine what exactly a data access regime should look like in order for it to remain fair and efficient for the entire term of the contract. In a not insignificant number of cases, it is accordingly found at a later date – after the contract has been put to the test in the real world – that the balance of interests has shifted in unpredictable ways to benefit one or the other party, with a major impact on the equilibrium of rights and obligations that was originally agreed. Since one of the parties typically stands to benefit from the new state of affairs, contracts are often not renegotiated, and so there is no opportunity to regulate data access properly and efficiently.

Particularly in complex value creation systems, there is frequently **no direct contractual relationship** between the party requesting access and the party that effectively controls the data (because there is an interposing link in the distribution chain, for example); in the interests of fairness and efficiency, however, data access arrangements would be desirable. In the B2B sector, incursions into freedom of contract in the form of an obligation to contract currently result almost exclusively from the provisions of anti-trust law, as well as a small number of general provisions of law in the case of essential commodities and monopoly positions; generally speaking, however, they are restricted to a small number of extreme situations.

5.3.2 Presence of a contractual relationship

In the estimation of the Data Ethics Commission, the steps that should initially be taken with a view to ensuring fair and efficient data access arrangements include **raising awareness and promoting digital skills** (→ see section 5.2.1 above), practical support in the form of **model contracts** that provide for a fair distribution of data access, and **infrastructures and intermediaries** that facilitate shared data use while ensuring protection of trade secrets, for example (→ see section 5.2.2 above).

In cases where a contractual relationship already exists, the principles of fair data access can be enforced primarily through **contract interpretation** (including by way of gap-filling by the courts), for example by assuming the existence of appropriate contractual ancillary obligations, and through **standard contract terms control** pursuant to Section 307 of the Civil Code ("fairness test"). However, one of the problems inherent to substantive fairness tests is the virtual absence of default provisions that can be used as a benchmark for these tests. Specific abusive contractual practices could therefore be spelt out explicitly as **blacklisted contract terms** (→ see section 3.2.3 above for corresponding recommendations for B2C contracts). If significant changes occur since the conclusion of the contract it may be possible for a party to invoke the provisions on **frustration of contract** (Section 313 of the Civil Code).

In this connection, the Data Ethics Commission wishes to reiterate the **general basic principles governing business-to-government (B2G) data sharing** as formulated by the **European Commission** in its communication of April 2018 entitled "Towards a common European data space".[28] These basic principles provide for the following:

a) transparency regarding access rights and the purposes for using the data;

b) recognition that several parties have contributed to shared value creation;

c) respect for each other's commercial interests;

d) undistorted competition; and

e) minimised data lock-in.

Particularly with regard to repositories that potentially include personal data as well as non-personal data, consideration could also be given to expanding these principles to include a right to informational self-determination for data subjects and the principle of "do no harm".

28 European Commission: Towards a common European data space, COM(2018) 232 final, 25 April 2018, p. 10 (available at: *https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-232-F1-EN-MAIN-PART-1.PDF*).

5.3.3 Absence of a contractual relationship

Where there is no contractual relationship at all between participants in a value creation system, despite any support provided, neither the rules on the interpretation of contracts nor the substantive fairness tests for standard contract terms apply, and it is also impossible to rely on frustration. In the view of the Data Ethics Commission, however, the simple fact that the party requesting access has contributed to generation of the data means that a special legal relationship exists between this party and the party that effectively controls the data (→ see section 2.1 above); this is all the more true if the relationship exists within a value creation system that is primarily shaped by contracts. This special legal relationship may give rise to certain duties of a fiduciary nature, including a **duty to enter into negotiations about fair and efficient data access arrangements**. The future legal framework should make explicit reference to this fact.

The Data Ethics Commission therefore recommends **amending Section 311 of the Civil Code** to include a new subparagraph mentioning the special relationship that exists between participants in a value creation system (e.g. as suppliers, manufacturers, brokers or end users), which would entail certain relevant duties, including with regard to data. The enormous significance of data for general legal and economic relations means that there are justified grounds for inserting a subparagraph in the law rather than subsuming such relations under the general heading of "similar business contacts". This would neither constitute a separate legal basis for the processing of personal data, nor would it restrict data protection law in any way.

Beyond this, consideration could be given to introducing **data-specific rules in the law of obligations** based on the principles referred to above (→ in section 2), aimed at judicial "gap-filling" and for use as a benchmark when carrying out substantive fairness tests on standard contract terms.[29] In particular, provisions for data contracts of this kind might define the conditions under which parties are entitled to access data and/or to request desistance from data access or data use and/or to request the rectification of data. However, the Data Ethics Commission was also concerned that, if such rules were specifically spelt out in the law (albeit as default rules only) this might give rise to additional disputes.

5.3.4 Sector-specific data access rights

If a need is identified for more extensive data access rights within existing value creation systems, priority should be given to sector-specific solutions. The Data Ethics Commission therefore recommends that the Federal Government should pay greater attention to data access issues when adopting and/or revising sector-specific regulations.

29 For a discussion of personal data, see Louisa Specht: Datenrechte – Eine Rechts- und Sozialwissenschaftliche Analyse im Vergleich Deutschland – USA, Teil 1: Rechtsvergleichende Analyse des zivilrechtlichen Umgangs mit Daten in den Rechtsordnungen Deutschlands und der USA, ABIDA-Gutachten [Data rights, an analysis from the perspective of the legal and social sciences based on a comparison between Germany and the USA, Part 1: Comparative law analysis of data governance under civil law within the framework of the German and US legal systems], 2017, pp. 89 et seqq. (available at: *http://www.abida.de/sites/default/files/ABIDA_Gutachten_Datenrechte.pdf*); for a discussion of non-personal data, see ALI-ELI Principles for a Data Economy (above, footnote 1).

5.4 Open data in the public sector

5.4.1 Preliminary considerations

The recently revised Directive (EU) 2019/1024 on open data and the re-use of public sector information (PSI Directive) and (at national level) the [German] Information Reuse Act (*Informationsweiterverwendungsgesetz*, IWG), the [German] E-Government Act (*E-Government-Gesetz*, EGovG) and additional special acts provide a sound legislative basis for the disclosure of public-sector data on the basis of OGD concepts. The premise underlying the concept of open government data is that citizens and companies have already subsidised the generation of the data **through the taxes they pay**, and should therefore be allowed to share in the associated benefits rather than incurring a double financial burden. Making public-sector data available for reuse by the private sector also benefits the European data economy; since open government data often hold **enormous potential for private-sector value creation**, companies can use them to develop new and innovative products and services, helping to increase the general level of prosperity in the process.

Looking beyond the economy, access to government data is also vitally important for **democracy and an open debate involving all of society**, since it increases administrative transparency, facilitates participation and allows oversight and fact-based discussions. Open government data can also be used in many different shapes and forms for social initiatives and innovations (for social or ecological purposes, for example).

As a basic principle, therefore, the Data Ethics Commission supports the **Open Data Charter** adopted at the G8 Summit in 2013, which defines the following as central principles for the governance of administrative data:

a) open by default (the expectation that administrative data will be made public without compromising the right to privacy);

b) quality and quantity (high-quality, timely and fully described open data);

c) usable by anyone (as much data as possible, in as many open formats as possible);

d) improved governance through open data (transparency and sharing of expertise regarding data collection, standards and publication procedures);

e) innovation (user consultations and support for future generations of creative minds).

Ethically speaking, if a government body decided to provide commercial operators with free-of-charge access to its data instead of selling them for a profit or otherwise exploiting them for economic purposes, this decision would need to be justified (on an approximated basis) by corresponding **increases in prosperity** at the macrosocial level.

The Data Ethics Commission also wishes to draw attention to a degree of tension between calls for privacy by default on the one hand and open by default on the other, and – in a broader sense – between the **debate on data protection** and the **debate on open government data**. If personal data are made public in a legally compliant manner on the basis of open-data concepts, there is no guarantee that the security mechanisms put in place to protect the right to informational self-determination (in the form of explicit or implicit restrictions on reuse or in the form of technical and organisational protection measures) will continue to be applied. The general provisions of data protection law concerning reuse can also be an issue. Furthermore, since Article 30 GDPR requires only the "categories of recipients" to be documented, and government bodies are almost never in a position to monitor compliance with the "appropriate safeguards" required pursuant to Article 89 GDPR, the disclosure of data that are, or might at some point become, personal data can be regarded as a potentially high-risk measure for data subjects.

When applying OGD concepts in this area, the right to informational self-determination that is protected as a fundamental right must always be weighed up carefully against the public-good interests pursued under the OGD banner, the right to freedom of information (which is also protected as a fundamental right), and the freedom of OGD recipients to exercise a trade or profession. The Data Ethics Commission submits that, in cases of doubt, priority should be given to the **State's duty of protection**. Compliance with this duty is all the more important because individuals may not be able to decide freely which data they entrust to government bodies, or may be **particularly apt to believe** that government bodies will refrain from forwarding personal data to third parties.

5.4.2 Legal framework and infrastructures

The Data Ethics Commission welcomes the Federal Government's National Action Plan to implement the G8 Open Data Charter and the efforts on the part of the Federal Government and the governments of the *Länder* to include OGD concepts when digitalising their administrations. It recommends that the Federal Government should take action to ensure across-the-board implementation of an **obligation to publish structured unprocessed data (open by default)** and to allow these data to be used without limitations and, in principle, free of charge, which already applies to the direct federal administration (Section 12a(1) E-Government Act). Given the aforementioned tension between open government data and data protection, the obligations imposed by Section 12a of the E-Government Act should only apply in relation to certain types of data (in particular those that have undergone effective anonymisation procedures).

The Data Ethics Commission welcomes the legislator's attempts to change the data governance culture within the administration, and acknowledges that this is a task made significantly more challenging by the **highly fragmented nature of the current legal situation**. It is often difficult – both for authorities and for potential OGD users – to forge a path through a tangled regulatory thicket made up of different legal regimes that set out general and specialised rules on access to data, reuse of data and e-governance at both Federal Government and *Land* level. A further complicating factor is the interplay between these regulations, data protection law and intellectual property rights (in particular copyright law), which is often fiendishly complex in practice. In this connection, the Data Ethics Commission recommends **merging** and **synchronising** the various legal bases that exist in Germany, as well as **clarifying** the demarcation lines between the various legal arrangements.

Another obstacle that stands in the way of the culture change that needs to take place is the fact that it is currently all but impossible to verify reliably whether the authorities are, in fact, complying with the data provision obligations already in force. For example, Section 12a(1) of the E-Government Act imposes an obligation on the direct federal administrative authorities to provide public access to data, but explicitly states that parties requesting access have **no enforceable right** to the data being made publicly available. Companies that wish to access data are therefore deprived of effective avenues for forcing the authorities to comply with the statutory obligation of making data open by default. In the view of the Data Ethics Commission, the introduction of a **right to request publication of data** might encourage a more proactive approach to the provision of open data on the part of the administrative authorities, within the limits placed on their obligation to do so by the E-Government Act and the Information Reuse Act.

The **quality standards** that must be achieved in respect of the data provided by government bodies are another question that is left open under the current legal situation. In particular, the E-Government Act states that the obligation to provide access to data can be met by handing over unprocessed data, but data can be reused easily and in a manner that complies with the OGD objectives only if a high level of data quality is guaranteed.

Aside from the legal framework, establishment or expansion of an **infrastructure framework** (e. g. open government data portals such as GovData) is also essential, particularly at local level (e. g. in the form of municipal platforms), and the same applies to investments in appropriate quality assurance tools.

5.4.3 The State's duty of protection

Keeping in mind the State's duty to protect all of the data entrusted to it, **appropriate precautions** must be taken to ensure that central interests of private individuals (e. g. those relating to personal data, operating and trade secrets or other sensitive data, such as confidential information relating to public procurement procedures) are given the same comprehensive level of protection as key public interests (e. g. security interests or interests relating to national sovereignty). The ethical premise underpinning the OGD concept – that citizens and companies have already paid for the data through their tax contributions – places certain **constraints on reuse**. In particular, care must be taken to ensure that data are not used by the private sector to develop services and products that may ultimately restrict the freedom of citizens and businesses and/or be available only to them under unfair conditions.

The Data Ethics Commission therefore recommends that the Federal Government should make use of the opportunity afforded by Article 8 of the recast PSI Directive by developing **model conditions** for standard licences, including restricted-use agreements and conditions for the transfer of data to third parties; alternatively, it should lobby for such conditions to be introduced at European level. It may even be advisable to make these model conditions **mandatory**, at least on a sector-specific basis, and they should be based on a number of key considerations, including the following:

a) pursuant to Article 8(1) of the PSI Directive, the conditions must be objective, proportionate, non-discriminatory and justified on grounds of a public interest objective; they shall not unnecessarily restrict possibilities for re-use and shall not be used to restrict competition;

b) the rules imposed on companies should contain clearly defined safeguards for the rights of affected third-parties, and mechanisms that allow compliance with these rules to be verified;

c) any intellectual property developed using the data must not be used to disallow activities carried out by government bodies in the fulfilment of their public remit or to make these activities subject to payment of a licence fee;

d) any product or service developed using the data should be offered to government bodies under preferential conditions;

e) companies with a large market share should be subject to a reciprocal obligation to make data generated by their operations available (under identical conditions);

f) the data should be used only for business activities that take place in the EU (or at a minimum for product or service development processes that take place in the EU).

As a basic principle, compliance with the agreed safeguards and restrictions on use can no longer be reliably verified once data have been transferred and once the copies of the data sent to the recipient have been stored on infrastructure controlled by the latter. Given the duty incumbent upon government bodies to protect data that may be used to harm third parties or the public (even if such harm would be possible only after de-anonymisation of the data or linking of the data with other data sets), special consideration must be given to a model under which government bodies allow only supervised data access and **supervised processing** of data on infrastructures that they control. Any costs incurred in this connection should be passed on to the companies seeking access.

5.5 Open data in the private sector

5.5.1 Platforms and data use

Operating data are generated by companies at all levels of the German economy in the course of their everyday business, and these data are enormously valuable for innovation, particularly when combined with data generated by other participants in the value creation chain. The German economy has already established sector-specific platforms for the express purpose of linking these different types of data.

Examples of different platform models include: (1) merger of several different companies into a GmbH (limited liability company); (2) in-house operation by a single company with the involvement of partners; (3) proprietary platform operated as a service for third parties.

The various sectors of the economy are increasingly coming around to the idea not only of shared platforms, but also of common regulatory approaches to data use.

The Data Ethics Commission believes that it is reasonable to assume that data use within value creation systems will continue to be organised by industrial players themselves on a sector-specific basis, and that new market entrants and start-ups will continue to find opportunities to innovate within this landscape, since market participants themselves stand to benefit from working together with trailblazing start-ups to develop disruptive digital innovations, and from sharing their data to this end. The trend for companies to club together to establish platforms modelled along various lines should be welcomed, as it allows them to build on the industrial know-how that already exists in Europe and fosters higher-quality data use (including higher standards of data protection and information security). The Data Ethics Commission proposes that the Federal Government should lend its support to the emergence of an increasing number of **private-sector platforms**, with a view to achieving the necessary market size and effects of scale and allowing German businesses to harness their shared strength to compete on the international stage.

5.5.2 Additional incentives for voluntary data sharing

There is already a large number of business models based on private providers voluntarily allowing the public to access their data.

Example 14

Example 14 One case in point is the geoinformation industries, which take basic geodata (in some cases from official sources) and enrich them with other information, allowing users to access specialist geodata for a wide range of purposes. Examples include both map services such as OpenStreetMap or Google Maps, which feature not only purely topographical and administrative information but also a wide range of other interesting details, and also tailored offerings such as weather forecasts or traffic predictions.

The Data Ethics Commission recommends that voluntary data access arrangements of this kind should be supported; in addition to the **practical support measures** recommended in → section 5.2 above, consideration should therefore be given to **additional incentives** for voluntary data sharing. For example, data transfers or releases and open access strategies should be favourably viewed:

- under tax legislation;
- under procurement law;
- when making grant awards (either inside or outside the research sector); or
- when carrying out authorisation procedures.

Voluntary data sharing, data transfers or releases and open access strategies should, however, be envisaged in the fields referred to above only if there is no risk of infringing confidentiality requirements under procurement law, operating and trade secrets, or the provisions of data protection law as a result.

5.5.3 Statutory data access rights

By way of contrast to the debate on voluntary data sharing, the main idea underpinning the discussion on statutory data access rights is that a society should "get something back" from the large repositories of data that many members of that society have helped to build up (in the case of social networks, for example). When viewed in conjunction with the fundamental value of social solidarity and the public-good interests that may be relevant in specific cases, this concept could serve as a basis for granting more extensive rights in respect of **data access and disclosure obligation**s on the part of private individuals.[30]

One potential measure that is often discussed in the context of improving general access to privately held data repositories is the introduction of a **general right to portability** for non-personal data, modelled along the lines of Article 20 GDPR. This would mean that a business that has supplied raw data to a controller would have a right to request the controller to make the data available to the business in a commonly used and machine-readable format, or to ask the controller to forward them directly to a third party. For reasons that are essentially similar to those cited in its arguments against an extension to the scope of Article 20 GDPR (→ section 4.4.2 above), the Data Ethics Commission recommends that the Federal Government should initially adopt a **"wait-and-see" approach** to developments relating to the **use and interpretation of Article 20 GDPR**. The complexity of this issue is exacerbated yet further by the fact that the issue of proper allocation of the portability right (i.e. who is the equivalent to the "data subject" with regard to non-personal data) would raise its head again.

A range of other measures that are ultimately synonymous with statutory data access rights are also being discussed with a view to improving general access to privately held data repositories. **Potential options** in this respect include a statutory obligation to publish reports containing internal data analytics, access rights for private individuals (e.g. mandatory licensing that complies with the FRAND[31] principles and/or incorporates the three-factor or four-factor test under copyright law[32]), or the disclosure of data to the general public (open access) based on either a general model or a market-share model.

The Data Ethics Commission believes that at least the following factors should be taken into account during an initial examination of these options:

30 For further details, see Viktor Mayer-Schönberger / Thomas Ramge: Das Digital [english title: Reinventing Capitalism in the Age of Big Data], pp. 195 et seqq.
31 FRAND = Fair, Reasonable and Non-Discriminatory.
32 The "three-factor test" features in several international agreements as a basis for assessing whether an exemption (i.e. a limitation on copyright) represents an acceptable encroachment on the copyright holder's rights. According to the test, exemptions of this kind are subject to three conditions: (i) they may apply only to certain special cases; (ii) they may not be in conflict with normal exploitation; and (iii) they may not unreasonably prejudice the legitimate interests of the right holder. Calls are increasingly being made for the test to include (iv) mandatory consideration of third-party interests and general interests.

a) the need to protect the personal data or the operating and trade secrets to which access may be given or which may be disclosed;

b) the need to ensure that any encroachment on the fundamental rights of private entities affected by a data access or disclosure obligation is proportionate; this relates in particular to the freedom to exercise a trade or profession;

c) the need to avoid any negative impacts on competition resulting from access to data or the disclosure of data, for example owing to strategic use by competitors that may not themselves be obliged to disclose data in return;

d) the need to ensure that incentives still exist to invest in business models for the data economy; and

e) the need to protect the strategic interests of German or European companies in the face of global competition; in particular, consideration must be given to whether these companies would still be able to compete effectively on the international stage if they were forced to provide access to their data repositories and the digital giants – which already stand head and shoulders over other companies in terms of their data proficiency, their data infrastructures and (in particular) the volumes of data they hold – were to exploit this open-door policy.

Having regard to the above, the Data Ethics Commission recommends that preference should be given to a **sector-specific approach**. As far as spatial information is concerned, the INSPIRE Directive and the provisions transposing it into national law already set out sector-specific data access rules; these rules apply only to government bodies, however. One of the first private-enterprise applications of an sector-specific data access right can be found in the payment services industry, and the Data Ethics Commission proposes that steps should be taken to identify the level of demand and implementation options in a number of other selected industries, for example the **media, mobility or energy sectors**.

5.5.4 Role of competition law

Although the framework of competition law that is currently in place contains almost no provisions relating to data, its general thrust also applies to the data economy. For example, the **essential facilities doctrine** (EFD) can be used (in a slightly modified form if necessary) if a market-dominant company holds exclusive control over a resource (e.g. a network/infrastructure) that is crucially important for competition on a neighbouring market. The **aftermarket doctrine** relates to cases in which lock-in effects mean that consumers of a primary product are unable to exercise in full their freedom to choose on a secondary market (e.g. market for repairs/spare parts), or in which a third-party provider on a secondary market of this kind faces anti-competitive barriers.[33] Yet the uncertain legal situation, the stringent requirements that apply, and the amount of time and money involved in the relevant procedures means that supervisory efforts to prevent abuse cannot currently be regarded as a fix-all solution to data access problems. The applicable provisions of competition law (either individually or in their entirety) could, however, act as a central building block in a new framework of **digital economic law**, one of the crucial components in which should be a range of solutions to data access problems. The findings of the Commission of Experts on Competition Law 4.0 should be taken into account in this respect.[34]

33 Jacques Crémer / Yves-Alexandre de Montjoye / Heike Schweitzer: Competition policy for the digital era, Special Advisers' Report for the European Commission, pp. 87 et seqq. (available at: *https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf*).

34 A New Competition Framework for the Digital Economy, Report by the Commission 'Competition Law 4.0', September 2019 (available at: *https://www.wettbewerbsrecht-40.de/KW40/Redaktion/DE/Downloads/a-new-competition-framework-for-the-digital-economy_.pdf?__blob=publicationFile&v=3*).

5.6 Data access for public-sector (B2G) and public-interest purposes

Thought should be given to whether controllers should be subject to an obligation to grant access to specific subsets of data in order to allow their use either by **public-sector bodies** or for certain **public-good purposes**, and the scope of any such obligation. Rights to access data belonging to private entities or obligations to disclose data might be particularly relevant in the **research** sector, and easier access to data might lead to general advances in science, provided that the access arrangements are designed appropriately and take due account of data subjects' rights. Corresponding access rights to private-sector data might also make it easier for NGOs, the media and similar institutions to deliver on their social remit, thereby helping to protect the **democratic polity**. Particularly priority must also be given at all times to the **averting of risks** (e. g. issuing storm warnings).

In the view of the Data Ethics Commission, preference should again be given to a **sector-specific approach** that tailors the design of data access and disclosure obligations to the specific requirements of constitutional law that come into play on the one hand, and to the practical circumstances that characterise the relevant sphere of activity on the other. The **health sector**, the **mobility sector** and the **energy sector** should be regarded as particular priorities for action in this respect. The Data Ethics Commission also calls for a broad-based, society-wide debate as a precursor to decisions on more general obligations to provide access to data, e. g. in connection with research projects that serve the public good.

The Data Ethics Commission wishes to reiterate the **basic principles governing business-to-government (B2G) data sharing** set out by the European Commission in its communication of 25 April 2018 entitled "Towards a common European data space":[35]

a) proportionality (i. e. justified by clear and demonstrable public interest and proportionate in terms of details, relevance and data protection);

b) purpose limitation (i. e. clearly limited for one or several purposes and assurances that the data obtained will not be used for unrelated administrative or judicial procedures);

c) "do no harm" (i. e. respect for legitimate interests such as data subjects' right to informational self-determination, trade secrets, commercially sensitive information and exploitation interests);

d) acknowledgement of the public interest goal when agreeing on conditions for data reuse (preferential treatment for government bodies, non-discriminatory conditions for government bodies, reduction in the overall burden on citizens and companies);

e) data quality management (an obligation to offer reasonable and proportionate support to help assess the quality of the data for the stated purposes, but no general obligation to improve the quality of the data);

f) transparency and societal participation in respect of parties to the agreement, their objectives, insights and best practices.

These basic principles may serve as a **good starting point** not only when drafting the provisions of freely negotiated contracts on data exchanges, but also when designing more extensive sector-specific statutory measures to improve data access.

35 European Commission: Towards a common European data space, COM(2018) 232 final, 25 April 2018, pp. 13 et seq. (available at: *https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-232-F1-EN-MAIN-PART-1.PDF*).

Summary of the most important recommendations for action

Debates around access to non-personal data

24

Access by European companies to appropriate non-personal data of appropriate quality is a key factor for the growth of the European data economy. In order to benefit from enhanced **access to data**, however, stakeholders must have a sufficient degree of data-awareness and have the data skills that are necessary to make use of the data. Also, access to data proves to be disproportionately advantageous to stakeholders that have already built up the largest reserves of data and that have the best data infrastructures at hand. The Data Ethics Commission therefore wishes to stress that the factors referred to should always receive due attention when discussing whether and how to improve data access, in keeping with the **ASISA principle** *(Awareness – Skills – Infrastructures – Stocks – Access)*.

25

The Data Ethics Commission therefore supports the efforts already initiated at European level to promote and improve **data infrastructures** in the broadest sense of the term (e.g. platforms, standards for application programming interfaces and other elements, model contracts, EU Support Centre), and recommends to the Federal Government that these efforts should continue to be matched by corresponding efforts at national level. It would also be advisable to set up an ombudsman's office at federal level to provide assistance and support in relation to the negotiation of data access agreements and dispute settlement.

26

The Data Ethics Commission ascribes enormous importance to a holistically conceived, sustainable and strategic **economic policy** that outlines effective methods of preventing not only the exodus of innovative European companies or their acquisition by third-country companies, but also an excessive dependence on third-country infrastructures (e.g. server capacities). A balance must be struck in this context between much-needed international cooperation and networking on the one hand, and on the other a resolute assumption of responsibility for sustainable security and prosperity in Europe against the backdrop of an ever-evolving global power dynamic.

27

Also from the perspective of boosting the European data economy, the Data Ethics Commission does not see any benefit in introducing new exclusive rights ("data ownership", "data producer right"). Instead, it recommends affording **limited third-party effects to contractual agreements** (e.g. to restrictions on data utilisation and onward transfer of data by a recipient). These third-party effects could be modelled on the new European regime for the protection of trade secrets. The Data Ethics Commission also recommends the adoption of legislative solutions enabling European companies to cooperate in their use of data, for example by using data trust schemes, without running afoul of anti-trust law (**"data partnerships"**).

28

The data accumulated in existing value creation systems (e.g. production and distribution chains) are often of enormous commercial significance, both inside and outside that value creation system. In many cases, however, the provisions on data access that appear in the contractual agreements concluded within a value creation system are unfair and/or inefficient, or lacking entirely; in certain cases, there is no contractual agreement at all. Efforts must therefore be made to **raise awareness among businesses** in sectors far outside what is commonly perceived as the "data economy", and to provide practical guidance and support (e.g. model contracts).

29

The Data Ethics Commission furthermore recommends cautious **adaptations of the current legislative framework**. The first stage in this process should be to make explicit reference in Section 311 of the [German] Civil Code (*Bürgerliches Gesetzbuch*, BGB) to the special relationship that exists between a party that has contributed to the generation of data in a value creation system and the controller of the data, clarifying that such parties may have certain quasi-contractual duties of a fiduciary nature. These duties should normally include a duty to enter into negotiations about fair and efficient data access arrangements. Consideration should also be given to whether additional steps should be taken, which could range from blacklisting particular contract terms also for B2B transactions, to formulating default provisions for data contracts, to introducing sector-specific data access rights.

30

The Data Ethics Commission believes that **open government data (OGD) concepts** hold enormous potential, and recommends that these concepts should be built on and promoted. It also recommends a series of measures to promote a **shift in mindset among public authorities** (something that has not yet fully taken place) and to make it easier in practice to share data on the basis of OGD concepts. These measures include not only the establishment of the relevant **infrastructures** (e.g. platforms), but also harmonisation and improvement of the existing **legal framework** that is currently fragmented and sometimes inconsistent.

31

Nevertheless, the Data Ethics Commission identifies a degree of tension between efforts to promote OGD (relying on principles such as "open by default" and "open for all purposes"), and efforts to enhance data protection and the protection of trade secrets (with legally enshrined concepts such as "privacy by default"). The Data Ethics Commission submits that, in cases of doubt, **priority should be given to the duty of protecting** individuals and companies who have entrusted their data to the State (often without being given any choice in the matter, e.g. tax information). The State must deliver on this duty by implementing a range of different measures, which may include technical as well as legal safeguards against misuse of data.

32

In particular, it would be beneficial to develop **standard licences and model terms and conditions** for public-sector data sharing arrangements, and to make their use mandatory (at least on a sector-specific basis). These standard licenses and model terms and conditions should include clearly defined safeguards for the rights of third parties who are affected by a data access arrangement. Provision should also be made against data being used in a way that ultimately harms public interests, and also against still greater accumulation of data and market power on the part of the big players (which would be likely to undermine competition) and against the taxpayer having to pay twice.

33

As regards **open-data concepts in the private sector**, priority should be given to **promoting and supporting voluntary data-sharing arrangements**. Consideration must be given not only to the improvement of infrastructures (e.g. data platforms), but also to a broad range of potential incentives; these might include certain privileges in the context of tax breaks, public procurement, funding programmes or licensing procedures. Statutory data access rights and corresponding obligations to grant access should be considered as fall-back options if the above measures fail to deliver the desired outcomes.

34

Generally speaking, the Data Ethics Commission believes that a cautious approach should be taken to the introduction of statutory data access rights; ideally such rights should be developed only on a **sector-by-sector basis**. Sectors in which the level of demand should be analysed include the media, mobility or energy sectors. In any case, before a statutory data access right or even a disclosure obligation is introduced, a full impact assessment needs to be carried out, examining and weighing up against each other all possible implications; these include implications for data protection and the protection of trade secrets, for investment decisions and the distribution of market power, as well as for the strategic interests of German and European companies compared to those of companies in third countries.

35

The Data Ethics Commission recommends considering enhanced obligations of private enterprises to grant access to data **for public interest and public-sector purposes** (business-to-government, B2G). A cautious and sector-specific approach is, however, recommended in this respect as well.

Part F

Algorithmic systems

1. Characteristics of algorithmic systems

Numerous products and applications these days, from voice assistants and automated lending right through to "autonomous" (driverless) cars, are based on more or less "smart" algorithms. Due to the many different forms that these types of technical systems can take, it seemed advisable to the Data Ethics Commission to base the considerations on the **general concept of "algorithmic systems"**. (→ see Part C, section 2.2.5 above). The key questions presented by the Federal Government regarding the topics of "algorithmic prognosis and decision-making processes" as well as "artificial intelligence" will therefore be discussed below together as questions concerning the use of algorithmic systems.

However, the **following distinctions** in particular must be taken into account as part of any ethical and legal **assessment of individual algorithmic systems**:

- From a **technical perspective**, different algorithmic systems have different characteristics. The spectrum ranges from systems which operate on a completely deterministic basis right through to systems which use machine learning to develop action plans independently in order to achieve the goal specified by the operator of the algorithmic system.

- Where algorithmic systems are used as social informatics systems, ethically and legally relevant processes can be established at **different system levels**, i. e. from the level of the pool of data used or the algorithm in the technical sense right through to the level of human individuals involved in the development, implementation, assessment or correction of the system.

- **The purpose and consequences** of using algorithmic systems can vary considerably. Where algorithmic systems support or replace human decision-making and prognoses, they often have a direct impact on individuals' rights and interests. Examples include automated lending and automated administrative acts. However, algorithmic systems are also used where such a link to human decision-making can, at most, be indirectly established. This is the case, for example, with various processes which constitute "autonomous" driving or with predictive maintenance in mechanical engineering.

- Algorithmic systems affect different **ethical and legal principles** depending on the context in which they are used. As such, the externally visible and discernible "action" of "autonomous" cyber-physical systems, for example, typically raises questions. This is a key aspect, for example, in the debate surrounding the use of robotics in healthcare. Principles such as that of human-centred design are essential for the assessment of such systems. Where algorithmic systems are not "physically embodied" in a similar way, it is conversely often the system's externally invisible method for making the "decision" that is the focus of attention. Discussions may, for example, centre on the system's transparency or the principle that the final decision should be made by a human in accordance with Article 22 GDPR. An example of this is automated credit checks. However, the distinction between "action"-oriented and "decision"-oriented perspectives becomes relative upon closer inspection, because every visible "action" by a system is, at some point, preceded by a human "decision", for example in the construction of the system, and every "decision" has an impact because another system component (including a human) will base its "action" on it.

The Data Ethics Commission believes that **further distinctions** should be made in particular where algorithmic systems are closely involved in human **decision-making processes**. An algorithm itself cannot make a decision in an ethically substantial sense, since it has no value-based preferences of its own accord. Three different levels of the involvement of algorithmic systems in human decision-making can be distinguished based on the specific distribution of tasks between humans and machine:

- **Algorithm-*based*** decisions are human decisions based either in whole or in part on information obtained using algorithmic calculations. Examples include clinical decision support systems which provide a doctor with treatment recommendations using patient data from electronic medical records and based on an assessment of scientific literature. Taking this recommendation into consideration, the doctor then makes the decision together with the patient as to which treatment option should ultimately be selected. Algorithm-based decisions can nevertheless subtly yet significantly influence human decisions, for example if the algorithmic system collates information on humans/objects/procedures which contain a value judgment of which the user may not necessarily be aware.

- **Algorithm-*driven*** decisions are human decisions shaped by the outputs of algorithmic systems in such a way that the human's decision-making abilities and capacity for self-determination are effectively restricted, in particular because the decision can be made only within algorithmically determined and prescribed paths. One such example is an Industry 4.0 application whereby, as part of human-machine interaction, a robotic system provides the human involved in the production process with only limited room for manoeuvre.

- **Algorithm-*determined* and hence fully automated** decisions are, prima facie, made independently of a human. In fact, the outputs of an algorithmic system trigger consequences automatically; no provision is made for an explicit human decision. Examples of applications range from price differentiations in e-commerce and fully automated administrative acts up to what are known as autonomous weapons systems. Human decisions are nevertheless involved in the sense that a human must have decided to use the algorithmic system for such a purpose and in such a way.

Example 1

The differences can be illustrated by an algorithmic system being used in the process of selecting candidates for a job: if an algorithmic system simply collates information on the individual candidates for the employer in question on the basis of which the employer will then make their decisions, this constitutes an algorithm-based decision-making process. The system will lead to algorithm-driven decisions if the information provided to the employer contains an evaluation of the individual candidates (for example a ranking), as this could significantly influence the likelihood of the individual candidates being selected. The actual restriction of the employer's ability to make decisions becomes even more apparent if the system already screens the candidates in advance, meaning that the employer no longer even sees some of the applications. In the case of an algorithm-determined selection process, each notification regarding the acceptance or rejection of an application would be automatically provided by the algorithmic system without a human ever checking the selection.

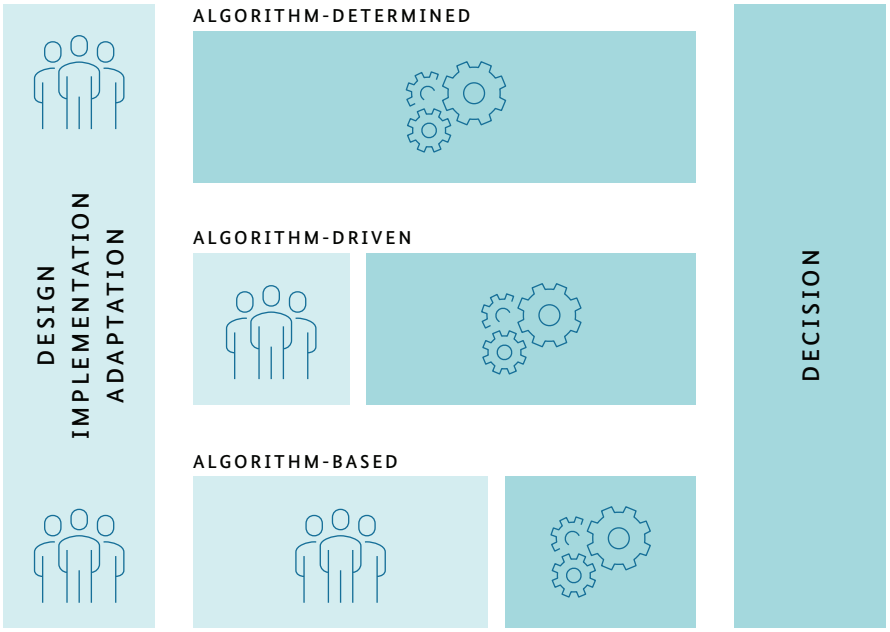


Figure 7:
Characteristics of algorithmic systems

Classifying an algorithmic system as one of these three types is often difficult, and **hybrids** are possible within complex software architecture. The level of determination for humans at the same point can also be different depending on the way the system works: in the example above, a decision-making process in which an algorithmic system filters out individual candidates in advance and rejects them is algorithm-determined from the point of view of the candidates who were filtered out but algorithm-driven for all the remaining candidates.

There can be **overlaps** in the practical operation of the systems on account of what are known as **automation bias and default effects**. Even in the case of algorithm-based decisions where humans have full decision-making authority, they may tend to simply go with the algorithmic system's recommendation without carrying out a sufficiently critical check, as otherwise they would feel an uncomfortable need to justify their decision and would get the impression that the risk of being blamed for any wrong decision would increase. Nevertheless, the fundamental distinction is relevant for assigning responsibility for a risk assessment and therefore also for regulation.

2. General standards for algorithmic systems

General ethical and legal principles, primarily human dignity (→ see Part B, section 3 above), constitute the benchmark for the design and use of algorithmic systems. In terms of the **principle of prospective responsibility**, the intentional and unintentional effects on the users and the individuals affected by the use of an algorithmic system must be taken into consideration as part of the assessment of specific algorithmic systems. It is also necessary to think about and plan for social consequences depending on the intended purpose and context of their use, especially with regard to network effects, effects of scale and effects of scope. These consequences range from the positive effects of social innovations right through to the (sometimes subtle) negative effects, for example on diversity and the culture of social debate as an essential condition for a functioning democracy. On that basis, the Data Ethics Commission believes that the following key requirements for the design and use of algorithmic systems can be set out which, in terms of the **governance perspective** taken up here, must be met in the interplay of, especially, developers, companies, users and state bodies.

2.1 Human-centred design

At the centre is the requirement to strive for algorithmic systems with a **human-centred and value-oriented design** which takes fundamental rights and freedoms into consideration. The Data Ethics Commission believes that the human-centred approach must permeate the entire design process. It must be ensured by means of a wide range of different measures, which may also and in particular involve i**nclusion and participation in the development** of algorithmic systems.

Human-centred design requires in particular taking into account changes in self-perception and self-design resulting from the individual's confrontation with algorithmic systems. Gains and losses in expertise in using the systems, effects on people's own lifestyles and the formation of opinions as well as on physical well-being must be taken into consideration as early as in the system development stage.

Attention should also be paid to the **emotional state** of the affected individuals which may differ (in both directions) depending on whether humans and conventional technology or algorithmic systems are used. This is significant not only for the individual affected by a decision but also for the user. Consideration should be given for example to the fact that direct interpersonal interaction fulfils a variety of functions which go far beyond "good decision-making".

Example 2
Where medical diagnoses are supported by algorithmic systems, the accuracy of diagnosis can be identified first and foremost as the intended purpose. However, the need for human care and contact in consultations concerning treatment (with corresponding significance for the success of the treatment) can be strong and must not be disregarded, nor should the need for doctors to be able to contribute their own medical experience. Conversely, in certain situations, for example in case of embarrassing symptoms, they may find it more comfortable not to have to confide primarily in another human person.

These functions include, for example, the satisfaction of a basic human need for **communication**, the feeling, in principle, of being able to assess the other person's line of thinking and reactions and to be understood by the other person, the opportunity to convince the other person of one's own point of view, as well as the certain control effect arising from the fact that the other human being is directly confronted with the reaction of the individual affected by the decision.

Example 3
Emotional aspects also play a major role where algorithmic systems are used in human-machine interaction. For example, the use of a system which is intrinsically intended to support employees may be perceived by the employees to be invasive or patronising, since the system analyses employees' behaviour, takes certain tasks off their hands which they have actually come to enjoy or makes them think that their own performance is inferior to that of their "robotic colleague".

The well-being of all individuals affected by technology, including for example in the use of robotics in nursing, is a central guiding value which absolutely must be taken into consideration as part of an ethical approach to technology design. It is important to note here that well-being is extremely subjective and not static but can change depending on the context and over time and therefore needs to be **constantly reassessed**.

2.2 Compatibility with core societal values

Depending on their area of application, the impacts of algorithmic systems may be relevant for society as a whole: for example they may affect the **democratic process**, **citizen-centred state action**, **competition**, the **future of work** and also the **digital sovereignty** of Germany and Europe.

Example 4
In the development of smart systems, providers which are able to build their business model on large amounts of data have a privileged starting position, since many applications of algorithmic systems depend on such amounts of data. The more data that can be analysed, the more likely correlations and findings are to be generated. Taken together with the network effects, effects of scale and effects of scope which are typical for platform markets, the market power of companies begins to strengthen and monopolies are formed, once a certain threshold is reached. This ultimately enables companies to prevent new players from entering the market and to interfere with the market-regulating forces of competition. Depending on the area of application, companies can then control social opinion-forming processes and market behaviour. In order to counteract that and create framework conditions for fair competition, the competition law control mechanisms must be readjusted and, where necessary, subsequently tightened.

The Data Ethics Commission is of the view that these supra-individual consequences often cannot be handled by state bodies or with legislative measures alone. Instead, they need to be taken into consideration in all phases of the design and use of algorithmic systems. **To that extent, developers, companies and users have a (shared) social responsibility.** In particular where corresponding consequences seem likely, for example in the case of algorithmic systems which affect communication between people which is relevant to democracy, it is necessary already in the design process to thoroughly assess the purposes and the unintended indirect consequences of the system in question and to examine the extent to which the system can affect democracy, fundamental rights, secondary law and the basic principles of the rule of law. As far as possible, a culture of "incorporating" the basic principles of democracy, the rule of law and fundamental rights into the system architecture should be established for the process of designing technology.

Many aspects of the interplay between technology and society are admittedly still unclear. The Data Ethics Commission believes that more research is therefore necessary to shed light on the social impacts of algorithmic systems and develop corresponding strategies to limit any negative effects.

2.3 Sustainability in the design and use of algorithmic systems

Any assessment of the personal and social effects of algorithmic systems must also be global in nature and not limited with regard to time. For this reason, when deciding on the use and design of algorithmic systems, **sustainability** and **human skills retention** in particular must also be taken into consideration. These are important for remaining human control functions (e.g. the "human-in-the-loop" principle), for the failure of algorithmic systems in exceptional circumstances (e.g. in the event of a disaster or cyber attacks) and for ensuring the innovative prowess of future generations (e.g. development of new digital technologies). It is, first and foremost, a question of basic and advanced training, as well as education in the sense of lifelong learning, ensuring that future generations also have the necessary general skills and not limiting training only to the user's perspective.

Teaching and developing digital skills also promotes **social sustainability**. Social framework conditions, for example in institutions and procedures, must be organised in such a way as to ensure the promotion of the participatory and inclusive design of algorithmic systems and their use to serve the public interest.

Sustainable development also includes the **ecological dimension**. Irrespective of the positive contribution which algorithmic systems can make to environmental protection, a key ethical requirement is reducing the need for electricity and for certain resources such as "rare earths" and using them efficiently.

Economic sustainability requires a perspective which looks beyond exclusively short-term economic profits and also takes the long-term effects into consideration. Short-term commercial success can have long-term disastrous consequences, as demonstrated by the global financial crisis several years ago. This should not limit the freedom of economic activity but should focus attention on the responsibility associated with economic activity within the context of a social market economy.

The principle of prospective responsibility as well as considerations of fairness and solidarity must, with regard to sustainability, be specifically taken into consideration in the design and use of algorithmic systems. As is the case with the handling of data, the **risk assessment** is of crucial importance for ecological, economic and social sustainability in the design and use of algorithmic systems.

2.4 High level of quality and performance

Algorithmic systems must work well and reliably in order to achieve the goals pursued with their help. If the systems are also used to promote ethical aims, then technical and legal specifications, designed to **improve, further develop and safeguard the state of the art**, will take on an ethical quality. Where such systems support or replace human activities, they are deemed, irrespective of the intrinsic value of human activity, to be implementing ethical principles better than previously.

Example 5

Any ethically sound use of algorithmic systems in the healthcare sector firstly requires the technology to have the necessary medical quality, i.e. the accuracy of the assessment of findings, the accuracy of the diagnosis, the probability that the recommended treatment will be successful or the success rate of a medical intervention, etc. must, when the system is used, be at least as good as and (in view of the sensitive usage context) ideally better than if conventional technology and humans were used.

Quality and performance can be improved through a wide range of different measures. These include, for example, appropriate risk models, the, as inclusive and participatory as possible, development of standards, systemic management and control approaches, and process design which is aimed at the continuous improvement of the entire system. The role of humans who are part of an algorithmic system understood as a social-informatic ensemble (→ see section 1 above) must always be taken into consideration in this context. After all, a number of algorithmic systems still rely on input from critical experts to perform optimally. Quality-oriented system design therefore also includes mechanisms which help **enhance human capabilities** and prevent or counteract any reduction in skills and any critical ability and readiness to reflect, for example in connection with automation bias. Examples of productive interaction between humans and machines which is also designed to ensure skill retention can be found in algorithm-supported diagnostic imaging in the healthcare sector.

2.5 Guarantee of robustness and security

Algorithmic systems must be robust and secure, otherwise the legitimate goals they are used to pursue will not be achieved or will be achieved only at the expense of potential harm to ethically and legally protected rights and interests. From an ethical perspective, it can be said that robust and secure system design and appropriate system usage therefore affect the respective purposes of a system and the need to protect the data used by the system. As a result, the robustness and security requirements are not identical for all systems. The specific requirements can differ based on the **specific need for protection and the usage context**.

Example 6

Systems which are not robust or secure which are used in control systems can pose an immediate threat to people or the environment, for example if they control the emission of pollutants from industrial plants, control robots or steer autonomous (driverless) cars in traffic. A failure here could even cause harm to important legally protected rights such as life and limb. In order to prevent this, processes should be put in place to define the current state of the art, legal rules and regulations should be enacted which make it mandatory to follow the state of the art, and measures should be implemented which guarantee the effective enforcement of standards.

Robust and secure system design involves not only **securing the system** against external threats (e. g. by means of encryption or anonymisation, etc.), but also **protecting humans and the environment against any negative influences from the system** (in particular through a systematic risk management approach, e. g. on the basis of a risk assessment). It must also incorporate all phases of data processing and all technical and organisational components. Risks can arise not only in the technical design but also as a result of errors caused by human decisions taken when using algorithmic systems. As algorithmic systems and the way they are incorporated in an organisation's other information technology are not static, a **management system** is also required which checks and ensures the effectiveness of the measures in view of changing conditions, for example newly discovered risks.

2.6 Minimising bias and discrimination as a prerequisite for fair decisions

A key aim in regulating algorithmic systems is to ensure that the decision-making patterns upon which the algorithmic systems are based do not have any systematic distortions (bias) leading to discriminatory and unfair decisions. It should, first of all, be noted that biased, discriminatory and unfair decisions can also be found where conventional technology and humans are used. Conversely to prejudiced decisions of individual humans, algorithmic systems however bear the danger that using the system on a large scale will have a broad impact which individual human decision-makers could never cause. With that in mind, the discussion surrounding bias and discrimination by algorithmic systems should, in the view of the Data Ethics Commission, **also be seen as an opportunity** to detect existing problems in existing decision-making contexts and, in general, achieve better decision-making processes.

Example 7

An algorithmic system used to detect skin cancer was trained predominantly on patients with white skin, and so the probability of its correctly detecting skin cancer is therefore significantly higher in the case of patients with white skin than in the case of patients with different coloured skin. As a medical device, such a system would be permitted for use only on patients with white skin. The same effect would admittedly also be noted if a dermatologist did their training and practised as a clinical professional exclusively in a specific cultural environment. Ultimately, in both cases, steps would need to be taken to ensure that all patients, irrespective of their skin colour, receive proper medical care.

Even in cases where there is no direct intention to discriminate when developing algorithmic systems, discriminatory decisions may still be made, i. e. decisions which systematically put certain groups at an unfair disadvantage. In particular in the case of machine learning, the problem is rather that the systems learn models by using available data. The resulting predictions and recommendations **extrapolate the past into the future**, whereby existing social injustices can be obscured through incorporation into seemingly neutral technology, and potentially amplified.

Example 8

An algorithmic system used to assess applications for a managerial position was trained with data of managers who had proven themselves at the relevant company over the past few decades. Since predominantly male managers had been employed over the past few decades, the system, which was trained with this data set, consistently assesses male candidates as being better than equally qualified female candidates.

The keyword **bias** covers a **range of different types of systematic distortions** with a range of different causes. In the case of human decision-makers, both cognitive bias and social preconceptions, prejudices or stereotypes can negatively affect the decision-making process. In the case of algorithmic systems, bias can refer to the technical reproduction of those social preconceptions, prejudices or stereotypes. This reproduction can take place at various points primarily within the context of machine learning. Often, an insufficient level of representation or a low number of cases of a social group in the training data leads to distortions whereby the specific characteristics of this group are not sufficiently recognised during the development process and are therefore not taken into account. In addition to the training data used, other technical and methodological decisions, e.g. regarding the target variables or labels, can also lead to discriminatory models and therefore to unfair decisions. Lastly, problems may not arise until the systems are actively used in practice for example if algorithmic systems are used in changing social framework conditions or in unforeseen usage contexts.

Algorithmic systems which **directly** use categories of data which are legally explicitly recognized to be **highly sensitive**, such as gender or origin, are particularly critical from the point of view of discrimination. Direct use of sensitive information may, depending on the area of application, be important for correct data processing and is also often permissible within legal limits.

Example 9
Many systems for diagnosing diseases know the patient's gender and age and take them into account. Sensitive characteristics may also be used within the context of a business decision for implementing business strategies, for example where a business is expanding into a specific age group, occupational group or region, if the characteristics define a customer segment, for example, for which simplified acceptance criteria apply.

The use of information which **indirectly** codes sensitive categories can, however, also be problematic.

Example 10
Household income is used as information in creditworthiness assessments. In Germany, the average income varies between genders. As a result, an algorithmic system which uses household income may incorrectly assess the creditworthiness of the men and women involved in terms of the distribution between them.

Fully preventing discrimination even in terms of legally recognised categories such as gender or origin is difficult within the context of algorithmic systems. Furthermore, the use of algorithmic systems can lead to **totally new groups being thrown together based on coindicing characteristics** being excluded from socially protected rights due to a certain classification system and without any just cause, or being confronted with other negative consequences. In the light of this, all those involved in the development and use of such a system must be made aware of the complex conditional discriminatory effects so that they can prevent or counteract them as far as possible (→ see section 4.2.4 below).

However, technical measures designed to minimise discrimination have their limitations even where continuous improvement processes are used, partly because different technical fairness targets cannot be achieved simultaneously. Which criteria for non-discrimination and fairness are appropriate in which context is not a technical but a social and political question. Accordingly, as such, these decisions must not be entrusted to technology developers alone. Instead, they should be part of a future regulation of algorithmic systems and be included in the operational obligations of data controllers. The prerequisite for that is that the **criteria must be decided on specifically based on context as well as democratically**.

Algorithmic systems are difficult to analyse precisely. In order to be able to detect and prevent discrimination, the data controllers and oversight bodies must have the opportunity to gain an idea of any undesirable discrimination effects that occur within an algorithmic system, both within the context of its development and its productive deployment. Such effects can be identified through processes such as **risk assessments and output analyses**.

There is a tension between specifications to limit the collection and storage of discriminatory characteristics and the concern to retain the possibility to detect any discriminatory effects or be able to prove non-discrimination. These different requirements must be balanced on a case-by-case basis, which may have an influence on tests in different phases of the system development lifecycle; standard collation of all potentially discriminatory and therefore sensitive information for the sole purpose of proving that, as a result, no discrimination is taking place would not be justified. Greater efforts are needed here to produce **practical concordance between anti-discrimination law and data protection law**.

2.7 Transparent, explainable and comprehensible systems

In order to be able to carry out a reliable ethical and legal assessment of an algorithmic system, it is essential that enough information be available about its scope, functionality, pool of data and data analysis. **Only a truly transparent system can be examined to determine whether it is pursuing a legitimate purpose.** The transparency principle can have further key functions depending on the type and addressee of possible transparency obligations. With regard to the public, sufficient transparency must be created so that sufficient information is available for socio-political discourse on algorithmic systems. Supervisory authorities or other oversight bodies must be able to decide whether the legal and technical specifications are being or have been met where algorithmic systems are being used. Individual citizens must be able to take informed and confident decisions regarding the use of algorithmic systems and, in the event of negative effects on their freedoms and rights, be able to assess whether and to what extent they wish to exercise their rights. That too is a consequence of the ethical principle of digital self-determination.

In view of the increasing complexity of systems, the demand for transparency is, in practice, confronted with the fact that even experts are hardly able to go through all the individual components of a system fully, look at how they interact and **comprehend** everything within a reasonable amount of time. In particular in the case of individual machine learning methods, it is difficult, with today's state-of-the-art science and technology, to state which input led to a specific output of the system. There is also the fact that even technically simple algorithmic systems are often incorporated into complex social informatics ecosystems, i. e. information and work-sharing processes in which numerous manufacturers and operators are involved.

Example 11
The visual display of a personalised online advert is the result of complex processes in which the advert is delivered and paid for on the basis of behaviour-based analysis and segmentation. In particular, analytics services are used which are deployed by site owners across websites by incorporating the corresponding program code (such as JavaScript code for tracking). The components of such systems are also not fixed but can change, for example if manufacturers provide new versions or if they are adaptive and/or self-learning systems.

Legal aspects can also **limit** certain forms of information disclosure via algorithmic systems. Source codes and hardware designs are often protected as trade secrets. Operators also often have a legitimate interest in preventing their systems from being manipulated. Where algorithmic systems process personal data, data protection law can also limit the interest of the public or other affected citizens in information. However, where the transparency requirement regarding the system concerns the disclosure of the source code, which as such does not contain any personal data, data protection law does not stand in the way of disclosure.

However, the ever-present complexity cannot refute the goal of designing algorithmic systems to be transparent, nor can it justify any lack of transparency. Just like the aforementioned legal grounds, these aspects must nevertheless be taken into account in the drafting of any information rights and transparency obligations, which must be based on what is legally and actually possible. **The principle of transparency** also requires continuously developing technology to make the disclosure of information easier (for example through the use of open-source software and open hardware) and developing approaches which reduce complexity. Research is also required here. Under the banner of "explainable AI", researchers are working with increasing success on producing meaningful findings on the internal processes of algorithmic systems.

The demand for transparency must always take the **different levels of expertise** of the parties potentially interested in transparency into account. For example, the disclosure of the computer code to supervisory authorities carrying out necessary checks, may make it much easier for them to understand the system. Conversely, laypersons often need clearly and comprehensibly prepared information on a system's basic characteristics which enables them to carry out a risk assessment suitable for everyday purposes. At the same time, their interest is seldom limited to the system "itself". In order to prevent any negative decisions in the future, an **explanation** is rather also required as to how the decision specifically concerning them came about and which factors had what weighting. The specific drafting of the specifications on transparency and explainability should be based on the affected individuals' level of understanding and always be **comprehensible** for them. In that sense, rules on transparency and explainability will safeguard citizens' capacity to act and their self-determination.

2.8 Clear accountability structures

Just as having control over data implies the obligation to be accountable for such power, the opportunity to control algorithmic systems must also be accompanied by willingness **to answer for one's own actions**, i. e. to **be liable** where necessary.

Again, it is the complexity of algorithmic systems which, in practice, can make it difficult to assign responsibility. Hardware or software manufacturers, data providers, algorithm developers, operators of individual components, clients and users (either as the organisation or its individual employees) contribute to the system. Components are often used which can change without the knowledge or control of the user, for example as a result of important updates required for information security purposes. Those involved are often also located in different parts of the world. Efforts are required at all levels in order to prevent any diffusion of responsibility and **establish accountability structures**, starting with the technical design of the systems right through to legal specifications, for example in the form of the concept under data protection law of "joint control" (Article 26 GDPR).

2.9 Result: responsibility-guided consideration

Assessing the ethical aspects of algorithmic systems is, **in practice, extremely complex**. This is due to the large number of factors which need to be taken into account as well as the fact that, in a specific area of application, different individuals may be put in a "better" or "worse" position. The same can be said of social consequences and sustainability aspects which can rarely be unequivocally classified as either "positive" or "negative". However, this does not mean that humans can surrender all judgment. In cases where it is difficult to weigh everything up, everyone is required to take particular care with their assessments and decisions. Where algorithmic applications may potentially develop such phenomenally impressive performance and scope that questions are raised concerning the future of mankind, weighted assessments of the opportunities and risks will increasingly reach their limits, and more fundamental anthropological and ethical discussions will be required. This is precisely where the principle of prospective responsibility is of fundamental importance.

With regard to all this, the **democratic process** provides ways and means for balancing conflicting convictions, ideally supported by special **deliberative processes and institutions** through which society can ensure, in as inclusive and participatory a way as possible, that the challenges presented by algorithmic systems are addressed.

It should only rarely be the case that human activity and the use of an algorithmic system do not need to be weighed up against each other because the latter, in all ethically relevant respects, achieves a "better" result than humans using conventional technology. Where this is the case however, the Data Ethics Commission believes that the use of algorithmic systems is **ethically commanded**, because a general ethical preference for human activity over the use of machines at the expense of the protection of important legally protected rights is not justified in the view of the Data Ethics Commission. However, with regard to the question as to whether human or machine activity is preferable (→ see Part B, section 1 above), other factors will routinely need to be taken into consideration, such as the emotional well-being of people, human skills retention and sustainable development, which ultimately requires weighing up the options. This may go against or in favour of the algorithmic system.

Example 12
If the use of a diagnostic algorithmic system in a specific clinical area leads to just 2% of patients dying, whereas 10% of all patients would die as the result of human misdiagnoses, the use of the system would, depending on the circumstances of the specific case, be ethically advisable even if, as a result, minor but tolerable reductions in patients' emotional well-being occurred and additional measures would have to be taken to ensure human skills retention.

However, if, taking all circumstances into account, the use of an algorithmic system, at the expense of important legally protected rights, leads to an inferior result than the use of conventional technology and humans (for example because more wrong decisions are made) and there is only an increase in efficiency or convenience, the use of algorithmic systems must, in principle, be **rejected for ethical reasons**. However, ethically defensible exceptions could be made in this case based on economic considerations if there would be only a minimal impairment but an exceptionally high potential saving which would benefit the public good.

3. Recommendation for a risk-adapted regulatory approach

From a regulatory point of view, the fact that algorithmic systems need to be assessed very differently from an ethical perspective, depending on their intended purpose, performance, robustness and security as well as in terms of their impacts, suggests that a **risk-adapted regulatory approach**[1] is required. It follows the principle that the **greater the potential** of algorithmic systems **to cause harm**, **the more stringent the requirements and the more far-reaching the intervention** by means of regulatory instruments. The risk spectrum of algorithmic systems therefore ranges from systems, the application of which involves low risk, right through to systems which could lead to irreversible harm for individuals and society. Causes of risks can, for example, be inadequate models, an unsuitable pool of data, in particular in the case of self-learning systems, or inappropriate basic assumptions and weighting (→ see sections 2.3 and 2.6 above).

Potential **harm** caused by algorithmic systems can vary in nature and can include financial loss, non-material damage and physical harm. For example, individual applications can cause potentially serious financial loss (for example lending or insurance terms), affect opportunities for participation (for example discrimination in hiring) and involve violations of fundamental rights and risks to the life and health of consumers (for example in the case of robotic nurses or mobility applications).

The overarching objective of regulating the use of algorithmic systems is to prevent detrimental effects at the individual and supra-individual level. In particular where algorithmic systems affect matters which are sensitive in terms of fundamental rights, legal provisions concerning the design of the systems are also needed. Regulation should strive to intervene as much as necessary and as little as possible in order not to hamper innovation and creativity while at the same time ensuring the protection of fundamental rights, freedoms and values. **Efficient and proper regulation** can help increase public trust in the use of algorithmic systems: The public perception of self-learning systems in particular is that they are not controllable, which adds to corresponding scepticism towards technology.[2]

The Data Ethics Commission takes the view that the primary addressees of regulation should be the **manufacturers** and **operators** of algorithmic systems. Due to the State's direct obligation to uphold fundamental rights, it is necessary to differentiate, however, between **private and state use** of algorithmic systems (→ see section 7 below in particular) when the regulation is drawn up in more detail. Given the model and role model character of state action, the Federal Government is advised to exercise particular care when using algorithmic systems for state purposes.

3.1 System criticality and system requirements

A risk-adapted regulatory approach can be made more concrete by orienting it towards the criticality model of an algorithmic system. **System criticality** is based on the system's potential to cause harm, which is determined based on the likelihood that harm will occur and on the severity of that harm.

1 Compare in particular Tobias Krafft / Katharina Zweig: Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse [Transparency and traceability of algorithm-based decision processes], Studie im Auftrag des Verbraucherzentrale Bundesverband e.V. (vzbv) [Study commissioned by the Federation of German Consumer Organisations (vzbv)], 22 January 2019, pp. 18 et seqq. (available at: *https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf*).

2 Sarah Fischer / Thomas Petersen: Was Deutschland über Algorithmen weiß und denkt – Ergebnisse einer repräsentativen Bevölkerungsumfrage [What Germany knows and thinks about algorithms – results of a representative population survey], Bertelsmann Stiftung, 2018 (available at: *https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/was-deutschland-ueber-algorithmen-weiss-und-denkt/*).

The **severity** of harm that could potentially result, for example from a faulty decision, depends among other things on the significance of the legally protected rights and interests affected (in particular, for example, the right to determine the use of one's personal data, to freedom of expression, the fundamental right to life and physical integrity, as well as to equal treatment) and the extent of the potential harm resulting from an infringement. Furthermore, the assessment of the severity of the potential harm must take into account the specific sensitivity of the data used, the level of potential harm for individuals or groups (including non-material harm or loss of utility that are hard to calculate in monetary terms), the number of individuals affected, the total figure for the potential damage and the harm to society as a whole, which may go well beyond a straightforward summation of the harm suffered by individuals. The consequences of using an algorithmic system should, based on its area of application, be considered in terms of its ecological, social, psychological, cultural, economic and legal dimensions. The general ethical values and principles (→ see Part B above) set the standard with regard to the assigned value.

The **likelihood** that harm will occur is also influenced by the following system properties, and factors:

- the role of algorithmic calculations in the decision-making process (from the mere inspiration of humans without any claim to accuracy up to algorithm-determined decisions, → see section 1 above);
- the complexity of the decision to be made (from a simple deterministic depiction of reality or a probabilistic appraisal of reality up to the multifactorial and non-determinate prediction of a future reality);
- the effects of the decision (from a purely abstractly conceivable context of action or a specific context of action up to direct implementation); and
- the reversibility of the effects (from full reversibility up to irreversibility).

The likelihood of the potential harm and the severity of that harm may also depend on whether it is a **state or private** party taking action and, particularly in economic contexts, on the **market power** of the party using the algorithmic system. This is due to the fact that the state or private nature of the action and market power are not only relevant in terms of the obligation to uphold fundamental rights and potential harm to society as a whole. They also determine possible alternative options for those affected. Where **affected persons depend** on an algorithmic system, for example in terms of access to markets, goods and services, the criticality increases. The limitation of options can be due to various different causes, for example network effects, effects of scale and effects of scope which can, in turn, be reflected in market power and (a lack of) equivalent alternatives.

The greater the system criticality, the stricter the **requirements** that have to be imposed on the system from a regulatory perspective. These requirements are being formed out, in particular, by

a) corrective and oversight mechanisms;

b) specifications regarding the transparency of algorithmic systems and the explainability and comprehensibility of the results; and

c) rules on the assignment of responsibility and liability within the context of the development and use of algorithmic systems (→ see sections 4, 5 and 8 below).

The variety, complexity and dynamics of algorithmic systems pose major challenges for regulation which cannot be based on a limited toolbox but must, depending on the system's criticality, implement **very different corrective and control instruments at different regulatory levels** in order to achieve the objectives of regulation and ensure that the risks involved in the systems are manageable. The spectrum of possible instruments ranges from forgoing special legal provisions and "soft" incentives for self-regulation, giving authorities the right to monitor, and requiring any final decision to be taken by a human, up to banning certain intended purposes and contexts for using algorithmic systems.

Provisions regarding the **transparency** of systems and the **explainability** and **comprehensibility** of their results (→ see section 2.7 above) are key components of a corrective and control regime for algorithmic systems. Also, to that extent, the criticality of a system determines the scope of any rights to information and obligations to provide information. How the information requested can be comprehensibly communicated varies depending on the addressees of the system and hence also the intended purpose and usage context.

From an ethical and legal perspective, it is crucial, for all dealings with algorithmic systems, that **responsibility** for their impacts can be clearly assigned to human decision-makers at all times. Rules on **liability** are, in particular, also of key importance here, while the question of the proper organisation of a liability regime for certain digital products, content and/or services must also be addressed with a view to the criticality of the system (→ see section 8 below).

In terms of the governance perspective adopted by the Data Ethics Commission, **all relevant stakeholders** (the State, companies, developers and the public) must participate in **specifying and drawing up** these differentiated **regulatory requirements**. The Data Ethics Commission points out that, even without any special regulation, the use of algorithmic systems must be measured against general legal norms. These include in particular civil liability law, which fundamentally states that compensation is mandatory in the event of action which infringes legally protected interests. The provisions of existing regulation against unfair competition also apply, for example in the event that consumers are misled, as well as criminal law if crimes are committed with the help of algorithmic systems. When examining the conditions of these norms, the criticality of the systems and the resulting system requirements also have legal significance in accordance with general standards.

Algorithmic systems are used in order to fulfil specific functions. In order to assess system criticality, the **ethical assessment of the intended purpose** is therefore also of crucial importance. If the intended purpose is ethically indefensible, for example because it infringes fundamental rights and freedoms or breaches the free democratic basic order, then there are "red lines" and "absolute limits" – both for algorithmic systems and for humans. For example, an algorithmic system used for political manipulation, fraud or collusive price-fixing must be seen per se as ethically objectionable.

The intended purposes are often multifaceted, and individual facets, in particular regarding secondary purposes, may each need to be assessed differently from an ethical perspective. Identifying an intended purpose which is decisive for the assessment often, in that sense, requires difficult **value judgments**. Assessing the intended purpose of algorithmic systems is further complicated in the case of digital products because the development and market launch phases increasingly overlap; the intended purpose of a product may also change after it has been launched on the market due to updates or deployment in other usage contexts.

Complex intended purposes in the case of media intermediaries

A number of media intermediaries, such as search engines, are essential in the Internet age because they provide access to information online, channel the flood of information and actually enable individuals to use the Internet in the first place. To that extent, their purposes are desirable and unproblematic in ethical terms. However, media intermediaries can be ethically problematic in terms of their specific design. Their systems provide users with a personalised selection of information which leads to selection of the displayed content. However, since as a result the overwhelming majority of content is not displayed or is only displayed with a lower priority, the individual's spectrum of perception is narrowed. As such, the intermediary decides, through programming, over the user's head as to what the user sees. As far as the business models of media intermediaries are driven by advertising, as is the case with major social networks, there is a risk that operators will have an economic interest in disseminating also ethically questionable or even extremist content because it promises to keep users on the platform longer, thus increasing advertising revenue. Due to the interplay of the sorting and narrowing of what is seen and the additional danger of influencing the user through non-transparent third-party interests, there is the possibility that influence will be non-transparently exerted, for example over the political decision-making process, and could even result in political manipulation. This is a significant danger for the free formation of opinions as a basic foundation of democracy.

3.2 Criticality pyramid

The Data Ethics Commission recommends consistently determining the degree of criticality of algorithmic systems using **an overarching model**. The degree of criticality should guide legislators and society when seeking suitable regulatory thresholds and instruments, but can also provide developers and operators with guidance for assessing their products and systems themselves and finally also be used in basic, advanced and further training to educate and **increase awareness amongst** various stakeholders. To that extent, with regard to the potential of algorithmic systems to cause harm, the Data Ethics Commission differentiates, both for private and for state operators, between **five levels of criticality**:

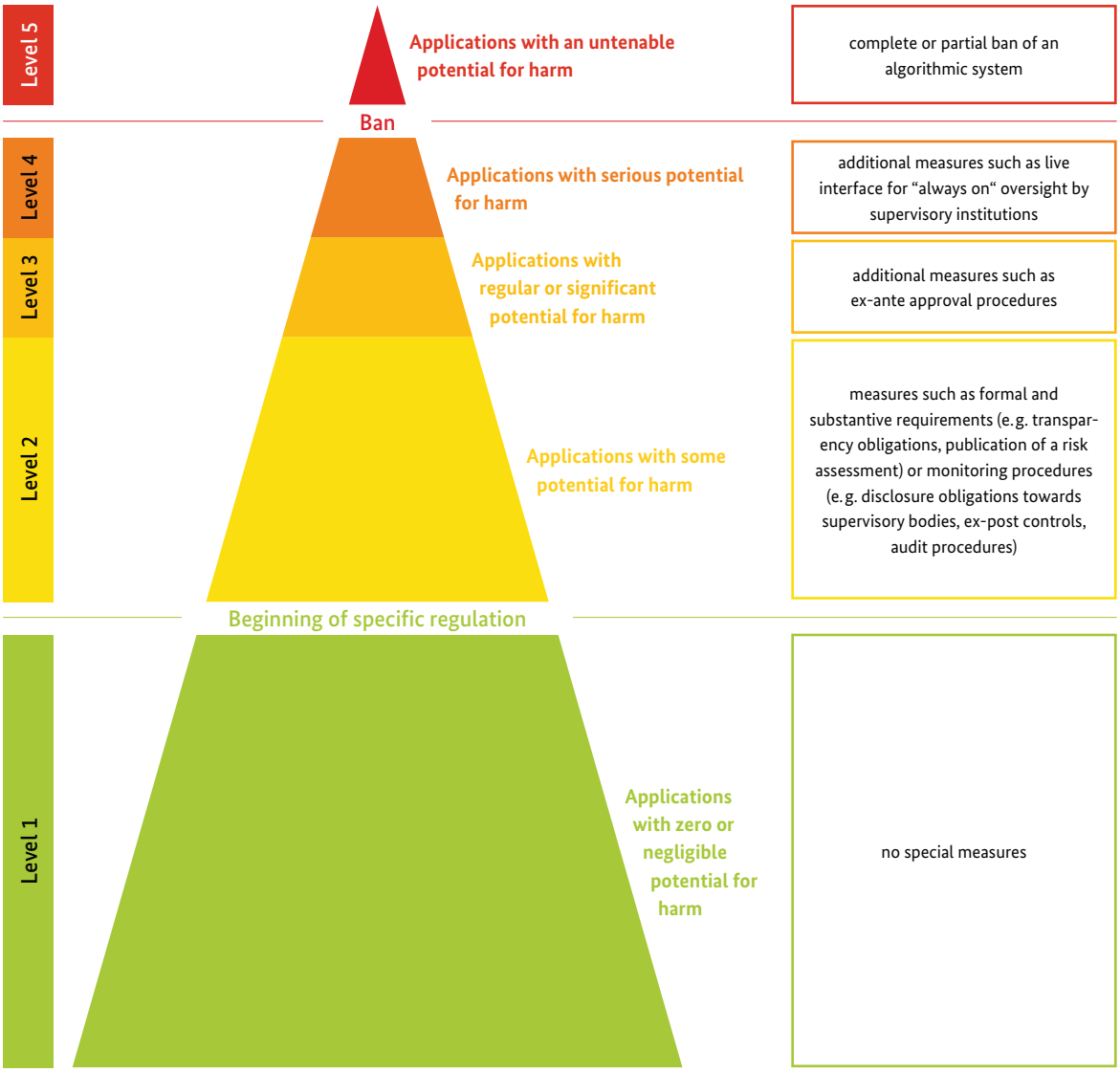


Figure 8:
Criticality pyramid and risk-adapted regulatory system for the use of algorithmic systems

In unproblematic usage contexts, it will normally not be necessary to require developers, clients or operators to go through specific ethical and legal oversight procedures. For the many **applications with zero or only negligible potential for harm**, i. e. on the lowest level (Level 1) of the criticality pyramid, the Data Ethics Commission sees no need for special oversight which would go beyond the general quality requirements which apply even to products without algorithmic elements.

Example 13
The algorithms used in a drinks vending machine do have a certain potential for harm, since a user could, for example, not receive any goods and lose his or her money. However, this potential for harm does not exceed the threshold for specific potential for harm within the algorithm context. It is sufficient here to rely on the general mechanisms which oblige contractual partners to fulfil their contractually undertaken performance obligations or manufacturers to produce devices which function properly.

In the case of **applications with some potential for harm** (i. e. on Level 2 of the criticality pyramid), regulation can and should be implemented. However, the scope of the necessary measures is limited here. In view of the low level of criticality, any excessive burden on manufacturers and operators should specifically be avoided here in order not to excessively hinder technological or social innovations or market development. Measures which could be offered at Level 2 include for example ad-hoc ex-post controls (for example in the form of an input-output control), if there is reason to suspect that the system is malfunctioning. Furthermore, there should be an obligation to produce and publish an appropriate risk assessment (→ see section 4.1.3 below). In addition, on a sector-specific basis, obligations to disclose information to supervisory institutions (including establishing an interface for a supervisory institution to carry out input-output controls), increased transparency obligations as well as access rights for individuals affected (→ see section 4.1 below for more details) may be useful. Codes of conduct should also be considered which would be developed specifically for each industry and then approved by the competent supervisory authorities. Compliance would then need to be tested by the supervisory authorities using spot checks as well as on an ad-hoc basis (→ see section 5.2 below).

Criticality in the case of smart mobility applications

A provider of smart mobility applications has access to a data pool generated using all vehicle and mobility data. If these data are used exclusively for predicting traffic jams, the level of criticality should be classified as negligible. However, the flow of traffic can also be controlled using smart mobility. If algorithms can, for example, identify which route is the optimum route for travelling from A to B based on the overall usage of the mobility system consisting of road, rail, water and air transport determined in real time using the vehicle data, a corresponding route can be suggested to the user based on the user's preference (e. g. fastest/most environmentally friendly/cheapest, etc. route). However, there is also the question as to whether the State can stipulate certain routes for the user in consideration of state-prescribed criteria. Here, in view of the changed potential for harm, the level of criticality would be higher and would therefore require stricter regulation as appropriate.

Example 14
Dynamic pricing (for example based on the criteria of supply and demand) in e-commerce, which however does not involve any personalised pricing, has a potential for harm that is generally low but still exceeding the threshold of relevance, for example concerning covert discrimination.

Example 15
Price algorithms for setting personalised prices (i. e. setting a price based on criteria which are tailored to the individual customer and usually estimate their maximum personal willingness to pay) involve appreciable potential for harm, for example concerning discrimination against particularly vulnerable groups. At best, it should be possible to use them only after they have undergone a licensing procedure.

In the case of **applications with regular or tangible potential for harm** at Level 3 on the criticality pyramid, in specific cases, in addition to the mechanisms already required for Level 2, an ex-ante control in the form of a licensing procedure may be justified (→ see section 4.2.5 below). On account of the fact that many algorithmic systems are highly dynamic, a regular review will be required in the event that a licence is granted.

The same must apply for **applications with significant potential for harm** at Level 4 as applies for Levels 2 and 3. However, here, additional oversight and transparency obligations, which may extend all the way through to the further publication of information on the factors that influence the algorithmic calculations and their relative weighting, the pool of data used and the algorithmic decision-making model in a comprehensible format, should be required or even “always-on” oversight via a live interface should be provided for. Further protective measures to prevent harm are also necessary.

Differentiated criticality in the case of media intermediaries

With the help of their algorithmic filtering systems, media intermediaries process and communicate both content relevant for the formation of opinions, which is relevant for the democratic decision-making process, and content used for advertising, purchase recommendations or entertainment. They therefore represent the perfect example of situations in which the use of the same algorithmic system has differing potential for harm. In the case of user interaction in the consumer goods sector (in particular advertising or purchase recommendations), depending on the personalisation model used, there will be a low to high potential for harm. As soon as balanced variety must be produced (in particular in the case of topics relevant to the formation of opinions) on account of overarching interests in maintaining the free democratic basic order, the potential for harm is already higher right from the outset due to the content. As a result, the regulatory requirements change simultaneously. In the case of consumption and entertainment offerings, depending on the personalisation criteria used, the usage contexts or the welfare effects to be expected, more or less stringent regulation must ensue.

Example 16
Algorithmic systems, for example of players with huge market share, which are used to determine the creditworthiness of an individual consumer or company must be classified as Level 4. Whether a person receives a loan or not can have a decisive bearing on that person's fate. The high level of system criticality is also justified by the market concentration with few providers and the tendency for a lender to rely on the judgment of a particular player.

With regard to the system criticality criteria, it may ultimately be worth considering a complete or partial ex-ante **ban** on the use of an algorithmic system for **applications with untenable potential for harm** (Level 5). An ex-post ban may also be used as a consequence for breaches of applicable law or non-fulfilment of the system requirements set out for the specific system criticality.

Example 17
Lethal autonomous weapons systems are often seen as a "red line", as machines should not be allowed to kill people. However, that can apply only on the basis that they are algorithm-determined killings. Where lethal autonomous weapons simply provide human soldiers with support in recognising objects or are merely used to keep a missile on track in the face of crosswinds, an ethical "red line" is not being crossed.

The classification of an algorithmic system in the criticality pyramid must, where necessary, be **regularly reviewed** in the light of the dynamic nature of these systems.

3.3 EU regulation on algorithmic systems enshrining horizontal requirements and formed out in sectoral instruments

Algorithmic systems are infiltrating more and more areas of our personal and social lives. The purposes of algorithmic systems and the areas in which they could potentially be used are therefore not set in stone. For example, a facial recognition system developed for use with private photos could also be used by state investigative authorities for law enforcement purposes or to prevent threats. This suggests addressing the challenges posed by algorithmic systems following the example of data protection law in the form of **horizontal regulation**, i. e. through a legal instrument, the material scope of which covers algorithmic systems in general, and which applies to **private and public players** alike. In addition to the considerable symbolic power, another point in favour of horizontal regulation is the fact that gaps in protection would be eliminated and dangerous situations which currently cannot be foreseen would be covered. One of the main arguments in favour of such overarching regulation which sets out basic principles for all algorithmic systems is also the fact that citizens would, as a result, have a clear idea of what to expect in all areas, and (European) legislators could complete this task within a reasonable period of time.

As a result, **the Data Ethics Commission recommends** that the Federal Government should work towards drawing up horizontal basic regulation at the European level in the form of an **EU Regulation on Algorithmic Systems (EU-ASR)**. In addition to the key basic principles for algorithmic systems developed here as requirements for algorithmic systems, the horizontal legal instrument should group together general substantive rules – informed by the concept of system criticality – on the admissibility and design of algorithmic systems, transparency, the rights of individuals affected, organisational and technical safeguards and supervisory institutions and structures.

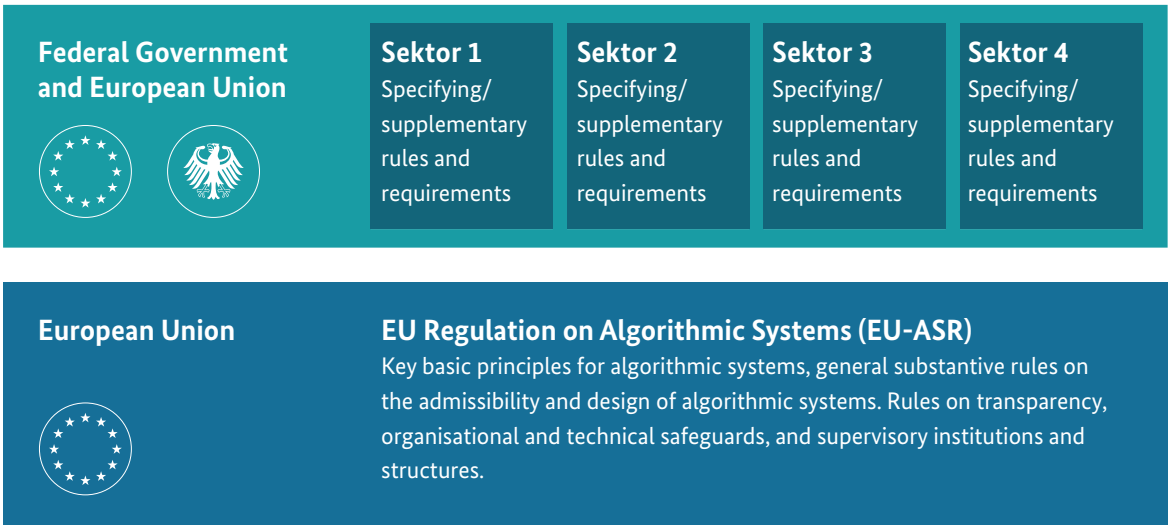


Figure 9:
EU regulation on algorithmic systems enshrining horizontal requirements and specified in sectoral instruments

At the same time, the Data Ethics Commission recommends that the Federal Government should also advocate sectoral rules on the European level and, outside the competences of the EU itself and within its own legislative and administrative competences, enact appropriate sectoral legal acts which are oriented towards system criticality. (Fig. 9).

An **overarching EU-ASR** will have to be limited to few **basic principles**, as otherwise European legislatory powers would be overburdened. Legislators would, if rules were too detailed, in particular face the issue of how to deal, in a general legal instrument, with the wide variety of systems of which it is now almost impossible to keep track and the highly dynamic development of technology. From the perspective of those affected, general legal instruments also carry the risk that the administrative obligations will also apply in cases where there is not sufficient potential for harm, because a horizontal legal instrument cannot distinguish between risky and less risky operational aims (as well as potential exceptional configurations) with the same level of detail that they have in reality. With regard to both points, the **supplementary** recourse to **sector-specific legislation** which would be limited in terms of scope but would therefore be easier to form out would relieve some of the burden. Any supplementary sector-specific approach would also have to take into consideration the legislative and administrative powers distributed in accordance with applicable law between the EU, the Federal level and the States *(Bundesländer)*. An additional fact is that, with regard to the official oversight and supervisory institutions and structures, for various reasons there could be no question of consolidating and assigning the "overall task" to one single authority (→ see section 5.1 below).

Therefore, in addition to the EU-ASR it will be necessary to enact several **legal instruments with specific provisions for individual sectors or potentially harmful situations**. In the view of the Data Ethics Commission, combining a general basic regulation with further sector-specific legal instruments has the major advantage of enabling differentiation between the different needs for protection involved for individual systems and usage contexts. This is in line with the basic concept behind risk-adapted regulation, according to which the regulatory requirements for algorithmic systems should be determined based on the specific system criticality.

Even in data protection law, in the public sector, there are numerous special laws which supplement the general provisions of the GDPR for different sectors. The basic idea behind data protection law is that, in the case of automated data processing, there is no longer such a thing as "inconsequential" data, which is why it is hardly possible any more to differentiate meaningfully between personal data on the basis of worthiness of protection or criticality in the absence of common basic rules. Nonetheless it is also true that a variety of special provisions ensures an increased level of protection in the wide range of areas of state activity. Similarly, there is an according need for supplementary sectoral provisions for algorithmic systems. The application of such regulation also does not have to fall short as a result of the fact that their purpose and usage context could change. After all, firstly, such a change would be, especially in more complex systems, inherently limited. Secondly, the issue could be addressed from a regulatory perspective by the fact that the legal instruments would not be materially linked to the original purpose or original usage context but the **current functionality of the system or the new intended purpose** of the system. In this way, any changes in purpose and context would, if necessary, result in the application of a differentiated regulatory framework.

However, these primarily pragmatic considerations in no way affect the requirement for the standard-setting body or bodies to ensure the greatest possible **coherence between legal instruments** in their respective undertakings. This will apply not only to the regulatory approaches developed here, i. e. in particular the notion of system criticality, and the rights of data subjects; regulatory infrastructures and processes should also be designed as uniformly as possible.

Summary of the most important recommendations for action

Risk-adapted regulatory approach

36

The Data Ethics Commission recommends adopting a **risk-adapted regulatory approach** to algorithmic systems. The principle underlying this approach should be as follows: the greater the potential for harm, the more stringent the requirements and the more far-reaching the intervention by means of regulatory instruments. When assessing this potential for harm, the **sociotechnical system as a whole** must be considered, or in other words all the components of an algorithmic application, including all the people involved, from the development phase – for example the training data used – right through to its implementation in an application environment and any evaluation and adjustment measures.

37

The Data Ethics Commission recommends that the potential of algorithmic systems to harm individuals and/or society should be determined uniformly on the basis of a **universally applicable model**. For this purpose, the legislator should develop a **criteria-based assessment scheme** as a tool for determining the criticality of algorithmic systems. This scheme should be based on the general ethical and legal principles presented by the Data Ethics Commission.

38

Among other things, the **regulatory instruments and the requirements that apply to algorithmic systems** should include corrective and oversight mechanisms, specifications of transparency, explainability and comprehensibility of the systems' results, and rules on the allocation of responsibility and liability for using the systems.

39

The Data Ethics Commission believes that a useful first stage in determining the potential for harm of algorithmic systems is to distinguish between **five levels of criticality**. Applications that fall under the lowest of these levels (Level 1) are associated with zero or negligible potential for harm, and it is unnecessary to carry out special oversight of them or impose requirements other than the general quality requirements that apply to products irrespective of whether they incorporate algorithmic systems.

40

Applications that fall under Level 2 are associated with **some potential for harm**, and can and should be regulated on an as-needs basis; regulatory instruments used in this connection may include ex-post controls, an obligation to produce and publish an appropriate risk assessment, an obligation to disclose information to supervisory bodies or also enhanced transparency obligations and access rights for individuals affected.

41

In addition, the introduction of licensing procedures may be justified for applications that fall under Level 3, which are associated with **regular** or **significant potential for harm**. Applications that fall under Level 4 are associated with **serious potential for harm**; the Data Ethics Commission believes that these applications should be subject to enhanced oversight and transparency obligations. These may extend all the way through to the publication of information on the factors that influence the algorithmic calculations and their relative weightings, the pool of data used and the algorithmic decision-making model; an option for "always-on" regulatory oversight via a live interface with the system may also be required.

42

Finally, a complete or partial ban should be imposed on **applications with an untenable potential for harm** (Level 5).

43

The Data Ethics Commission believes that the measures it has proposed should be implemented in a new EU Regulation on algorithmic systems enshrining general **horizontal requirements (Regulation on Algorithmic Systems, EU-ASR)**. This horizontal regulation should incorporate the fundamental requirements for algorithmic sytems that the Data Ethics Commission developed. In particular, it should group together general substantive rules – informed by the concept of system criticality – on the admissibility and design of algorithmic systems, transparency, the rights of individuals affected, organisational and technical safeguards and supervisory institutions and structures. This horizontal instrument should be fleshed out in **sectoral instruments** at EU and Member State level, with the concept of system criticality once again serving as a guiding framework.

44

The process of drafting the EU-ASR (as recommended above) should incorporate a debate on how best to demarcate the respective scopes of this Regulation and the **GDPR**. A number of factors should be taken into account in this respect; firstly, algorithmic systems may pose specific risks to individuals and groups even if they do not involve the processing of personal data, and these risks may relate to assets, ownership, bodily integrity or discrimination. Secondly, the regulatory framework introduced for the future horizontal regulation of algorithmic systems may need to be more flexible and risk-adapted than the current data protection regime.

4. Instruments: obligations of data controllers and rights of data subjects

In order to provide individuals and groups with effective protection against the dangers of algorithmic systems, the Data Ethics Commission believes that both transparency requirements (→ see section 4.1 below) and further specifications for algorithmic systems with a view to effective protection against substantively inappropriate decisions and unfair decisions (→ section 4.2.) are advisable.

4.1 Transparency requirements

4.1.1 Mandatory labelling ("if")

A key tool for creating transparency is **mandatory labelling**. Because a mandatory labelling scheme requires little detailed information, infringements of the fundamental rights of system operators, in particular with regard to their business secrets, are also less serious than in the case of access rights. The Data Ethics Commission believes that this justifies establishing labelling in the case of critical systems (as from Level 2) as a blanket obligation for system operators and not as a request-based right for the individuals affected.

Due to the comparatively narrow scope of Article 22 GDPR (relating to a decision based solely on automated processing), to which the duties to provide information refer, the Data Ethics Commission believes that the existing labelling obligations of the GDPR[3] are **insufficient**. In particular, significant impacts for affected individuals can arise even below the threshold of Article 22 GDPR. That applies for algorithm-based and algorithm-driven decisions, i.e. situations in which the humans taking the decisions run the risk of accepting algorithmic information and proposed decisions without reflection and by default (in particular in areas where human assessment is expected) or only following algorithmically determined and prescribed paths.

Because the Data Ethics Commission sees the authenticity of interpersonal communication as a fundamental condition for trustworthy interaction within society, a mandatory labelling scheme should always apply if there is any **risk of confusion** between human and machine and should therefore apply irrespective of system criticality. This applies, for example, to digital voice assistants and chatbots which these days are sometimes hard to identify as such. Labelling may, in the case of voice assistants for example, be carried out both by means of a regular reminder of the assistant's mechanical nature (even during ongoing communication) and also through the use of a mechanical-sounding voice. Conversely, the Data Ethics Commission considers that there is no risk of confusion (and therefore also no need for a mandatory labelling scheme) in areas where the nature of the information is irrelevant or the recipient expects a mechanical voice anyway, such as in the case of loudspeaker announcements at railway stations.

4.1.2 Duties to provide information, duties to provide an explanation and access to information ("how" and "what")

Whilst mandatory labelling schemes require system operators to ensure transparency regarding as to whether and the extent to which algorithmic systems are used ("if"), duties to provide information and **rights of access** are regularly focused on more detailed information regarding the decision-making mechanism ("how") and the data used ("what") by the algorithmic system.

3 Article 13(2)(f), Article 14(2)(g) and Article 15(1)(h) in conjunction with Article 22 GDPR.

Duties to provide information and rights of access regarding the behaviour of algorithmic systems and the way that decisions are made inside the systems are important from the perspective of citizens for them to be able to understand decisions and review them and/or have them reviewed individually. Only with their help can data subjects exercise their rights and challenge a decision on an informed basis. The following transparency requirements apply equally to private and state operators of algorithmic systems. Special requirements with regard to the transparency of systems used by the State will be covered in more detail in section 7 below.

4.1.2.1 Duties to provide information and rights of access

Articles 13, 14 and 15 GDPR already set out duties to provide information and rights of access where personal data are processed. In the event of automated decision-making within the meaning of Article 22 GDPR, the GDPR grants the data subject a right to "meaningful" information about the "logic involved", as well as the "significance" and the "envisaged consequences" of the processing.[4]

The Data Ethics Commission takes the view that, just as in the case of the mandatory labelling scheme (→ see section 4.1.1 above), the legal concept behind these norms should also apply outside of the narrow scope of Article 22(1) GDPR and be an integral part of the EU-ASR suggested here (→ see section 3.3 above). The extent of such a duty to provide information will depend on the **criticality of the system**. In the case of applications with negligible potential for harm, brief statements on the logic behind decisions will suffice, for example on the pool of data used or the general weighting of certain factors with regard to the result. The more risk a system involves, the more extensive the duties to disclose information will essentially be.

The more sensitive a decision is in terms of personality, the more detailed information relating to the individual case is needed. However, it should also be borne in mind that providing detailed information regarding the factors and their weighting could also have potentially ethically questionable influence on the private lifestyle of the data subject. Furthermore, the data subject could also use the acquired information to undermine an algorithmic system which performs an important function.

The **technical and organisational requirements** which must be met in order to be able to fulfil these extensive duties to provide information must be incorporated in the design of algorithmic systems right from the outset, as it will be possible to ensure that the systems are operated lawfully only if the corresponding necessary "meaningful" information can also be provided when the system is used.

When defining duties to provide information and rights of access in order to increase the transparency of algorithmic systems, care should be taken to ensure that no special technical skills or knowledge are required of consumers. Whenever rights of access are expanded, it should be borne in mind that, from the perspective of data subjects, this will increase transparency only if the information is prepared **in a way which is suitable for the recipient**.

4 Article 13(2)(f), Article 14(2)(g) and Article 15(1)(h) GDPR.

4.1.2.2 Duties to provide an explanation

At least in certain areas of complex algorithmic systems, it may be appropriate, in addition to the general explanation regarding the system's logic and significance, to require an explanation of the specific reasons why the system made a recommendation or decision. Such a specific explanation is required above all if the decision concerns areas which are sensitive in terms of personality or otherwise is of particular significance in terms of fundamental rights or socioeconomics. It is important, in such cases, for data subjects to be informed in a comprehensible, relevant and clear manner. The Data Ethics Commission therefore welcomes the technical efforts to improve the explainability of algorithmic (in particular self-learning) systems (explainable or explicable AI), and encourages the Federal Government to promote such projects.

The Data Ethics Commission believes that, in certain situations, it is worth considering an entitlement to **"counterfactual explanations"** as is sometimes discussed in the literature.[5] In such cases, data subjects are informed of the factors in the decision-making process which, in the case of a (negative) decision for them, would have made the positive difference, i. e. would have actually led to the desired outcome. In a case where an application for a loan has been rejected based on the use of an algorithmic system, the data subject would, for example, be entitled to learn from the system operator which of the factors taken into consideration by the system would have had to have been different, and in what way, for the application to have had a positive outcome. However, the Data Ethics Commission points out that this approach quickly reaches its limits in the case of more complex systems, as the data subject would have to be provided with a whole host of different "counterfactual" scenarios here in order to be given a reasonably complete picture; otherwise there would be a danger of misinformation, questionable steering or even manipulation by focusing on certain aspects for strategic or educational reasons.

In the view of the Data Ethics Commission, given the current state of technical development, the concept of "counterfactual explanation" is therefore not suitable for use as a general component of any regulation of algorithmic systems; however, it could be considered for special processing situations.

4.1.2.3 Access to information for not directly affected persons

In addition, the Data Ethics Commission considers that, in certain sectors in which not only individual but also social interests are affected to a significant extent, it is advisable even for individuals not directly affected to be granted a right of access to information regarding the algorithmic systems. This would apply, in particular, if their use were **relevant for public opinion-forming** or had major **welfare effects** for the population. Such rights would, first and foremost, be worth considering for journalistic and research purposes and would also have to be accompanied with adequate protective measures for any affected interests of system operators.

Under certain circumstances, in particular in the event of the State's use of systems with significant potential for harm, **unconditional rights of access to information** and **publication requirements** are also conceivable in the view of the Data Ethics Commission.

5 Sandra Wachter / Brent Mittelstadt / Chris Russel: Harvard Journal of Law & Technology 2018 (31), pp. 841 et seqq.

4.1.2.4 Requirements for defining duties and rights, in particular in consideration of system operators' rights

When defining duties to provide information and explanations and rights of access, it must always be borne in mind that these may also affect the **legally protected interests of the operators** of algorithmic systems, as well as of those who use their outputs. This includes, most notably, the protection of business secrets and the interest in preventing any manipulation of the systems and manipulative use of the systems. Private system operators can, in principle, invoke the fact that they define their own free-will decisions and contractual decisions based on the outputs of an algorithmic system. However, that does not release them from monitoring required to check whether they are acting in accordance with the law, as the fundamental right to freedom of action is restricted by bans on discrimination (in particular the General Act on Equal Treatment), the fundamental rights of the data subjects or third parties and, in general, the provisions (and specific contractual provisions) of the legal system. Furthermore, transparency rights must always be balanced with the provisions of data protection law relating to the protection of the personal data of third parties stored in the system.

The Data Ethics Commission therefore believes that it is appropriate for legislators to accompany transparency obligations with rules which, at the initiative of the system operators or also possibly affected third parties, enable **the conflicting rights and interests to be weighed** against the transparency interests of the data subjects or other private individuals entitled to claim rights. However, in the view of the Data Ethics Commission, **rigid rules of priority**, for example a general preference for the protection of business secrets over transparency interests, are **not appropriate for the matter concerned**, despite the increase in legal certainty they might bring. Where system operators or third parties invoke conflicting interests, meticulous checks must be carried out to see whether such interests cannot be taken into account with specific protective measures before a transparency obligation is completely rejected. If private individuals have rights of access to information, the requirements regarding the protective measures and the demonstration of their existence must be devised so that they do not act as a barrier preventing vulnerable consumers and/or citizens from acquiring information. Interests of third parties must be protected for example by means of anonymisation.

4.1.3 Risk impact assessment

The impact assessment within the meaning of Article 35(1) GDPR concerns only information on the impacts for the protection of *personal data*; however, it does not include a comprehensive risk analysis of an algorithmic system. In the case of algorithmic systems, as from a certain level of potential for harm, it is, however, appropriate and reasonable legally to require the provider/user to produce and publish an appropriate risk impact assessment in order to assess the risk involved with the system. The more critical the system is, the more comprehensive the risk impact assessment must be. It should also cover an assessment of the **risks relating to self-determination, privacy, bodily integrity and personal integrity, as well as assets, property and non-discrimination**, and also include methods for gauging the quality and fairness of the data and the model accuracy, for example the bias or the rates of (statistical) error (overall or for certain sub-groups) exhibited by a system during forecasting/category formation.

Use Case: Personalised prices I – transparency requirements

The increasing use of pricing algorithms in e-commerce presents challenges not only for consumer protection law but also for competition law: pricing algorithms can review the market in order to adjust prices in line with demand and competitors' offers in real time.

In e-commerce, providers can therefore apply personalised prices (for individual users or groups) directly or via individual discounts. Algorithmic systems can, for example, be used specifically to cash in on consumers' maximum willingness to pay or encourage users not to abort a purchase transaction. This personalisation is based on scoring processes, for example using real-time analyses of users' surfing habits or data collected in another way. The underlying algorithmic systems are usually "black boxes", meaning that the pool of data used and logic behind the decisions on pricing are not comprehensible to outsiders. There is therefore a risk of price discrimination, for example relating to protected population groups within the meaning of the General Act on Equal Treatment.

The potential for harm to be caused by the implementation of higher personalised prices for individual consumers can vary greatly. Nevertheless, even small price increases for individual goods and services can, when added together, lead to significant welfare losses for the individuals and population groups affected. In particular, learning systems, which may, for example, use signalling, can also lead to quasi-collusive high market prices. If competitors deviously collude on prices or conditions via algorithms, this has a negative effect on competition, the innovative prowess of the economy and ultimately consumers; this applies both to the intentional use of algorithms to influence prices and also where parallel behaviour and high prices (tacit collusion) occur by means of learning algorithms without such specific intention and where no direct price-fixing was undertaken by humans.

It would not suffice for this overall high level of criticality to merely trigger transparency requirements and labelling obligations for pricing systems. A comprehensive impact assessment could also help to identify the discrimination risks of an algorithmic pricing system: if the pool of data being used to calculate personalised prices is known, independent experts should be able to check whether they correlate with protected population groups (known as proxies), i. e. whether, for example, women or certain religious groups have to pay higher prices. If consumers are also made aware, via labelling obligations, that prices and/or discounts are personalised, the affected parties could exercise rights of access to check the data used for "their" price for accuracy or potential discriminatory factors.

Transparency regarding price-relevant factors is also important in order to observe the steering effects of personalised pricing on the behaviour of individual consumers, as they may be of a scale which is relevant for freedom.

4.1.4 Duty to draw up documentation and keep logs

The more complex, dynamic and dispersed the process is by which individual IT systems convert an input into an output, the more important it is, from a regulatory perspective, to make the specific causes of a particular decision comprehensible. Only then can errors be detected and infringements of rights be penalised effectively. One approach to better understand how software-based processes work is to record individual program steps digitally and use them for test purposes. This may be required for personal data processing in accordance with data protection law in order to fulfil the accountability requirement.

Firstly, such a requirement to document and log the data sets and models used, the level of granularity, the retention periods and the intended purposes should be specified in data protection law so as to provide controllers and processors with greater legal clarity. Secondly, systems which have a significant potential for harm (Level 4) should be required to document and log program processes. The data sets and models used should be described in such a way that they are comprehensible to the supervisory institutions carrying out oversight measures (as regards the origin of the data sets or the way in which they are prepared, for example, or the optimisation goals pursued using the models).

4.2 Other requirements for algorithmic systems

4.2.1 General quality requirements for algorithmic systems

System operators should be required by standards to **guarantee a minimum level of quality, from both a technical and a mathematical-procedural perspective**. The procedural criteria imposed must ensure that algorithmically derived results are obtained in a correct and lawful manner. For this purpose, quality criteria can be imposed, in particular as regards the mathematical model, specific processing methods, corrective and control mechanisms or data quality and system security. To strike a balance between the conflicting fundamental rights of the software operator and the subjects of decisions, the requirements for the validity of mathematical models and the relevance of the underlying data should become stricter as **the potential of algorithmic systems to cause harm increases**.

In the case of algorithm-based and algorithm-driven decisions, **skill sensitivity** should also be built into the **design**, for example by deliberately mandating the completion of certain **training modules**. In situations where decision assistants are used, for example, it has proven particularly helpful to introduce system-imposed **role changes** at certain intervals, or in other words to assign the user the task of making the initial decision before he or she sees the algorithmically derived proposal. **Attention tests** are another option, albeit one which the individual user may perceive as more onerous; they require him or her to detect incorrect decisions which the computer has deliberately interspersed among correct ones – and therefore also require the true nature of the proposals in question to be identified in good time before anyone suffers harm.

Steps should also be taken to ensure that improvement processes are carried out fairly and with regard to the interests of everyone affected; particular attention should be paid to ensuring that suitable **feedback loops** take the interests of the data subjects and not just of the system operators into account. With regard to data quality, it would also be advisable to specify the extent to which the use of estimated or "proxy" data (→ see Part C, section 2.2.2 et seq. above) should be permitted or forbidden for certain areas of application.

In addition to the requirements placed on the algorithmic system by the actual processing purpose, the **security** requirements should also be fulfilled at the design stage. The individual requirements of all parties involved should be taken into consideration in order to ensure that appropriate design-related decisions are taken as part of conceptualisation, implementation and operation. Although the system operator usually has the main responsibility for the risk assessment, the system operator can fulfil this responsibility only with access to sufficient documentation, e.g. the manufacturer's risk impact assessment. There also needs to be clarity as to who is responsible for which area. For areas identified as critical, the Data Ethics Commission recommends setting out legal specifications relating to

- minimum standards for the required security and the measures to be taken;
- specific details regarding how and under what conditions manufacturers or system operators must design and conduct test procedures (for example to identify bias and/or discriminatory distortion);
- legal consequences in the case of security gaps or other errors;
- duties to draw up documentation on functionality and on tests which users receive in order to be able to assess risks;
- obligations to carry out system updates within a specified time frame and to report on them.

4.2.2 Special protective measures in the use of algorithmic systems in the context of human decision-making

Humans must not become an object of technology. This key principle for the regulation of algorithmic systems is particularly pertinent where algorithmic systems are used in order to support human decisions or automate decision-making processes, i.e. replace human decision-making with technical processes.

Article 22 GDPR codifies this principle in applicable existing law for certain algorithmic systems which fall within the scope of the GDPR: no one can be subject to a decision based solely on automated processing, including profiling, which produces legal or other significant effects concerning him or her – unless it is necessary for entering into, or performance of, a contract, is based on the data subject's explicit consent or is authorised by law. Where such a fully automated decision is permitted, the data controller must implement protective measures in order to safeguard the data subject's rights and interests[6] Stricter duties to provide information and rights of access also apply.[7]

6 Cf. Article 22(3) GDPR.
7 Cf. Article 13(2)(f) GDPR, Article 14(2)(g) GDPR and Article 15(1)(h) GDPR.

The Data Ethics Commission believes that various aspects of these rules currently **require further clarification**. The duties to provide information and rights of access connected with Article 22 GDPR ("including profiling") should refer to automated **profiling as such**. Individual credit reference agencies, for example, do not consider themselves subject to these rules, claiming that they apparently simply conduct profiling, while the "decisions" are made by the companies which, for example, request a credit score. The Data Ethics Commission believes this argument does not sufficiently take the intention of the GDPR into account, as the long-term effects on the data subjects of such profiling could, firstly, be significant, and secondly, the GDPR particularly emphasises profiling. Where the data protection authorities and the courts are able to apply the applicable law to the appropriate extent by means of an interpretation based on the protective purpose of the GDPR, this is to be welcomed. However, at the same time, given how sensitive this issue is in terms of fundamental rights, the democratically legitimised legislator is called upon to further specify the legal framework conditions soon in order to create legal certainty as quickly as possible. The Data Ethics Commission recommends that the Federal Government should advocate for this as part of the evaluation of the GDPR.

Clarification and specification is also needed regarding the question as to when a decision pursuant to **Article 22 GDPR** is "based solely" on automated processing of personal data and the scope of the term "similar effect" and of the protection rights under Article 22(3) GDPR. The Data Ethics Commission recommends that the Federal Government should advocate, in the evaluation of the GDPR, for the scope of Article 22 GDPR to be fleshed out. The potential for harm caused by the algorithm-determined decision-making systems, which was the original guiding principle of Article 22 GDPR, does not, in particular, categorically differ from that of many algorithm-driven decision-making systems. In particular, the tendency of the humans involved simply to accept the recommendations of algorithmic systems and not exercise discretion plays a role.

In view of the fact that the potential for harm of algorithm-based systems varies heavily in the detail, the Data Ethics Commission does not believe that it would be appropriate to generally broaden the prohibitory principle of Article 22 GDPR. In particular, the principle of human final decision-making pursuant to Article 22(3) GDPR is not suitable for all algorithmic systems in equal measure. As such, for algorithmic systems where no "decision" is taken by the system within the meaning of the current wording of Article 22(1) GDPR, a right to having the final decision made by a human would often not be very practical and also often not desirable. Instead, the Data Ethics Commission recommends a risk-adapted regulatory regime which provides individuals with appropriate safeguards (in particular against profiling) and opportunities to defend themselves if mistakes are made or if their rights are jeopardised.

The legal notion that humans must not become a mere object of technical systems should also form a **central legislative anchor point** within the horizontal EU legal instrument of a EU-ASR (→ see section 3.3 above) on the risk-adapted regulation of algorithmic systems, which the Data Ethics Commission recommends, and within the accompanying sectoral legal instruments. These legal instruments should therefore include provisions which also set out specifications for algorithm-based decision-making systems outside of the scope of Article 22 GDPR. In so far as the new layer of regulation also covers algorithmic systems which also fall within the scope of Article 22 GDPR (which may have been modified in light of the recommendations made here), the **regulatory systems** must be precisely **synchronised**.

4.2.3 Right to appropriate algorithmic inferences?

The Data Ethics Commission believes that the processes involved in the data-based generation of **algorithmic inferences** on the supposed interests, tendencies and character traits of individuals, in particular consumers, deserve maximum social and political attention. The digital economy is awash with such inferences. They are very characteristic of many digital business models which are geared towards the detailed personalisation of certain offers or services. Many consumers appreciate the convenience of such offers and services; however, they can also lead to risks if inferences are made based on an incorrect pool of data or if results with inappropriate contents are obtained on account of the inadequacy of other system components.

In order to prevent the risks which could be posed by certain algorithmic inferences, many want to grant data subjects a legal "right to appropriate inferences".[8] That proposal sets out a comprehensive package of measures which would give each data subject an effective tool for monitoring the inferences concerning them generated by operators of algorithmic systems. In addition to a substantive right to be subject to appropriate inferences, it sets out an obligation on the part of the system operator, without having to be requested for the information, to inform the individual concerned that the inferences drawn were "appropriate" and the reasons why that is the case.

The Data Ethics Commission welcomes the debate which the proposal of such a "right to appropriate inferences" has triggered. However, it points out that such a right could affect constitutionally protected interests of operators of algorithmic systems. In the view of the Data Ethics Commission, any regulatory development of the proposal should take these protection aspects into consideration, for example by limiting the scope to systems which have a high level of criticality due to their relevance in terms of participation and fundamental rights.

4.2.4 Legal protection against discrimination

One of the main aims of the regulation of algorithm-based, algorithm-driven and algorithm-determined decision-making systems is to prevent discrimination against an individual based on a characteristic set out in Article 3(3) of the Basic Law for the Federal Republic of Germany and/or Article 21(1) of the Charter of Fundamental Rights of the European Union, as well as any objectively unjustified discrimination, and to protect the personal integrity of individuals concerned. Whilst state bodies have a direct **obligation to uphold fundamental rights** when undertaking any kind of state activity and are therefore subject to a comprehensive prohibition on discrimination, a sub-constitutional basis is required for private actors. The technical legal starting point for this is essentially the **German General Act on Equal Treatment**, also serving to incorporate according EU directives into German law, alongside general clauses in German private law, for example on unconscionable contracts.

For discrimination between private individuals to fall under the General Act on Equal Treatment, firstly the discrimination must be on the grounds of a **sensitive characteristic** (race, ethnic origin, gender, religion, disability, age or sexual orientation); secondly, the **situational scope** must be open (employment context or access to goods and services, including housing, which are available to the public).

8 Omer Tene / Jules Polonetsky: Northwestern Journal of Technology and Intellectual Property, 2013 (11:5), pp. 279 et seq.; Sandra Wachter / Brent Mittelstadt: Columbia Business Law Review, 2019 (2), p. 1, et seqq. The proposal consists of a material component and a procedural component.

In principle, the provisions of the General Act on Equal Treatment already cover discrimination by algorithmic systems in accordance with applicable law. However, not all matters susceptible to discrimination are included in the scope of the General Act on Equal Treatment, and that Act does not cover all sensitive situations where algorithmically established results trigger or facilitate discrimination (e.g. in the case of a mortgage offer based on an individual risk assessment). It is therefore worth considering to, for example, broaden the **situational scope of the General Act on Equal Treatment** to include all automated decision-making processes or additionally incorporating individual areas relating to algorithmic inferences which are particularly sensitive in terms of personality.[9] This primarily concerns areas which could have a long-lasting negative effect on a person's way of life, such as consumer contracts drawn up based on scoring or on high-risk procedures, facial recognition methods or price discrimination in certain areas of life such as healthcare. The contractual partner's general freedom of action which is equally constitutionally protected must also be properly taken into consideration.

It is also necessary to discuss whether, in the context of algorithmic systems, legislators should remove the restrictive reference to specific grounds for discrimination. The discriminatory effects of algorithmic systems only sometimes reflect bias which exists within society with regard to **classic grounds for discrimination**, for example in so far as the bias is in the training data or in the model used. This would, for example, be the case if a system which is used to select candidates was trained using the data of successful managers from the past who were overwhelmingly male. However, the potential for algorithmic systems to discriminate extends far beyond this, for example if a disadvantage is systematically associated with group attributes against which discrimination is not prohibited by law (e.g. home address in a specific district) or with correlations determined by means of pattern recognition but which are really more random. To some extent, these situations can already be managed in the form of **indirect discrimination**. In that respect, a suitable relaxation of the rules relating to the burden of proof may also possibly be required. To some extent, however, entirely new issues of fairness also arise. These concern not only the distribution of opportunities to the detriment of traditionally marginalised communities but also the exclusion of groups which have been thrown together based on more or less coincidental attributes: the specific characteristics of machine learning are creating **new grounds for discrimination** which, however, could have enormous widespread impacts on account of the fact that trained algorithms are also used in other areas of application.

9 Mario Martini, Juristenzeitung (JZ), 2017, p. 2021.

It is therefore appropriate to consider broadening protection to include every systematic and objectively unjustified type of discrimination based on a group attribute. The Data Ethics Commission recommends that the Federal Government should also **examine appropriately adjusting the General Act on Equal Treatment or alternatively anchoring protection in any future specific algorithm legislation**. A particular regulatory problem is that there is a (fundamentally ever-growing) plethora of group attributes which could lead to such algorithmic discrimination, and hence the systematic nature would be the sole criterion for differentiating between prejudices which are relevant and irrelevant in terms of discrimination law. Any corresponding regulation for substantive protection against discrimination would therefore, in any case, have to be accompanied, on the one hand, by corresponding duties of disclosure and duties to state reasons and, on the other, by various internal and external oversight mechanisms for which the new regulation would provide the substantive examination criteria. The consequences of such regulation on all the parties involved would, in any case, have to be meticulously assessed and weighed up.

Irrespective of the issue of broadening the definition of the offence, thought should be given to whether the **rules on the burden of proof** already sufficiently reflect the characteristics of algorithmic systems. Ascertaining indirect discrimination requires neither proof of any intent to discriminate nor any unambiguous proof of causality. In fact, all the injured party has to prove is a correlation between the decisions and sensitive criteria. Where algorithmic systems are used, however, this proof is generally difficult for the affected parties to provide.

The Data Ethics Commission therefore recommends that legislators should enact legislation clarifying the requirements for providing proof of discrimination by operators of algorithmic systems and lower such requirements further for affected parties as needed. For this reason, the General Act on Equal Treatment should always be considered together with **rights of access and duties to state reasons** (→ see section 4.1.2) without which the injured party would often be unable to exercise his or her rights. The protection interests of third parties and of system users affected as a result must be given sufficient consideration.

4.2.5 Preventive official licensing procedures for high-risk algorithmic systems

In the case of algorithmic systems with regular or appreciable (Level 3) or even significant potential for harm (Level 4), in addition to existing regulations, it would make sense to establish licensing procedures or preliminary checks carried out by supervisory institutions in order to prevent harm to data subjects, certain sections of the population or society as a whole.

Summary of the most important recommendations for action

Instruments

45

The Data Ethics Commission recommends the introduction of a **mandatory labelling scheme** for algorithmic systems of enhanced criticality (Level 2 upwards). A mandatory scheme of this kind would oblige operators to make it clear whether (i.e. when and to what extent) algorithmic systems are being used. Regardless of system criticality, operators should always be obliged to comply with a mandatory labelling scheme if there is a risk of confusion between human and machine that might prove problematic from an ethical point of view.

46

An individual affected by a decision should be able to exercise his or her right to "meaningful **information** about the logic involved, as well as the scope and intended consequences" of an algorithmic system (cf. GDPR) not only in respect of fully automated systems, but also in situations that involve any kind of **profiling**, regardless of whether a decision is taken on this basis later down the line. The right should also be expanded in the future to apply to the algorithm-based decisions themselves, with differing levels of access to these decisions according to system criticality. These measures may require the clarification of certain legislative provisions or a widening of regulatory scope at European level.

47

It certain cases, it may be appropriate to ask the operator of an algorithmic system to provide an **individual explanation** of the decision taken, in addition to a general explanation of the logic (procedure) and scope of the system. The main objective should be to provide individuals who are affected by a decision with comprehensible, relevant and concrete information. The Data Ethics Commission therefore welcomes the work being carried out under the banner of "Explainable AI" (efforts to improve the explainability of algorithmic systems, in particular self-learning systems), and recommends that the Federal Government should fund further research and development in this area.

48

In view of the fact that, in certain sectors, society as a whole may be affected as well as its individual members, also particular **parties who are not individually affected** by an algorithmic system should be entitled to access certain types of information about it. It is likely that rights of this kind would be granted primarily for journalistic and research purposes; in order to take due account of the operator's interests, they would need to be accompanied by adequate protective measures. The Data Ethics Commission believes that consideration should also be given to the granting of unconditional rights to access information in certain circumstances, in particular when algorithmic systems with serious potential for harm (Level 4) are used by the State.

49

It is appropriate and reasonable to impose a legal requirement for the operators of algorithmic systems with at least some potential for harm (Level 2 upwards) to produce and publish a proper **risk assessment**; an assessment of this kind should also cover the processing of non-personal data, as well as risks that do not fall under the heading of data protection. In particular, it should appraise the risks posed in respect of self-determination, privacy, bodily integrity, personal integrity, assets, ownership and discrimination. It should encompass not only the underlying data and logic of the model, but also methods for gauging the quality and fairness of the data and the model accuracy, for example the bias or the rates of (statistical) error (overall or for certain sub-groups) exhibited by a system during forecasting/category formation.

50

To provide controllers and processors with greater legal clarity, further work must be done in terms of fleshing out the requirements to **document and log** the data sets and models used, the level of granularity, the retention periods and the intended purposes. In addition, operators of sensitive applications should be obliged in future to document and log the program runs of software that may cause lasting harm. The data sets and models used should be described in such a way that they are comprehensible to the employees of supervisory institutions carrying out oversight measures (as regards the origin of the data sets or the way in which they are pre-processed, for example, or the optimisation goals pursued using the models).

51

System operators should be required by the standard-setting body to guarantee a minimum level of **quality, from both a technical and a mathematical-procedural perspective**. The procedural criteria imposed must ensure that algorithmically derived results are obtained in a correct and lawful manner. For this purpose, quality criteria could be imposed, in particular as regards corrective and control mechanisms, data quality and system security. For example, it would be appropriate to impose quality criteria on the relationship between algorithmic data processing outcomes and the data used to obtain these outcomes.

52

The Data Ethics Commission believes that a necessary first step is to clarify and flesh out in greater detail the scope and legal consequences of Article 22 GDPR in relation to the use of algorithmic systems in the context of human decision-making. As a second step, the Data Ethics Commission recommends the introduction of additional **protective mechanisms for algorithm-based and algorithm-driven decision-making systems**, since the influence of these systems in real-life settings may be almost as significant as that of algorithm-determined applications. The prohibitory principle followed to date by Article 22 GDPR should be replaced by a more flexible and risk-adapted regulatory framework that provides adequate guarantees as regards the protection of individuals (in particular where profiling is concerned) and options for these individuals to take action if mistakes are made or if their rights are jeopardised.

53

Consideration should be given to expanding the **scope of anti-discrimination legislation** to cover specific situations in which an individual is discriminated against on the basis of automated data analysis or an automated decision-making procedure. In addition, the legislator should take effective steps to prevent **discrimination on the basis of group characteristics** which do not in themselves qualify as protected characteristics under law, and where the discrimination often does not currently qualify as indirect discrimination on the basis of a protected characteristic.

54

In the case of algorithmic systems with regular or significant (Level 3) or even serious potential for harm (Level 4), it would be useful – as a supplement to the existing regulations – for these systems to be covered by **licensing procedures or preliminary checks** carried out by supervisory institutions, in the interests of preventing harm to individuals who are affected, certain sections of the population or society as a whole.

5. Institutions

The Data Ethics Commission takes the view that the burden of responsibility for the ethically justified and lawful use of algorithmic systems must be shared and rest on several sets of shoulders. The institutions and supervisory structures which currently exist are not sufficiently prepared to effectively oversee monitoring of algorithmic systems at various levels. The Data Ethics Commission therefore urges the Federal Government to expand and reorient the competences of existing supervisory institutions and structures and set up new institutions and structures where necessary.

5.1 Regulatory powers and specialist expertise

5.1.1 Distribution of supervisory tasks within the sectoral network of oversight authorities

The Data Ethics Commission recommends that the Federal Government should in principle entrust regulatory supervisory tasks and oversight powers in each case to authorities which already have **sector-specific expertise**. In the view of the Data Ethics Commission, the same should apply to matters which fall within the administrative competence of States (*Bundesländer*).

Specifically, the Data Ethics Commission believes that it would make sense to entrust oversight of the use of algorithmic systems by private parties in the sectors of the digital economy in which authorities with sector-specific responsibility already exist to those **existing authorities**. As examples, authorities such as the Federal Financial Supervisory Authority (*Bundesanstalt für Finanzdienstleistungsaufsicht*, BaFin), the Federal Network Agency (*Bundesnetzagentur*, BNetzA), the Federal Office for Information Security (*Bundesamt für Sicherheit in der Informationstechnik*, BSI) and the Federal Motor Transport Authority (*Kraftfahrtbundesamt*, KBA) come into mind. Furthermore, the Federal Cartel Office (*Bundeskartellamt*, BKartA) and the data protection supervisory authorities would have special status, as they both have horizontal responsibilities, i.e. responsibilities which span the various different sectors of the economy.

The Data Ethics Commission believes that a national and EU-level **"oversight network for critical algorithmic systems"** should be set up in order to coordinate the activities of the authorities entrusted with algorithm supervisory tasks. In particular, rules on the distribution of responsibilities within the network, the exchange of information, the organisation of administrative procedures carried out by the network and legal protection would be appropriate for such purposes.

In order to prevent any gaps in supervision, the Data Ethics Commission urges the Federation and the *Länder* to identify areas where there is **currently no sector-specific authority with sufficient expertise** to which oversight tasks could be assigned for monitoring critical algorithmic systems. In the view of the Data Ethics Commission, in such cases, it will often be appropriate, in the event of a corresponding need for oversight, to entrust matters to one of the existing authorities with horizontal responsibility. In the case of algorithmic systems which process sensitive personal data, the data protection authorities, for example, may have the adequate expertise. However, the Data Ethics Commission believes that, in particular cases, it may be necessary to create completely new regulatory control structures. In the light of ever-changing technical developments, the Federation and the *Länder* should regularly review the situation.

Authorities are faced with a structural challenge in effectively executing their algorithmic system oversight tasks: the object which is the focus of their oversight work is technically highly complex and is subject to dynamic change. The Data Ethics Commission therefore believes that **providing the authorities with practical skills** will be particularly important. It firmly recommends that the Federal Government should provide the federal authorities with the financial, human and technical resources required. The draft Salary Structure Modernisation Act (*Besoldungsstrukturenmodernisierungsgesetz*), which is expected to increase the salaries and bonuses of public-sector IT professionals and establish new regulations for them as from 2020, is without doubt a welcome first step. However, in the light of how difficult it is to attract well-trained professionals to the public sector, further measures will soon be required.

The Data Ethics Commission also recommends that the Federal Government should set up an official unit in the form of a **competence centre for algorithmic systems** to provide the sectoral authorities with support in monitoring algorithmic systems. The responsibility of such a body should not only acquire, analyse, further develop and impart the technical methodological knowledge required for supervising critical algorithmic systems. It should (in coordination with and at the request of the sector-specific authorities) also primarily support the sector-specific supervisory authorities in building up the expertise needed to carry out their tasks and assess the criticality of algorithmic systems. This will extend in particular to the centre's task of further developing **criteria, processes and tools** for the oversight of algorithmic systems. This will also include **standards for assessing criticality** and checking the compliance of critical algorithmic systems. Such a centre of competence will also have an important **intermediary advisory role**: as far as possible, it will advise not only bodies of the Federation, the *Länder* and municipalities, but also manufacturers, system operators, system users and data subjects with regard to the use and development of algorithmic systems. It will also be involved in international and European initiatives designed to build up sufficient oversight expertise including standardisation procedures. However, the competence centre should not have its own supervisory powers. These remain with the sectoral supervisory authorities. The service unit should either be created as a new, autonomous federal authority or be attached to an existing cross-sectional authority, such as the Federal Office for Information Security.

The Data Ethics Commission considers that it would also make sense to establish a corresponding body at **European Union level** in the future, for example in the form of an agency, and the Federal Government should work towards achieving this.

In principle, the Data Ethics Commission sees no reason why state bodies should not be able to make use of the **expertise of private individuals or entities** in carrying out their tasks and in building up their own in-house expertise or to involve private individuals or entities in the execution of their tasks, as long as such cooperation complies with the general constitutional and administrative specifications applicable to such cooperation. Conversely, corresponding cooperation, for example also by entrustment, may be used in order to deal with the current lack of qualified specialists and expertise in the public sector.

5.1.2 Definition of oversight powers according to the tasks involved

The regulating body should, **by law**, clearly **assign** the relevant competent authorities the **powers of intervention**, including rights to information and rights of inspection and access, required for the supervision of algorithmic systems. Blueprints for such regulatory powers for content control can be found in various areas of the law.[10]

The competent supervisory authorities must, at all times, be able to **examine** algorithmic systems in sensitive areas of application or those with a high potential for harm. The audit and test procedures used in doing so must, in particular, cover systems where there is interaction with the user. This may, for example, take place via standardised interfaces. Such access can be used to carry out what are known as input-output tests, which check, for example, whether an algorithmic system systematically discriminates against groups. This is particularly useful in the case of learning systems which adapt their internal rules over time. Steps must be taken here to ensure that any testing of learning systems does not lead to a change in the system of rules whereby the system learns from the test data during the test.

10 For example, Article 58 GDPR governs the investigative powers relating to data protection supervision and Section 32e of the [German] Act against Restraints of Competition (*Gesetz gegen Wettbewerbsbeschränkungen*, GWB) governs sector inquiries by the Federal Cartel Office. Oversight of high-frequency trade by financial supervisory authorities is based on Section 6(4) of the [German] Securities Trading Act (*Gesetz über den Wertpapierhandel*, WpHG) and Section 3(4)(4)(5) of the [German] Stock Exchange Act (*Börsengesetz*, BörsG) amended version in conjunction with Section 7(3) of the Stock Exchange Act.

When assigning legal authority, steps must be taken to ensure that the supervisory authorities have the power, in the event of a proven breach of the law, to force operators of algorithmic systems to configure systems in compliance with the law (for example by adapting the pool of data used) and, where necessary, apply **penalties**. Provided that it is commensurate with the case in question, the supervisory authorities should also be able to impose official **bans** on the use of unlawful algorithmic systems (or their components).

5.1.3 Criticality-adapted extent of oversight

All elements of an algorithmic system must be taken into account in order for its behaviour to be effectively audited. An audit conducted by authorities may, and potentially must, extend to the training data and learning processes used, the final rule-based model as well as the input and output data underlying the decisions. Quality indicators regarding the pool of data used and model accuracy (training model, final decision model) can also be taken into consideration in order to identify a system's bias or rates of (statistical) error (overall or for certain sub-groups). From a methodological perspective, a test may be carried out by analysing large amounts of data, reviewing the weighting of factors in complex multidimensional systems and analysing input-throughput-output.

Due to the complex nature of the subject matter and amounts of data involved, the use of control algorithms can significantly increase the efficiency and effectiveness of the audit. They can systematically look for conspicuous patterns in the pool of data used and the results of an algorithmic system which can, for example, shed light on a case of discrimination.

The extent of oversight required in each specific case should be determined based on the area of application and system criticality. In the case of systems which have only some potential for harm (Level 2), it may suffice for legislators to limit regulatory oversight to an inspection of the results in the event of a system's documented failure. However, in areas with a high potential for harm, it may be necessary to stipulate that system operators must use a standardised interface.

In the view of the Data Ethics Commission, the question as to whether regulatory oversight affects system operators' trade and **business secrets** or third parties' **privacy rights** is not an issue at any level of the criticality pyramid. As supervisory authorities are obliged to treat all information obtained as part of their oversight work as confidential due to professional secrecy, these aspects do not represent any legal obstacle to far-reaching powers for full and detailed audits.

The proper interpretation of test results is, from a technical perspective, anything but trivial. In particular, it is not always clear whether they really unearth an error by an algorithmic system. This restricts their ability to provide evidence. The quality and informative value of the different test procedures and audits therefore also need to be agreed on – in particular with regard to their probative value in court proceedings in order to enforce the rights of the parties affected. The Data Ethics Commission therefore recommends that the Federal Government should support initiatives to **develop statistical technical standards for test procedures and audits**, where necessary differentiated by areas of application. The competence centre for algorithmic systems (→ see section 5.1.1) should take a leading role in such endeavours.

Use case: Personalised prices II – ex-post controls by supervisory institutions

Supervisory institutions could check whether algorithmic pricing systems used in e-commerce comply with the law or discriminate, for example, against protected population groups (within the meaning of the General Act on Equal Treatment). Supervisory authorities could look for conspicuous patterns in the pool of data used and the issued prices, which may shed light on a possible case of discrimination.

To do so, those carrying out the supervision do not have to comprehend the (potentially highly complex) rules of the underlying algorithm by analysing the code. Effective oversight can be carried out with the help of statistical tests which analyse how, all other things being equal, issued prices change depending on input data which are associated with certain population groups. If, for example, the system issues higher prices for consumers when only the gender is changed from "male" to "female" in the input data or if the issued prices correlate with attributes, protected under equality legislation, of individual population groups (for example via proxies), this can be mathematically statistically determined.[11]

11 Cf. Gesellschaft für Informatik: Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. Gutachten der Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V. im Auftrag des Sachverständigenrats für Verbraucherfragen [Gesellschaft für Informatik: Technical and legal considerations regarding algorithmic decision-making processes. Report by the Legal Informatics expert group of Gesellschaft für Informatik e.V. at the request of the Advisory Council for Consumer Affairs], Berlin, pp. 63 et seqq. (available at: *www.svr-verbraucherfragen.de/wp-content/uploads/GI_Studie_Algorithmenregulierung.pdf*).

5.2 Corporate self-regulation and co-regulation

It is neither possible nor necessary for the legislator to implement blanket regulations covering all algorithmic systems. Instead, various models of self-regulation and co-regulation could also essentially provide sufficient responses for certain situations. Co-regulation involves regulatory approaches which navigate between state regulation and private self-regulation and is characterised by the combination of a public/state component and a private/institutional component.

5.2.1 Self-regulation and self-certification

The Data Ethics Commission recommends self-regulation in the form of an internal audit conducted by the manufacturer or operator of the algorithmic system for the lowest level of the criticality pyramid. This could be supported by self-certification of manufacturers and operators on the basis of specific standards for algorithmic systems. The particular advantage of such a system would be that the self-certification bodies would have the necessary know-how on account of their **close connection to the specific topics**. As a result, experts, even from the companies in question, could take the legal standards and monitoring of compliance therewith into consideration, including at the development stage, and, where necessary, also incorporate their corporate expertise into the regulatory mechanisms institutionally. Admittedly, purely internal and voluntary self-regulation would not constitute an independent monitoring and, in the event of breaches, would not ensure any effective implementation of penalties.

The self-regulation architecture could be supplemented with a model involving regulated self-monitoring, which would set out external standards for quality and risk management of self-monitoring which could also be externally monitored. A similar system is set out in the GDPR, in which Article 40 establishes the option to specify general clauses of the GDPR and make them applicable to specific real-life circumstances which are significant to the parties subject to the codes of conduct as well as set minimum standards specific to the sector in question. In order to be able to guarantee that the regulation would be as effective as intended, effective monitoring must ensure actual compliance with the approved codes of conduct pursuant to Article 40 GDPR. Not only would the codes of conduct themselves have to be drawn up, but the procedural rules relating to monitoring, control and the implementation of penalties for cases of non-compliance would also have to be set out.

Where a provider signs up for voluntary self-monitoring and verifiably demonstrates compliance with the agreed procedures, the standard-setting body may grant privileges in terms of supervisory measures. Such an approach would be based on the condition that, in exercising their corporate responsibility and in cooperation with a private self-monitoring body, providers would have to develop procedural standards which would be recognised by the supervisory authority. The involvement of civil society organisations in the preparatory work would be essential in order to be able properly to represent the interests of citizens and consumers and take them into consideration.

5.2.2 Creation of a code of conduct

For the concept of regulated self-regulation, it would be worth considering including an **Algorithmic Accountability Code**, adopting a "comply or explain" approach which is well-established in other parts of the legal system. It could oblige parties subject to regulation to state whether or not and the extent to which they are following the recommendations of the code.[12] False statements would be subject to sanctions. As such, a code to be drawn up could be binding in nature by holding companies and authorities responsible for the consequences of the use of algorithmic systems. It could, for example, be developed based on corporate digital responsibility guidelines (→ see Part D, section 2 above) or conversely also help to shape such guidelines. What level of granular detail for codes and guidelines is practical and/or the sector-specific ethical challenges for which a specific code would be useful will become clear.

The quality of the defined requirements and the framework conditions, i. e. the opportunities for independent external parties to carry out checks and the ability to impose penalties in the event of breaches, would be essential in ensuring that a code had a control function. Responsibility for developing such a code should be assigned to an independent commission with equal representation of manufacturers, operators, the scientific community and civil society. It remains to be seen whether the Government Commission on the German Corporate Governance Code *(Regierungskommission Deutscher Corporate Governance Kodex)* (*www.dcgk.de*) could be a model for this.

In addition or alternatively, binding statements by and between manufacturers and operators of algorithmic systems could be considered.

12 Mario Martini, Juristenzeitung (JZ), 2017, p. 1022 et seq.

5.2.3 Quality seals for algorithmic systems

Establishing quality seals for algorithmic systems are sensible in order to support effective algorithm regulation. They could take the form of voluntary or mandatory evidence of protective measures which would make the extent to which an algorithmic system meets certain requirements clear to users. It would be important to clarify who would define the requirements of a quality seal and who would specifically be responsible for fulfilling the requirements connected with the quality seal and the extent to which breaches would be subject to penalties. As in the case of an Algorithmic Accountability Code, responsibility for defining the requirements of a quality seal should be entrusted to an independent commission with equal representation of operators of algorithmic systems, the scientific community and civil society.

5.2.4 Contact persons for algorithmic systems in companies and authorities

Companies and authorities which work with critical algorithmic systems (as from Level 2) should (at least starting at a certain size of company or authority) appoint a contact person responsible for communications with authorities and cooperation. In all cases, it must be ensured that such a contact person has **specific expertise**. He or she will monitor the use of algorithmic systems internally and provide the company's or authority's management team with advice and will be functionally independent. As is the case with data protection officers, the contact person could act as a link between the supervisory authority, operators of algorithmic systems and affected groups of people. This would also help to ensure proper awareness of problems within companies and authorities and increase oversight pressure from inside.

5.2.5 Involvement of civil society stakeholders

In order to ensure that the interests of civil society and affected companies are properly taken into account as part of audits of algorithmic systems, **advisory boards** should be set up within sector-specific competent authorities, and civil society stakeholders should also, for example, be involved in connection with a code. Such advisory boards should feature a balance of representatives of civil society organisations and individuals appointed by companies in order to ensure that both the interests of affected individuals and groups and the interests of affected companies are properly taken into account as part of audits.

5.3 Technical standardisation

In the view of the Data Ethics Commission, standardisation organisations such as ISO/IEC, IEEE, IETF, ITU, ETSI, W3C, CEN and DIN, which set technical standards for information and communications technologies, could significantly help with forming out sector-specific requirements for algorithmic systems. Technical standards which take ethical and legal requirements into consideration could provide legal certainty for companies which develop and use algorithmic systems. They could also easily translate the requirements for the legality of algorithmic systems into specific guidelines in individual sectors.

The Data Ethics Commission believes that technical standards would essentially be useful tools to bridge the gap between "classic" state regulation and purely private self-regulation. It therefore recommends that the Federal Government should suitably work to develop and adopt technical standards designed to prevent the risks posed by algorithmic systems.

However, in the view of the Data Ethics Commission, the Federal Government should also not lose sight of the fact that **technical standards have their limitations** (→ see Part D, section 6 above). Technical standards are no substitute for defining clear legal requirements for algorithmic systems or for regulatory supervision of the use of such systems. For constitutional reasons, the principle that the more citizens' fundamental rights are affected, the more detailed legal provisions should be, must be upheld. In practice, this means that legislators must, first of all, define the legal framework – not technical standard-setting committees. This will not least ensure that the integrity of decision-making will be protected, as the active participation of representatives of sectors and/or affected companies will ensure that, in addition to impressive technical expertise, the interests of such companies and/or sectors are, of course, also often taken into consideration first hand when the technical standards are drawn up.

5.4 Institutional legal protection (in particular rights of associations to file an action)

The system of granting competitors, competition associations and consumer associations the right to file an action has been an important feature of the German legal landscape for many years, and could play a key role in **civil society oversight** of the use of algorithmic systems. In particular, private rights of this kind allow civil society players with a legitimate mandate to enforce compliance with legislative provisions in the area of contract law and fair trading law without needing to rely on the authorities to take action and without needing to wait for individuals to authorise them. This civil law approach has particularly strong market focus and is characterised by swift responses and is therefore, by international standards, very successful. Associations are essentially politically and administratively independent and can therefore advocate, on their own authority and in the common interest of consumers and companies, for competition regulations and consumer rights to be efficiently protected against unfair business practices which are also damaging for consumers.

Anyone who does not comply with regulatory provisions will potentially benefit from an unfair competitive advantage. In order to prevent any competitive edge being gained by breaking the law, competition associations and consumer associations should be able to stop such legal infringements.

Summary of the most important recommendations for action

Institutions

55

The Data Ethics Commission recommends that the Federal Government should expand and realign the competencies of existing supervisory institutions and structures and, where necessary, set up new ones. Official supervisory tasks and powers should primarily be entrusted to the **sectoral supervisory authorities** that have already built up a wealth of expert knowledge in the relevant sector. Ensuring that the competent authorities have the financial, human and technical **resources** they need is a particularly important factor in this respect.

56

The Data Ethics Commission also recommends that the Federal Government should set up a **national centre of competence for algorithmic systems**; this centre should act as a repository of technical and regulatory expertise and assist the sectoral supervisory authorities in their task of monitoring algorithmic systems to ensure compliance with the law.

57

The Data Ethics Commission believes that initiatives involving the development of technical and statistical **quality standards for test procedures and audits** (differentiated according to critical application areas if necessary) are worthy of support. Test procedures of this kind – provided that they are designed to be adequately meaningful, reliable and secure – may make a vital contribution to the future auditability of algorithmic systems.

58

In the opinion of the Data Ethics Commission, particular attention should be paid to innovative forms of **co-regulation and self-regulation**, alongside and as a complement to forms of state regulation. It recommends that the Federal Government should examine various models of co-regulation and self-regulation as a potentially useful solution in certain situations.

59

The Data Ethics Commission believes that an option worth considering might be to require operators by law (inspired by the “comply or explain” regulatory model) to sign a declaration confirming their willingness to comply with an **Algorithmic Accountability Code**. An independent commission with equal representation – which must be free of state influence – could be set up to develop a code of this kind, which would apply on a binding basis to the operators of algorithmic systems. Appropriate involvement of civil society representatives in the drafting of this code must be guaranteed.

60

Voluntary or mandatory evidence of protective measures in the form of a specific **quality seal** may also serve as a guarantee to consumers that the algorithmic system in question is reliable, while at the same time providing an incentive for developers and operators to develop and use reliable systems.

61

The Data Ethics Commission takes the view that companies and authorities operating critical algorithmic systems should be obliged in future to appoint a **contact person**, in the same way that companies of a specific size are currently obliged to appoint a data protection officer. Communications with the authorities should be routed through this contact person, and he or she should also be subject to a duty of cooperation.

62

To ensure that official audits of algorithmic systems take due account of the interests of civil society and any companies affected, suitable **advisory boards should be set up within the sectoral supervisory authorities**.

63

In the opinion of the Data Ethics Commission, technical standards adopted by **accredited standardisation organisations** are a generally useful measure, occupying an intermediate position between state regulation and purely private self-regulation. It therefore recommends that the Federal Government should engage in appropriate efforts towards the development and adoption of such standards.

64

The system of granting **competitors, competition associations or consumer associations the right to file an action** has been an important feature of the German legal landscape for many years, and could play a key role in civil society oversight of the use of algorithmic systems. In particular, private rights of this kind could allow civil society players with a legitimate mandate to enforce compliance with legal provisions in the area of contract law, fair trading law or anti-discrimination law, without needing to rely on the authorities to take action and without needing to wait for individuals to authorise them.

6. Special topic: algorithmic systems used by media intermediaries

6.1 Relevance for the democratic process: the example of social networks

For many people, it would be impossible to imagine life these days without social networks, search engines and the like: they enable users to keep up to date on the latest news from around the world and from their circle of friends in real time, are platforms through which people can portray their lifestyles and communicate with each other, and can also be used for entertainment purposes and for business activity, including advertising.

On the whole, they are becoming increasingly important for private and public opinion-forming. In order to manage the wealth of information available, providers of such services use algorithmic systems which are designed, amongst other things, to identify the interests, tendencies and convictions of users, identify posts which are of potential relevance to them, present them with similar posts in order to encourage them to interact with the network, and filter out illegal or offensive posts. The economic aim is primarily to generate high advertising revenue.

Depending on their reach and content, media intermediaries can have a profound impact on the democratic process. More and more people are also using social networks to keep abreast of politics and world affairs. Social networks therefore offer users new opportunities to participate in the information society and, in that sense, constitute **media and factors for the exchange of information and opinions**.

At the same time, the fact that public debate is concentrated on only a few private platforms also poses a challenge for democracy. After all, as economic stakeholders, private operators of social networks have a vested interest in directing traffic to their networks and gearing activity on them primarily towards economic aspects rather than focusing on social interests in having a multi-faceted opinion-forming process for the benefit of the public good. The use of algorithmic systems which are **predominantly oriented on economic criteria** can have negative consequences for the diversity of opinions on social networks.

The use of services can also lead to the manipulation of opinions. On the one hand, this can happen unintentionally due to certain characteristics of underlying software, such as for example recommender systems. On the other hand, these systems can be used intentionally by various actors for manipulative purposes. Up to now, operators of social networks have not sufficiently guarded against such activities which threaten the foundations of democracy. What is more, a regulatory framework and social oversight are needed, in particular in view of their **high level of criticality**.

The Data Ethics Commission believes that, in the future, media intermediaries which have a gatekeeper role can ultimately develop a high potential for harm to our democracy and that there is a resulting **need for regulation**. The Data Ethics Commission believes that it is essential for legislators to create an appropriate regulatory framework for the use of algorithmic systems by media intermediaries. The Data Ethics Commission is of the opinion that, first of all, the operators of such platforms and providers of such services should themselves define and implement basic rules to ensure fairness in the opinion-forming process. However, this "digital domiciliary right" has its limitations, in particular where the integrity of the democratic process is affected. Depending on the market share and gatekeeper role of such platforms and services, operators have fundamental-rights-based obligations on account of the indirect third-party effect.[13] In the view of the Data Ethics Commission, these obligations should be specified more precisely in sub-constitutional law, in particular also with regard to the use of algorithmic systems by and on platforms and by and in services with a significant market share and a gatekeeper role. This is also relevant for the EU-ASR recommended by the Data Ethics Commission (→ see section 3.3 above).

Regulation is also needed to ensure regulatory fairness in comparison to broadcasters. The Data Ethics Commission recommends that the Federal Government should examine how risks posed by providers which have a particular power to influence opinions can be countered. A whole range of measures are possible, from greater transparency right through to ex-ante controls in the form of a licensing procedure for algorithmic systems which are relevant in terms of democracy.

6.2 Diversity and media intermediaries: the example of social networks

The wide variety of roles played by social networks and the predominantly high level of criticality of the algorithmic systems they use present particular challenges for the Data Ethics Commission's suggested approach of risk-adapted regulation of algorithmic systems. The Data Ethics Commission believes that positive legal provisions for social networks which, for example, increase the **transparency and range of discussions held there** and bolster the **rights of users** would be particularly constructive.

In any case, where social networks have dominant market share, the Data Ethics Commission calls for further measures to **safeguard diversity**, as defensive measures alone will not suffice. Algorithmic systems which operate in these types of networks and have impacts on the freedom and diversity of opinion-forming which are constitutive of democracy have an extremely high level of criticality on account of their reach alone. The Data Ethics Commission believes that legislators are therefore under an ethical and constitutional obligation to establish a **binding normative framework** for the regulation of media intermediaries in order to protect democracy. This may require transforming the regulatory framework governing the media.

Legislators must take suitable measures to ensure that the total range on offer reflects the variety of opinions that exist and guarantees **balance, neutrality and freedom from bias in the information society**.[14] This applies in particular to media intermediaries with a gatekeeper role and power to influence opinions. According to the Federal Constitutional Court, to safeguard pluralistic diversity, substantive, organisational and procedural regulations are needed which are focused on creating freedom of communication and are therefore suitable for producing the desired effects of Article 5(1) of the Basic Law for the Federal Republic of Germany.

13 Decisions of the Federal Constitutional Court 128, p. 249 (FRAPORT); 148, p. 267 et seqq., margin no. 32 et seqq. (Stadionverbot).
14 Cf. decisions of the Federal Constitutional Court 136, 9, 28 with further references.

In the light of this, the legislators in the *Länder*, which are responsible for media law, are obliged to implement the aforementioned provisions. The same applies to the legislators of an EU Regulation on Algorithmic Systems (EU-ASR (→ see section 3.3 above). As media intermediaries, video-sharing platforms (VSPs) are already subject to the Audiovisual Media Services Directive[15] because they provide user-generated content for the general public. The draft Interstate Media Services Agreement also covers media intermediaries. The Data Ethics Commission once again welcomes, in that respect, the provisions for the transparency of social networks set out in the draft **Interstate Media Agreement (Medienstaatsvertrag, MStV-E)** as an initial step in this direction.

The **legislators for the *Länder*** have plenty of scope and freedom for drawing up the provisions. However, they must decide on the regulation model themselves and must not leave it to private individuals to agree on. The Data Ethics Commission is of the view that plurality obligations for media intermediaries should, in any case, include the obligation to use algorithmic systems which, at least as an additional option, also provide access to an unbiased and balanced selection of posts and information which reflect a diverse range of different opinions.[16]

Based on these considerations, the Data Ethics Commission also recommends that the Federal Government should investigate whether there are other areas where, irrespective of the situation relevant to democracy discussed here, a corresponding obligation to establish requirements for neutrality and provisions on diversity seems necessary. **Protecting minors** from being influenced by and on social networks, for example, is one such consideration.

6.3 Labelling obligation for social bots

The democratic process is, in essence, based on people's freedom to form their own opinions and make their own decisions. However, bots, i. e. software programs which **give the impression that they are human users**, are used on various platforms. In the view of the Data Ethics Commission, it is highly problematic if such bots are used to manipulate individual users and/or public debate or guide the result of a vote one way or the other where political decisions are to be made. Firstly, the simulation of human traits falsely suggests that the statements made are the result of independent thought and of the independent formation of political opinions. Secondly, automation can massively increase the number and frequency of expressions of opinion, making it harder or even impossible to assess actual majorities of opinions. The Data Ethics Commission believes that regulatory intervention is required here.

On that basis, the Data Ethics Commission recommends implementing a **measure to enhance transparency** in the form of a labelling obligation for social bots on social networks. Based on general considerations, the Data Ethics Commission recommends that such a labelling obligation should be implemented anywhere where there is a risk that social bots could be mistaken for human interlocutors (→ see section 4.1.1 above). Given the particular potential to jeopardise the democratic process, the Data Ethics Commission furthermore believes, in any case, that a labelling obligation for social bots which have an impact on political opinion-forming processes is essential, even irrespective of any real risk of confusion.

15 Directive 2010/13/EU of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive).
16 Rolf Schwartmann / Maximilian Hermann / Robin Mühlenbeck, Multimedia und Recht (MMR), 2019 (8), p. 498 et seqq.

6.4 Measures to combat fake news

A labelling obligation for social bots could help to combat the automated spread of fake news. However, the Data Ethics Commission also believes that the concept of fake news is **not suitable as a starting point for any regulation relating to media legislation**. The presentation of a legal definition of fake news, which draws an objective and distinct line between an exaggerated or satirical expression of opinion and an intentional misrepresentation of news is impossible due to the complexity of human communications. Disinformation and the manipulation of public opinion-forming, typically associated with the term "fake news", can also result from true facts being presented selectively.

The Data Ethics Commission also, in particular, recommends to legislators that operators of social networks should grant their users an easy-to-exercise **right of reply** requiring the network to post the correction of a statement proven to be false (e. g. an invented quote) on the timeline or newsfeed, etc. of all users whom the network, using available data, can trace back to have been shown the false statement.

The Data Ethics Commission emphasises that the State must not create any incentives for collateral censorship through social networks. To provide protection from "overblocking", the Data Ethics Commission therefore believes it is necessary, in parallel to the obligations imposed on the operators, to grant the affected individuals prompt and efficient procedural protection mechanisms. The Data Ethics Commission believes that these should include in particular a **right to an effective process to reinstate deleted posts** provided that they do not break any laws; any invocation by networks of their own rules alone cannot suffice as grounds for permanent deletion/blocking. In the view of the Data Ethics Commission, such rights must apply to users with respect to all social networks.

6.5 Transparency obligations for news aggregators

Where social networks use algorithmic systems which also aggregate, select and present journalistic/editorial content of third parties in a generally accessible way, they should have to allow users and interested third parties enough insight into the technical procedure they use to select and prioritise news to make clear how a recommendation is arrived at in an individual case. The democratic information interest would essentially take precedence over any business secrets of media intermediaries. In the interests of a fair opinion-forming process and a fair exchange of opinions, such duties to disclose information should also stretch to any economic ties. For that reason as well, the Data Ethics Commission welcomes the current thoughts on reforming the Interstate Media Agreement (Medienstaatsvertrag, MStV-E) which call for corresponding transparency obligations for media intermediaries as soon as they have a certain reach.

Summary of the most important recommendations for action

Special topic: algorithmic systems used by media intermediaries

65

Given the specific risks posed by media intermediaries that act as **gatekeepers to democracy**, the Data Ethics Commission recommends that options should be examined for countering these risks, also with regard to influencing EU legislation (→ see Recommendation 43 above). A whole gamut of risk mitigation measures should be considered, extending through to ex-ante controls (e.g. in the form of a licensing procedure).

66

The national legislator is under a constitutional obligation to protect the democratic system from the dangers to the free, democratic and pluralistic formation of opinions that may be created by providers that act as gatekeepers by establishing a binding normative framework for **media**. The Data Ethics Commission believes that the small number of operators concerned should be obliged to use algorithmic systems that allow users (at least as an additional option) to access an unbiased and balanced selection of posts and information that embodies pluralism of opinion.

67

The Federal Government should consider measures that take due account of the risks typically encountered in the media sector in respect of all media intermediaries and also in respect of providers that do not act as gatekeepers or whose systems are associated with a lower potential for harm. These measures might include mechanisms for **enhancing transparency** (for example by ensuring that information is available about the technical procedures used to select and rank news stories, **introducing labelling obligations for social bots**) and establishing a right to post countering responses on timelines.

7. Use of algorithmic systems by state bodies

7.1 Opportunities and risks involved in the use of algorithmic systems by state bodies

Citizens will rightly expect their State to **use the best technology available** to carry out its duties. Depending on the type of duties, this will also include algorithmic systems. Systems already exist which can relieve state bodies of repetitive tasks (thereby expediting processes and freeing up human resources for complex cases) and which, in certain set-ups, improve the consistency and quality of state activity or, in the form of chatbots or voice assistants, for example, can facilitate citizens' access to justice.

At the same time, when using algorithmic systems, state bodies must uphold particularly high standards: firstly, they have a direct obligation to uphold fundamental rights as public authorities and secondly, state activity is, in general, expected to **set an example** for the whole of society. The institutional capacity and expertise, which the State must build up in order to ensure sufficient oversight of algorithmic systems used by private parties, must therefore also be used in order to guide and oversee the work carried out by state bodies themselves. In particular, the competence centre for algorithmic systems called for by the Data Ethics Commission is likely to play a key role in this context.

The use of algorithmic systems by state bodies must be treated **in principle as particularly sensitive** within the meaning of the criticality pyramid (at least Level 3). Therefore, in the view of the Data Ethics Commission, a comprehensive risk impact assessment must be carried out as a mandatory requirement for any ethically sound use of algorithmic systems. Furthermore, depending on the criticality of the systems used by the State, where necessary, other instruments discussed above and designed to ensure that citizens are protected should be put in place for such algorithmic systems used by the State. Farther-reaching legal data protection requirements will remain unaffected, as will other constitutional and administrative specifications for the design of the systems. Additionally, in the view of the Data Ethics Commission, in certain sectors where the use of algorithmic systems conflicts with constitutionally protected rights of overriding importance, the use of algorithmic systems should, irrespective of the protective measures taken in the case in question, be permitted only under very restrictive conditions or prohibited. This in particular concerns the use of algorithmic systems for the purposes of law-making and jurisprudence.

7.2 Algorithmic systems in law-making

The use of algorithmic systems within the government context of law-making is subject to restrictions. The Data Ethics Commission sees the democratic process, in the sense of people being able to form their own opinions and make their own decisions as freely as possible, as essentially sacrosanct. Automated support in law-making is therefore acceptable **at most for low-level ancillary tasks** (e.g. detecting inconsistent use of terms) and/or **legal instruments which are far removed from the democratic decision-making process** (e.g. catalogues of technical specifications in subsequent regulations). In both cases, the systems must meet extremely strict requirements for quality and security.

In this context, the Data Ethics Commission also, in particular, opposes any demand that newly enacted legal instruments should already be formulated with a view to possible future automated application; **in that regard, technology must follow the law and not the reverse**. Only if, in accordance with conventional criteria for the assessment of legislation (compliance with fundamental rights and other higher-ranking law, impact assessment, etc.), two equivalent versions are conceivable may the argument that one version is easier to algorithmise tip the scales in its favour.

7.3 Algorithmic systems in the dispensation of justice

The Data Ethics Commission is of the view that the use of algorithmic systems in the dispensation of justice is permissible **only for peripheral tasks**. Justice is administered "in the name of the people", and that means, at least in contentious proceedings as well as in administrative court proceedings and criminal proceedings, always administered by human judges. The pacification effect of court proceedings is achieved not only through the judgment itself (fairness of the finding) but also through the hearing and weighing up of conflicting interests by humans and, in particular, the structural processing of the facts and legal consequences (procedural fairness), in contrast to an opaque black-box decision.

Due to the often high level of trust placed in the supposed "infallibility" of technical systems (automation bias) as well as the low level of willingness to make divergent decisions, in particular if this is associated with an additional burden of reasoning and proof and the risk of a "miscarriage of justice" (default effects), **even legally non-binding proposals for decisions** for judgments by algorithmic systems are generally **highly problematic** from the perspective of the parties concerned.

However, algorithmic systems can, provided that there are strict quality control and high security standards in place, be useful for **preparatory work** which does not directly affect the judicial decision (e.g. file management and document control).

Lastly, the use of systems which **retrospectively analyse judicial decisions**, are available only for voluntary use by judges and are protected against access by third parties with high-level security measures, is also conceivable. Such systems could, for example, work out whether decisions were influenced by external factors and, if so, which ones in order to provide judges in future with ways to prevent such distortions themselves and thus contribute to better and more consistent dispensation of justice. Researchers may also have a legitimate interest in access to such systems, though sufficient safeguards would be required here in individual cases. The use of systems for the purpose of monitoring the path of judicial decision-making or of checking the dispensation work of judges against external targets (e.g. average processing time for a case) is, however, in view of objective judicial independence, not permissible.

In the **pre-litigation domain** (for example, exercising of air passenger rights) or also in a dunning procedure or similar, in the view of the Data Ethics Commission, fully automated handling of legal claims is permissible provided that procedural rights of the individual parties concerned are safeguarded as a result. However, this is not the case if algorithmic systems create correlations which do not follow the legal provisions and procedural steps set out. With the current state of the art, only systems based on classic deterministic algorithms can therefore generally be considered which, for example, make decisions by meeting formal criteria (which are not open to assessment). From a systemic point of view, impending losses of expertise are compensated for here by the freeing up of human resources for complex individual cases.

7.4 Algorithmic systems in public administration

There is a potentially greater scope for the use of algorithmic systems in public administration. The increased **automation of authorities' routine cases**, which can be included subject to precisely defined conditions regarding facts and legal consequences, may be advisable in the interest of efficiency (Section 10(2) of the Administrative Procedures Act) in order to carry out administrative procedures as appropriately and swiftly as possible. Here in particular, it relieving administrative staff of routine tasks frees up human resources which can then be deployed to handle procedures which cannot be automated.

There is potential, in particular, in the **provision of services and benefits**. The Data Ethics Commission believes that algorithmic systems can and should be used here to expand proactive procedure management whereby, where all the required data are available for the authorities, services and benefits will be increasingly provided without the need for applications. Educationally disadvantaged individuals and the needy in particular could benefit from this (cf. family allowance in Austria provided when a child is born without the need to apply for it).

However, in the case of **intervention by authorities**, the use of algorithmic systems must be dealt with carefully because fundamental rights are particularly affected. As with judicial use, this applies not only to algorithm-determined administrative decisions but also where the use of the systems limits the authorities' scope for decision-making. In general, in assessing whether to permit the use of the systems, the extent of the resulting intervention and the reversibility of the decisions need to be taken into consideration. Essentially, in designing the systems, technology must be used which is most easily accessible to oversight. Therefore, in sensitive areas, public administration will often be allowed to use only systems which are based on classic deterministic algorithms. The use of proprietary software should be avoided for the same reason.

In the case of **discretionary decisions** by the executive and decisions with a margin of discretion which have an external legal effect, the Data Ethics Commission believes that it is currently necessary for humans to make the final decision where the decision has more than mere beneficial impacts. However, by forming groups of cases and through further specification, it is conceivable that discretion could be reduced to such an extent that, from the view of the algorithmic system, there is only one option in terms of the decision. The Data Ethics Commission is of the view that Section 35a of the German Administrative Procedures Act does not sufficiently reproduce the range of different possible types of cases and is too schematic. Taking into account the safeguards required by constitutional law and based on Article 22 GDPR, legislators should **carefully expand the scope of Section 35a of the Administrative Procedures Act** and/or set out provisions, differentiated in terms of specific legislation, for administrative acts supported partially or fully by automation. Regulations on the partial and full automation of administrative procedures should be further developed as part of the horizontal and sectoral regulations for algorithmic systems recommended by the Data Ethics Commission (→ see section 3.3 above).

7.5 Algorithmic systems in public security law

The public discussion is especially critical of the use of algorithmic systems by security authorities. As administrative measures in this area can have a particularly profound effect on fundamental rights, the use of algorithmic should generally be **restricted**.

If algorithmic systems are used to predict crimes or threat situations **(predictive policing)**, consideration must be given to the fact that even systems which do not use any personal data can directly have effects relevant to fundamental rights. This is the case in particular if a reference to a person can be (re-)created by means of especially detailed location information. In addition, "location-related risk prognoses" can lead to excessive police checks in certain neighbourhoods identified as hotspots and therefore to the ethnic or social profiling of population groups living there. Such measures can also trigger crime relocation and displacement effects. The Data Ethics Commission therefore recommends making the security authorities of such effects and incorporating randomisations into the prediction systems in order to reduce corresponding effects and other system-based distortions; steps must also be taken to ensure that the security authorities can still always carry out a human review of more cases other than the risk cases selected by the system (cf. Section 88 of the Fiscal Code of Germany (*Abgabenordnung*, AO)). Nor should the security authorities be allowed to order further discretionary intervention measures solely on the basis of location-related forecasts.

Where **risk forecasts relating to individuals** are allowed by law in the area of security, such forecasts must not be created fully automatically where doing so could have negative legal consequences for the parties concerned. On account of the risk of automation bias, even in the case of algorithm-based decisions, support for human decision-makers from algorithmic systems in such profiling may, if at all, only be permissible within very strict limits.

7.6 Transparency requirements for the use of algorithmic systems by state actors

State decisions made using algorithmic systems must remain **transparent and justifiable**. This is, generally speaking, even more important than in the private sector due to the obligation to uphold fundamental rights and the need for democratic accountability of all authority and power in the public sector. Therefore, not only do the general transparency requirements (→ see section 4.1 above) apply to state bodies, but state bodies must also strive particularly hard to ensure openness.

The Data Ethics Commission points out that, in many cases, algorithmic systems used by state actors already fall within the scope of existing freedom of information and/or transparency laws. The Data Ethics Commission also welcomes the position paper "Transparency in Public Administration in the Use of Algorithms" ("*Transparenz der Verwaltung beim Einsatz von Algorithmen*") adopted during the 36th Conference of Freedom of Information Officers (*Konferenz der Informationsfreiheitsbeauftragten*) in Germany. According to this paper, state bodies must have meaningful, comprehensive and generally comprehensible information regarding their own data processing and, where legally possible, should publish it, including information (i) on the data categories of the procedure's input and output data; (ii) on the logic involved, in particular on the calculation formulae used including the weighting of the input data, on the underlying expertise and on the individual configuration deployed by the users; and (iii) on the scope of the resulting decisions and on the possible consequences of the procedures.[17]

17 Position paper as part of the 36th Conference of Freedom of Information Officers in Germany – "Transparenz der Verwaltung beim Einsatz von Algorithmen für gelebten Grundrechtsschutz unabdingbar" ["Transparency Public Administration in the Use of Algorithms as Essential for the Protection of Fundamental Rights"], Ulm, 16 October 2018 (available at: *https://www.datenschutzzentrum.de/uploads/informationsfreiheit/2018_Positionspapier-Transparenz-von-Algorithmen.pdf*).

With regard to specifying the corresponding transparency obligations and/or duties to provide access to information, the Data Ethics Commission also points out that insufficient provisions on transparency can lead to a lack of trust in the systems, which can lead to greater numbers of appeals, thereby counteracting any efficiency gains intended with the use of algorithmic systems. For that reason, the Data Ethics Commission ultimately believes that it is justifiable in no more than very few cases to rule out access to information regarding public algorithmic systems across the board by citing a risk of manipulation or the protection of business secrets. As a rule, therefore, the particular interests must be weighed against each other.

The disclosure of information on a system's general functionality will not be sufficient in every case where algorithmic systems are used by public authorities. Often, decisions made by public authorities must also be justified to the parties affected, i. e. the **"main factual and legal reasons"** which led to the decision in the particular case must be provided (cf. Section 39(1)(2) of the Administrative Procedures Act). Where such an individual explanation is required under constitutional or sub-constitutional law but, due to the technical complexity of the system, is not possible or is not possible in a way which, in the course of an official complaint procedure or before the court, enables an effective review of the viability of the reasoning, the use of algorithmic systems must be prohibited. Apart from that, the Data Ethics Commission believes that the State is required to build up sufficient **expertise** within administration and the courts to be able to ensure the necessary oversight of the system-internal decision-making processes.

The Data Ethics Commission points out that the transparency of state activity can also be negatively affected if the State uses proprietary software (closed-source software) of private providers in carrying out its duties. Generally speaking, proprietary software makes it difficult for users to make changes and adaptations, which results in a dependent relationship. In addition, the use of proprietary software leads to a lack of transparency and can therefore threaten public acceptance of the systems. Especially in areas which are sensitive in terms of fundamental rights, such as public security law, the use of proprietary software should therefore be avoided if possible. Instead, state bodies should rely on **open-source solutions** or develop their own systems ideally through interdisciplinary teams of developers. Where this is not practical, the Data Ethics Commission recommends that the Federal Government should consider amending public procurement law to minimise the aforementioned negative effects of proprietary software. Where there is no need to fear that the effectiveness of the system will suffer as a result of transparency, i. e. exploitation effects can be ruled out, the software should be developed in an open and consultative process with the inclusion of civil society stakeholders.

7.7 The risk involved in automated total enforcement

The Data Ethics Commission refuses, from an ethical point of view, to acknowledge any general right to non-compliance with rules and regulations. However, an automated total enforcement of the law raises a number of ethical concerns. For example, citizens might feel that full enforcement in practice places everyone under suspicion, which in turn, reduces their general willingness to obey rules and regulations. Furthermore, with automated enforcement, there is the danger that the complexity of real-life situations will not be sufficiently portrayed and, in particular, that unforeseen exceptional situations (for example, speeding in a private vehicle taking a seriously injured individual to the hospital) will not be sufficiently taken into consideration. Finally, many laws were not originally enacted for total enforcement. As a general rule, therefore, systems should be designed in such a way that a human can override technical enforcement in a specific case. In addition, each law enforcement measure constitutes state intervention and, as such, must be based on the **principle of proportionality**.

Summary of the most important recommendations for action

Use of algorithmic systems by state bodies

68

The State must, in the interests of its citizens, make use of the best available technologies, including algorithmic systems, but must also exercise particular prudence in all of its actions in view of its obligation to preserve fundamental rights and act as a role model. As a general rule, therefore, the use of algorithmic systems by public authorities should be assessed on the basis of the criticality model as **particularly sensitive**, entailing at the very least a comprehensive risk assessment.

69

In the areas of **law-making** and the **dispensation of justice**, algorithmic systems may at most be used for peripheral tasks. In particular, algorithmic systems must not be used to undermine the functional independence of the courts or the democratic process. By way of contrast, enormous potential exists for the use of algorithmic systems in connection with **administrative** tasks, in particular those relating to the provision of services and benefits. The legislator should take due account of this fact by giving the green light to a greater number of partially and fully automated administrative procedures. Cautious consideration should therefore be given to expanding the scope of both Section 35a of the German Administrative Procedures Act (*Verwaltungsverfahrensgesetz*, VwVfG) (which is couched in overly restrictive terms) and the corresponding provisions of statutory law. All of these measures must be accompanied by adequate steps to protect citizens.

70

Decisions taken by the State on the basis of algorithmic systems must still be **transparent**, and it must still be possible to provide **justifications** for them. It may be necessary to clarify or expand the existing legislation on freedom of information and transparency in order to achieve these goals. Furthermore, the use of algorithmic systems does not negate the principle that decisions made by public authorities must generally be justified individually; on the contrary, this principle may impose limits on the use of overly complex algorithmic systems. Finally, greater priority should be accorded to open-source solutions, since the latter may significantly enhance the transparency of government actions.

71

From an ethical point of view, there is no general right to non-compliance with rules and regulations. At the same time, however, automated "total" enforcement of the law raises a number of different ethical concerns. As a general rule, therefore, systems should be designed in such a way that a human can override **technical enforcement** in a specific case. The balance struck between the potential transgression and the automated (and perhaps preventive) enforcement measure must at all times meet the requirements of the proportionality principle.

8. Liability for algorithmic systems

8.1 Significance

Criminal responsibility, administrative sanctions and liability for damages are vital components of any ethically sound regulatory framework, especially for algorithmic systems and other digital technologies. From an ethical perspective, the Data Ethics Commission also highlights, in particular, the role of tort law, which serves both for compensation for and prevention of damage and therefore very significantly contributes to **the protection of legally protected interests in line with fundamental rights**.

From an ethical perspective, the following requirements, *inter alia*, must be set out for a liability system which needs to keep up with new digital technologies:

a) sufficient **compensation** for victims, in particular in the case of legally protected interests which are highly relevant in terms of fundamental rights and where compensation in a comparable situation involving humans or conventional technology would be owed;

b) provision of the right **behavioural incentives**, whereby damage is paid for by the actors who caused the damage through avoidable and undesirable behaviour or out of whose sphere the risk in question resulted;

c) **fairness**, whereby the actors liable to pay damages are those who, for example, placed the system on the market or who exercise control over the system and benefit from its use;

d) **efficiency**, whereby costs are covered (internalised) by the actors who can avoid or insure such costs with the least amount of effort.

8.2 Harm caused by the use of algorithmic systems

8.2.1 Liability of the "electronic person"?

The Data Ethics Commission **expressly advises against** granting robots or autonomous systems legal personality (often discussed using the keyword **"e-person"**) with the intention of making the systems liable themselves (e.g. a self-driving car with no registered owner, which "operates itself" as a mobility service). Such a measure would not achieve allocation of responsibility and liability for harm to those who are responsible for the use of the system and ultimately benefit economically from such use. In fact, the measure could, conversely, be used to evade responsibility. The legal personality of machines as a new type of legal entity would not enable any desirable outcome to be achieved which could not be achieved more freely and easily another way (for example with the help of company law). Treating autonomous machines even in analogy to natural persons would be a dangerous mistake.

8.2.2 Vicarious liability for "autonomous" systems

The Data Ethics Commission believes, however, that harm caused by autonomous systems should be attributed to those operating the systems according to the same rules of **vicarious** liability as would apply in the case of human auxiliaries (cf. in particular Section 278 of the [German] Civil Code). An actor which uses such a system in order to broaden its range of activities (for example a hospital which uses a surgical robot) should, in the event of a malfunction, not be able to release itself from liability because an actor which uses a human vicarious agent (for example a human surgeon) will be liable for any culpable misconduct of the vicarious agent, which is treated as behaviour on the part of the actor. This becomes particularly important in the case of **liability for an algorithmic system**, where otherwise liability loopholes will easily occur if no breach of duty of care by the person behind it can be proven in the use and monitoring of the algorithmic system.

Example 18
A surgical robot at a hospital makes an operational incision which is too long and causes complications. Or: an algorithmic system incorrectly derives the score for the creditworthiness of a bank's customer, which is why the customer cannot take up a one-off attractive offer relating to a property.

It may occasionally be difficult to establish an adequate equivalent to "standard of care" for autonomous systems, in particular as soon as the abilities of a machine exceed those of a human. In the majority of cases, however, malfunctions will be distinguishable from normal functions, and therefore this cannot, in general, be cited against the operator's liability. The standard must then be defined based on comparable systems available on the market, whereby the question as to the use of which technology could be expected of the operator must be decided on based on general principles (e.g. in that respect, the question as to what quality of surgical robot was to be used does not differ from the question as to what quality of X-ray device was to be used).

8.2.3 Strict liability

It is essentially a well-known fact that the rules relating to classic fault-based liability are not always sufficient for resolving the legal issues which arise in the case of dangerous products. The legal system has so far come up with a range of different answers to this challenge. In particular, these include:

- **modification of fault-based liability** (for example through adaptations of the standard of care and various ways of easing the burden of proof right through to the reversal of the burden of proof);

- various bases of **strict liability** (i.e. for facilities and activities which typically cause harm but which, on account of their benefit for society as a whole, should not be prohibited); and

- **product liability** in accordance with the [German] Act on Liability for Defective Products (*Gesetz über die Haftung für fehlerhafte Produkte*, ProdHaftG); it acts as a special form of liability regardless of fault which differs from strict liability on account of the fact that it requires, *inter alia*, a product defect and therefore comes fairly close to fault-based liability.

Steps must be taken to ensure that these answers lead to legally watertight solutions in terms of compensation for harm caused by dangerous digital applications.

The operation of digital applications currently involves **legal uncertainties and liability loopholes**, which primarily result from the unpredictability of harmful events, including when the applications are placed on the market (and hence possibly a failure of classic fault-based liability). They also result from the fact that, when various different actors and applications interact, generally speaking, it is almost impossible to prove where an error occurred and/or what the cause of the error was. The open and dynamic nature of digital ecosystems and the close functional interplay of products, digital contents and digital services also present a challenge. These legal uncertainties are, from the perspective of both companies and consumers, **obstacles to innovation and to the acceptance of new technologies**. If harmful events cannot routinely be assigned in terms of liability and compensated for, the impact on the market intended to be achieved through liability provisions cannot be achieved. In order to create an appropriate balance of interests, the legislator must provide for transparency and responsibility. Only if the responsibilities are clarified will it be possible to insure against harm or damage in practice.

The Data Ethics Commission cannot solve at this point the complex technical legal questions that arise, and cannot pin down the right solutions in terms of liability law, especially as, in some instances, the chances of finding a solution at European level should be explored first. From an ethical perspective, it is crucial that **legal clarity and legal certainty, in particular with regard to the liability principles described above**, be created. However, as the debate currently stands, it appears highly likely that, in addition to appropriate amendments to the Product Liability Directive (→ see section 8.2.4 below), certain changes may need to be made to the rules relating to fault-based liability and/or new bases of strict liability may need to be introduced.

In the legislative process, it will firstly be necessary to determine the liability regime that is most appropriate **for particular types of products, digital content and digital services**, and the exact shape that this regime should take, depending, once again, on the criticality (→ see section 3.1 above) of the relevant system, but also on other criteria which are specifically relevant within the context of liability. As such, strict liability (for example based on the model involving the car owner's liability) could be most appropriate in cases regarding devices where the operational risk is similarly uncontrollable and could end up leading to harm to life and limb. As part of this, the question of insurability and/or possible compulsory insurance must always play a role. A decision must also always be taken on **which type of harm** should be the subject of the liability (e.g. personal injury and damage to property, data loss, pure financial losses and non-material damage).

Ultimately, in each case, a decision will need to be taken as to who, taking into consideration the liability principles described above, is the right **party to which liability should be assigned**. There will, in particular, be three possible parties to which liability could be assigned, of which two could possibly also be jointly and severally liable:

- the individual **registered owner** of the system (i.e. the owner or person who, in a similar position, uses the system for their own purposes);
- the **manufacturer** of the system;
- the **operator** of the system (i.e. whoever exercises greater control over the system's operation, the individual registered owner as the front-end operator or a back-end operator who may also be the manufacturer but does not have to be).[18]

Determination of that party and of the type of liability will always depend on the specific type of networked or autonomous system and the identification of the specific spheres of liability.

8.2.4 Product security and product liability

Overall, it is currently important to highlight a paradigm shift from a situation whereby products were simply placed on the market to a situation whereby products are placed on the market but additional services continue to be provided for the products thereafter. As such, ongoing **product monitoring** and **product maintenance** are becoming more and more important. IT security and data protection standards not only have to be fulfilled when a product leaves the production plant but also must continue to be met as part of subsequent software updates. Conversely, in the event that security gaps subsequently appear, the manufacturer should (in accordance with the provisions of the directives on digital content and digital services and trade in goods) be subject to a duty to provide **security updates** in line with consumers' reasonable expectations regarding service life.

18 For the liability concept of such differentiated liability of the operator in digital ecosystems, see the report entitled "Liability for Artificial Intelligence and other emerging digital technologies" by the European Commission's Expert Group on Liability and New Technologies (New Technologies Formation), September 2019, no. [11], p. 40 et seqq.

Example 19
No security updates are provided for a smart home system and, as a result, following a cyber-attack, the house is broken into.

The Product Liability Directive from the 1980s is no longer able to cover the features of networked, hybrid or autonomous products. The Data Ethics Commission recommends that, as part of the **evaluation and revision of the Product Liability Directive** at European level, the Federal Government should push for watertight and clear legal provisions, in particular for the following aspects:

a) inclusion of digital content and digital services, including algorithmic systems, under the term "product";

b) liability for product defects which do not appear until after the product has been placed on the market and are the result of self-modifying software, of provision of updates or a failure to provide them, or of product-specific data feeds;

c) liability for breaches of the product monitoring obligation;

d) inclusion of legally protected interests typically affected by digital product safety, in particular the right to informational self-determination, in compensation regimes;

e) adaptation of the development risk defence.

8.3 Need for a reassessment of liability law

Digital ecosystems throw up a variety of other issues in connection with liability and responsibility. For example, there is, to some extent, a liability loophole in current tort law in cases of **damage to data and digital products**, provided that neither a recognised 'absolute right' has been infringed (e.g. ownership of the storage medium), nor a statute that is intended to protect another person breached (e.g. provisions of criminal law), nor the conditions of intentional damage contrary to public policy met. New digital technologies often also involve the **opportunistic use of other people's infrastructures** (e.g. the systematic collation and use by third parties of sensor data generated by private IoT devices or the direct use of computing capacity or transmission functions), which can create complicated liability issues. In contexts with a stronger focus on contract law, major harm or damage (in particular at the expense of consumers) can be caused on account of the fact that the **usability of high-value goods** (real property, machines, cars, etc.) is becoming increasingly dependent on the long-term provision of digital services (software updates, user accounts, etc.) and the provision of such services is not guaranteed and/or can even be specifically suspended in order to put individuals under pressure **(electronic repossession)**.

Digital ecosystems are also, to some extent, characterised by the interaction of numerous components and operators, whereby it is often disproportionately difficult for the injured party to prove **which of several potential tortfeasors** (e.g. hardware supplier, suppliers of various software components, data feed provider or network operator) caused the harm. On the other hand, digital technologies not only create a new lack of transparency with regard to the cause of harm or damage but can, conversely, also help in documenting the course of causal events in an unprecedented way. The question therefore arises as to which actor is obliged to contribute to providing clarification on the cause of the harm by already **logging data** ex-ante and to whom the data actually recorded via logging should be disclosed in the event of harm.

The Data Ethics Commission therefore recommends overall that the Federal Government should investigate the extent to which current liability law has kept up with the **challenges of digital ecosystems** or needs to be reworked. Priority must be given to striving to achieve a solution at European level. The Data Ethics Commission advises in this context against any tendency towards a one-sided focus on specific technological features, in particular the feature of machine learning. Whilst machine learning creates certain additional dangers and involves certain additional issues regarding the assignment of liability, most challenges for liability law are attributable to other factors (e.g. intangibility, interaction of numerous components, networking and decentralisation).

Summary of the most important recommendations for action

Liability for algorithmic systems

72

Liability for damages, alongside criminal responsibility and administrative sanctions, is a vital component of any ethically sound regulatory framework for algorithmic systems. It is already apparent today that algorithmic systems pose challenges to liability law as it currently stands, *inter alia* because of the complexity and dynamism of these systems and their growing "autonomy". The Data Ethics Commission therefore recommends that the current provisions of liability law should undergo in-depth checks and (where necessary) revisions. The scope of these checks and revisions should not be restricted on the basis of too narrowly defined technological features, such as machine learning or artificial intelligence.

73

The proposal for a future system under which legal personality would be granted to high-autonomy algorithmic systems, and the systems themselves would be liable for damages (**"electronic person"**), should **not be pursued further**. As far as this concept is, by some protagonists, based on a purported equivalence between human and machine it is ethically indefensible. And as far as it boils down to introducing a new type of company under company law it does not, in fact, solve any of the pertinent problems.

74

By way of contrast, if harm is caused by autonomous technology used in a way functionally equivalent to the employment of human auxiliaries, the operator's liability for making use of the technology should correspond to the otherwise existing vicarious **liability regime of a principal for such auxiliaries** (cf. in particular Section 278 of the German Civil Code). For example, a bank that uses an autonomous system to check the creditworthiness of its customers should be liable towards them to at least the same extent that it would be had it used a human employee to perform this task.

75

As the debate currently stands, it appears highly likely that appropriate amendments will need to be made to the **Product Liability Directive** (which dates back to the 1980s), and a connection established to new product safety standards; in addition, certain changes may need to be made to the rules relating to **fault-based liability** and/or new bases of **strict liability** may need to be introduced. In each case, it will be necessary to determine the liability regime that is most appropriate for particular types of products, digital content and digital services, and the exact shape that this regime should take (once again depending on the criticality of the relevant algorithmic system). Consideration should also be given to innovative liability concepts currently being developed at European level.

Part G

A European path

The Data Ethics Commission has examined a great many different questions, and discussions on these questions have raised new ones in turn; this alone should indicate that this Opinion can serve only as one out of many building blocks in the larger edifice of a broad-based **debate on the future of ethics, law and technology** that we must return to again and again. This debate must be interdisciplinary from the outset, and encompass a broad range of sciences and a diverse mix of representatives from the worlds of the economy, civil society and politics. In view of the immense economic pressure and the fast-paced nature of technological change, the findings that emerge from this debate must be integrated on an ongoing basis into the activities of the parties involved at all levels, so that we can shape a technological future that is founded on values.

Data transfers and the use of algorithmic systems transcend national boundaries, which means that a forward-looking discussion of the ethical and legal issues arising in connection with data and algorithmic systems must not be restricted to the national level. We need to view problems from a **global perspective**, and accordingly strive to present our findings and perspectives more than before on a pan-European debate as well. Lessons learned from implementing the GDPR have shown that the economic clout of the European Economic Area and its significance as a market for the operators and providers of algorithmic systems may ultimately mean that these latter are prompted by economic interests to comply with the EU's basic requirements when developing and implementing their products and services. These European requirements are also being used by ever more non-European governments as a reference point when drafting their own regulatory frameworks.

The debate that needs to take place should therefore be a priority topic on the agendas of international forums (EU, OECD, Council of Europe, United Nations, G7 and G20). With this in mind, the Data Ethics Commission recommends that the Federal Government should make its voice heard within these international bodies. In particular, the **German Presidency of the EU Council** in the second half of 2020 should be utilised as an opportunity to promote the measures to deal with data governance and algorithmic systems as proposed in this opinion on the European level. The Data Ethics Commission also believes that the Federal Government should be actively involved (both in the early stages of the process and on an ongoing basis) in the establishment of an International Panel on Artificial Intelligence (IPAI) as initiated on the level of the G7.

In the global contest for future technologies, Germany and Europe are being confronted with value systems, models of society and cultures that differ widely from our own. This has prompted a debate whether Germany and Europe are to adapt to one or the other non-European models in order to remain competitive. The Data Ethics Commission supports the **"European path"** which has been followed to date. It is often referred to in debates as a "third way" that strikes a balance between the US and Chinese positions, and which asserts that the defining feature of European technologies should be their consistent alignment with European values and fundamental rights, in particular those enshrined in the European Union's Charter of Fundamental Rights and the Council of Europe's Convention for the Protection of Human Rights and Fundamental Freedoms.

In order to remain actively involved in the future debate on the interplay between ethics, law and technology, the digital sovereignty of Germany and Europe must be preserved to the greatest extent possible. When used in reference to nation states or organisations, the term "digital sovereignty" encompasses every aspect of data processing, i.e. control over the storage, transfer and use of the sensitive data held by these bodies, and autonomous decisions on who can access them.

A globalised world in which people, states and companies co-exist side by side requires cross-border flows of data, and the Internet – which serves as the conduit for these flows – is a global "network of networks"; this distributed global structure, which embraces very different legal and societal systems, renders complete sovereignty an impossible task. The debate on digital sovereignty must therefore tackle vital questions relating to technical infrastructure, including hardware, networks, control components such as routers or address servers, and data centres. With a view to preserving the digital sovereignty of Germany and Europe, and given the huge extent to which we are reliant on foreign products, the Data Ethics Commission believes that there is an urgent need to take action at German and European level through **investments into developing and safeguarding appropriate technologies and infrastructures**.

Virtually all of the most important and basic Internet infrastructure components that are used in Germany (and indeed in Europe as a whole) can be procured only from other continents at present, and so efforts to preserve sovereignty must be restricted for now to the two main avenues open to us; the first of these is the critical analysis and assessment of the basic components being used, and the second is the application of the highest possible security standards when operating them in order to minimise the risk of misuse by foreign states and organisations. Looking to the future, however, the Data Ethics Commission believes that it is important for Germany and Europe as a whole to develop a **higher level of digital sovereignty**, right down to the level of **technical infrastructure**. Support should be available for R&D work on systems that comply with the highest possible standards of security. Work of this kind would include both the design of new components to replace previous systems, and attempts to engineer integrated solutions that use existing components and that achieve the required level of protection in spite of known or suspected inadequacies or security risks.

The digital sovereignty of a nation state should be viewed not only in relation to other nation states, but also in relation to non-state actors that wield significant amounts of power. As the data economy grows, there is a trend for **economic power to be concentrated** in the hands of a few, and the emergence of **new power imbalances** is apparent. To an ever greater extent, R&D work on algorithmic systems and other digital technologies is being carried out within a framework established by a small group of digital giants; what is more, these companies often act as an important source of public research funding and therefore have a say in this research. Over the past few decades, intermediaries have played an increasingly important role in forming opinions, and therefore in influencing the sociopolitical discourse; this means that the associated risk of abuse has also increased. Given the importance of ethical and legal fundamental values and freedoms and to preserve the digital sovereignty of Germany and Europe, the Data Ethics Commission believes that there is an urgent need to closely monitor the shifts in power structures, as those are vital for the functioning of a democratic State and a social market economy, and to efficiently regulate the according areas wherever needed.

Excessive dependence on others turns a nation into a rule taker rather than a rule maker, resulting in the citizens of this nation being subject to requirements imposed by players elsewhere in the world. Embarking on **efforts to safeguard digital sovereignty in the long term** is therefore not only a politically far-sighted necessity, but also an expression of ethical responsibility.

Appendix

1. The Federal Government's key questions to the Data Ethics Commission

Coalition Agreement:

"We will set up a data ethics commission that within the next year will provide the government and parliament with proposals on how to develop data policy and deal with algorithms, artificial intelligence and digital innovation. Clarification of data ethics questions can add impetus to the process of digital development and can help define an approach towards resolving social conflicts within the area of data policy."

Key questions for the Data Ethics Commission:

Digitisation is fundamentally changing our society. New data-based technologies can be beneficial for people's everyday lives as well as for industry, the environment, science and society as a whole. Their potential is enormous.

At the same time, digitisation also clearly brings certain risks. Numerous ethical and legal questions are raised, particularly concerning the effects of these developments and the desired role of new technologies. If digital change is to benefit the whole of society, we need to examine the possible consequences of new technologies and establish ethical safeguards.

One challenge is to develop 21st-century law in a way that protects human dignity ("a human being must not become a mere object") and guarantees fundamental and human rights such as the general right of personality, the right to privacy, the right to informational self-determination, freedom from discrimination, freedom of science, freedom to conduct a business, and freedom of expression and information – bringing all of these rights into equilibrium with one another.

There are complex tensions between the principles of the common good, progress, innovation and solidarity.

The task of this Commission – having identified the current state of discussion and legislation at the European and international level, ascertained the possibilities for positive action at the national level, and given special consideration to sensitive areas – is to develop ethical standards and guidelines for the protection of individuals, the preservation of social cohesion and the safeguarding and promotion of prosperity in the information age. The Commission is also tasked with providing the Federal Government with recommendations and regulatory proposals on how ethical guidelines can be developed, respected, implemented and monitored. These proposals should also include a description of the underlying concepts used, as well as assessments of the possible consequences and side effects.

The public is to be appropriately involved in the work of the Commission.

In order to help the Data Ethics Commission carry out its work, the Federal Government has provided it with the following key questions in three areas:

I. Algorithmic decision-making (ADM)

Advanced automation systems are increasingly shaping economic and social realities and people's everyday lives. Data collection and analysis enable the development of innovative interpretation models, which are also used to make or prepare algorithm-based decisions. Algorithms make it possible, for example, to recognise patterns and differences in the behaviour of different groups. Whether it is a matter of setting individual prices in e-commerce, assessing creditworthiness or selecting candidates in recruitment procedures, people are being evaluated by technical processes in more and more areas of life. Data evaluation and predictions about individual behaviour can offer opportunities (e.g. aiding research, strengthening innovation within industry, increasing the efficiency of data processing processes), but they also harbour risks (e.g. for individual freedom and self-determination, for participation and equal opportunities among certain individuals and social groups). Social inequality and discrimination against individuals or groups of individuals can be perpetuated if biases are incorporated into the programming of an algorithm or its training data. These risks are particularly acute in participation-relevant and personality-sensitive ADM processes. Against this background, the following questions arise, especially with regard to consumer protection:

- What are the ethical limits to using ADM processes? Or what ethical limits should there be?
- Can it be ethically necessary to use ADM processes?
- Are there characteristics, criteria or certain kinds of data that should not be incorporated into ADM processes – due to their age or origin, for example?
- How can we determine which prejudices and distortions in which areas are ethically undesirable? What effects can the use of ADM processes have on social groups?
- What regulatory approaches could be used to prevent manipulation, unequal treatment and discrimination?
- Is it advisable to have a graduated regulatory framework based on the risk to social participation or the potential for discrimination?
- How can the reliability, reproducibility and scrutiny of ADM be guaranteed?
- Are there limits to the use of ADM if its use and criteria cannot be explained to the people affected?
- Are there test methods that can make self-learning ADM open to scrutiny?

II. Artificial Intelligence (AI)

With the development of AI, industrial and administrative environments are deploying more and more highly automated systems that use AI methods and have the ability to "learn" through the use of training data. In addition, work is being done on simulating the cognitive functions of the human brain. The developments in the field of artificial intelligence raise the question of how the dignity, autonomy and self-determination of the individual can be safeguarded and fostered. This leads to questions such as the following:

- What fundamental ethical principles must be observed when developing, programming and using AI?
- Where do the ethical boundaries lie for using AI and robots, especially in special areas of life such as care/assistance and dealing with particularly vulnerable groups (children, the elderly, people with disabilities)? Can it be ethically necessary to use AI?
- Is "ethics by design" possible for AI? If so, how could it be implemented and monitored?
- How can it be ensured that machines working on an AI basis can be controlled?

- To whom are the creations/inventions generated by AI to be ascribed? Who should bear the responsibility for malfunctioning systems? How can the responsibility of the actors involved in the development and use of AI systems (programmers, data scientists, clients, etc.) be made transparent?

- What else will be necessary in the future to sustainably guarantee the freedoms and fundamental rights upon which our society is based?

III. Data

Digitisation is characterised by an increase in the volume of data (big data), by a vast accumulation of data by individual actors, by the high speed of data processing (real time), by connectivity (internet, complex networks of actors, Internet of Things), by the increasing ubiquity and permanence of data, and by the further development of various methods of data analysis. As the amount of available data increases, so too does the ability to undertake more granular analyses. Data is used to develop new business models and change value-added chains and work processes. By some, data is regarded as a commodity that enables value creation ("data economy").

At both the national and European level, there are current laws (e.g. the General Data Protection Regulation, open data legislation) and numerous legislative initiatives that concern the handling of data (e.g. the ePrivacy Regulation, legislative proposals regarding the free flow of data). On the one hand, these are intended to safeguard fundamental rights such as the right to informational self-determination, while on the other hand they are intended to enable useful and innovative data processing. Further proposals are discussed as to whether and how access to data, use of data, trade in data, and rights to data could be regulated for the first time or be better regulated.

In the process, the following questions may arise regarding the handling of data in general, data access and the use of data:

- What are the ethical limits to the economization of data?

- Who should be permitted to derive economic benefit from data?

- Should there be an obligation to offer payment models?

- Is it advisable to have uniform rules that apply equally to all data? Or should preference be given to rules that apply to specific areas (e.g. for brain data)? What should be the connecting factor for rules that apply to specific areas?

- What consequences do existing access and exclusivity rights to data have for competition and innovation? And what consequences would additional access and exclusivity rights to data have?

- Is there a need for the state to offer support as part of its provision of general public services so that citizens can navigate the internet and social networks in a responsible, competent and confident manner and learn how to handle data? Can the provision of data, in particular open data, become part of the provision of public services by the state?

- How much transparency is necessary and appropriate to safeguard the right to informational self-determination and to enable citizens to participate in economic life in a self-determined manner?

- Do particular life circumstances require special protection concepts for specific user groups?

- Are the existing institutions in sensitive areas sufficient to ensure data is used ethically? How can adequate stakeholder representation be ensured in the long term?

- What effects can extensive data collections have on the functioning of the market economy (e.g. competitiveness, information asymmetry between suppliers and consumers, the possibility of developing innovative products) and democracy (e.g. recording and analysing behaviour in social networks)? If necessary, how can action be taken against data power/data silos (especially intermediaries)?

- Should data or access to data be declared a public good in certain cases? In which cases and under which ethical criteria?

- The use of non-personal data can have collective effects. For example, individuals or certain population groups may be placed at a disadvantage because data analysis shows that payment habits are worse in a particular neighbourhood. What regulatory instruments would be needed for this? In which sectors?

- Are statutory regulations on improving access to data possible, necessary and advisable?

- Should data processing be prohibited in certain cases for ethical reasons, for example in cases involving certain types of data (e.g. political views; brain data) or certain areas of use (e.g. profiling for political purposes or for use in elections)?

- Under what circumstances can there be an ethical obligation to use data?

- Does the legal system sufficiently recognise the possible benefits that data processing can have for the common good? If not, how can this be achieved?

- Is it possible and advisable to create experimentation clauses for testing new applications or new regulatory instruments?

- Does it make sense to invest in data infrastructures? If so, in which ones?

- How can the constitutionally protected interests of individuals, enterprises, science and art be reconciled with the public interest in the use of data?

2. Members of the Data Ethics Commission

Co-Spokespersons

Prof. Dr Christiane Wendehorst

- Professor of Civil Law at the University of Vienna
- Co-Head of the Department of Innovation and Digitalisation in Law at the University of Vienna
- President of the European Law Institute (ELI)

Prof. Dr Christiane Woopen

- Professor for Ethics and Theory of Medicine and Head of the Research Unit Ethics at the University Clinic of Cologne
- Executive Director of the Cologne Center for Ethics, Rights, Economics, and Social Sciences of Health (ceres) at the University of Cologne
- Chair of the European Group on Ethics in Science and New Technologies (EGE)

Members

Prof. Dr Johanna Haberer

- Professor of Christian Media Studies at Friedrich Alexander University Erlangen Nuremberg (FAU)
- Director of the Institute for Practical Theology at Friedrich Alexander University Erlangen Nuremberg (FAU)

Prof. Dr Dirk Heckmann

- Full Professor of Law and Security of Digitization at the Technical University of Munich (TUM)
- Director at the Bavarian Research Institute for Digital Transformation
- Judge at the Bavarian Constitutional Court

Marit Hansen

- Data Protection Commissioner of Land Schleswig-Holstein
- Head of Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (Independent Centre for Privacy Protection Schleswig-Holstein)

Prof. Ulrich Kelber

- Federal Commissioner for Data Protection and Freedom of Information
- Honorary Professor at Bonn-Rhein-Sieg University of Applied Sciences (H-BRS)

Prof. Dieter Kempf

- President of the Federation of German Industries (BDI)
- Honorary Professor at Friedrich Alexander University Erlangen Nuremberg (FAU)

Prof. Dr Mario Martini

- Professor of Public Administration, Public Law, Administrative Law and European Law at the German University of Administrative Sciences Speyer (DUV Speyer)
- Head of the Programme Area "Transformation of the State in the Digital Age" and Deputy Director of the German Research Institute for Public Administration (FÖV)

Klaus Müller

- Executive Director of the Federation of German Consumer Organisations (vzbv)
- Lecturer at Heinrich Heine University Düsseldorf (HHU)

Paul Nemitz

- Principle Advisor at the European Commission, Directorate-General for Justice and Consumers

Prof. Dr Sabine Sachweh

- Professor for Applied Software Engineering at Dortmund University of Applied Sciences and Arts (FH Dortmund)
- Spokesperson and Board Member of the Institute for the Digital Transformation of Application and Living Domains (IDiAL) at Dortmund University of Applied Sciences and Arts (FH Dortmund)
- Co-Spokesperson of the "Digitalisation and Education for the Elderly" Advisory Council at the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth

Christin Schäfer

- Founder and Managing Director of the company acs plus, a data science boutique
- Advisor of the Big Data Analytics Research Group at the German Economic Institute in Cologne (IW Köln)

Prof. Dr Rolf Schwartmann

- Professor of Civil Law and Economic Law at Cologne University of Applied Sciences (TH Köln)
- Head of the Research Centre for Media Law at Cologne University of Applied Sciences (TH Köln)
- Chairman of the German Association for Data Protection and Data Security (GDD)

Prof. Dr Judith Simon

- Professor for Ethics in Information Technology at the University of Hamburg (UHH)

Prof. Dr Wolfgang Wahlster

- Professor of Computer Science, Chair for Artificial Intelligence, Saarland University
- CEO/CEA of the German Research Center for Artificial Intelligence (DFKI)
- Head of the Steering Committee for the AI Standardisation Roadmap at the German Institute for Standardization (DIN)

Prof. Dr Thomas Wischmeyer

- Assistant Professor (Tenure Track) for Public Law and Information Law at the University of Bielefeld

Current as of: 10 October 2019

Imprint

Berlin, December 2019

Opinion of the Data Ethics Commission

Publisher
Data Ethics Commission of the Federal Government
Federal Ministry of the Interior, Building and Community
Alt-Moabit 140, 10557 Berlin
Federal Ministry of Justice and Consumer Protection
Mohrenstraße 37, 10117 Berlin

E-mail
datenethikkommission_gs@bmi.bund.de
datenethikkommission_gs@bmjv.bund.de

Website
www.datenethikkommission.de

Design
Atelier Hauer + Dörfler GmbH, Berlin

Photo credits
p. 53: shutterstock.com
p. 234: BMI (group photo), Studio Wilke (Christiane Wendehorst), Reiner Zensen (Christiane Woopen), BPA/Kugler (Ulrich Kelber)
p. 235: Christian Kruppa (Dieter Kempf), vzbv/Gert Baumbach (Klaus Müller), Markus Mielek (Sabine Sachweh), TH Köln/Schmülgen (Rolf Schwartmann), UHH/Nicolai (Judith Simon), Jim Rakete (Wolfgang Wahlster)

Printing
Brandenburgische Universitätsdruckerei und Verlags-gesellschaft Potsdam mbH (bud)

Gutachten der Datenethikkommission

daten
ethik
kommission

Gutachten der Datenethikkommission

daten
ethik
kommission

Ausschließlich zum Zweck der besseren Lesbarkeit wird im vorliegenden Gutachten der Datenethikkommission auf die geschlechtsspezifische Schreibweise verzichtet. Alle personenbezogenen Bezeichnungen sind geschlechtsneutral zu verstehen.

Inhaltsübersicht

Inhaltsverzeichnis

Executive Summary

Die Digitalisierung verändert unsere Gesellschaft tiefgreifend. Neuartige datenbasierte Technologien können für das Leben des Einzelnen und das gesellschaftliche Zusammenleben Nutzen stiften, die Produktivität der Wirtschaft steigern, zu mehr Nachhaltigkeit und zu grundlegenden Fortschritten in der Wissenschaft beitragen. Gleichzeitig zeigen sich jedoch auch Risiken der Digitalisierung für grundlegende Rechte und Freiheiten. Es stellen sich damit zahlreiche ethische und rechtliche Fragen, in deren Mittelpunkt die gewünschte Rolle und die Gestaltung der neuen Technologien stehen. Wenn der digitale Wandel dem Wohl der gesamten Gesellschaft dienen soll, müssen sich Gesellschaft und Politik mit der Gestaltung datenbasierter Technologien einschließlich der Künstlichen Intelligenz (KI) befassen.

Die Bundesregierung hat am 18. Juli 2018 die Datenethikkommission (DEK) eingesetzt. Sie erhielt den Auftrag, innerhalb eines Jahres ethische Maßstäbe und Leitlinien sowie konkrete Handlungsempfehlungen für den Schutz des Einzelnen, die Wahrung des gesellschaftlichen Zusammenlebens und die Sicherung und Förderung des Wohlstands im Informationszeitalter zu entwickeln. Dazu hat die Bundesregierung der DEK Leitfragen an die Hand gegeben, die sich auf die drei Themenfelder Algorithmenbasierte Prognose- und Entscheidungsprozesse (ADM), KI und Daten konzentrieren. Aus Sicht der DEK ist allerdings KI lediglich eine besondere Ausprägung algorithmischer Systeme und teilt viele ethisch und rechtlich relevante Eigenschaften mit anderen Arten solcher Systeme, weshalb die DEK ihre Ausführungen auf **Daten** und **algorithmische Systeme** allgemein bezieht.

Die DEK hat sich für ihr Gutachten an den folgenden **Leitgedanken** orientiert:

- Menschenzentrierte und werteorientierte Gestaltung von Technologie
- Förderung digitaler Kompetenzen und kritischer Reflexion in der digitalen Welt
- Stärkung des Schutzes von persönlicher Freiheit, Selbstbestimmung und Integrität
- Förderung verantwortungsvoller und gemeinwohlverträglicher Datennutzungen
- Risikoadaptierte Regulierung und wirksame Kontrolle algorithmischer Systeme
- Wahrung und Förderung von Demokratie und gesellschaftlichem Zusammenhalt
- Ausrichtung digitaler Strategien an Zielen der Nachhaltigkeit
- Stärkung der digitalen Souveränität Deutschlands und Europas

1

Allgemeine ethische und rechtliche Grundsätze und Prinzipien

Der Mensch ist moralisch verantwortlich für sein Handeln – er kann der moralischen Dimension nicht entkommen. Welche Ziele er verfolgt, welche Gründe er dafür hat und welche Mittel er einsetzt, liegt in seiner Verantwortung. Bei der Gestaltung unserer technologisch geprägten Zukunft ist dieser Dimension sowie der gesellschaftlichen Bedingtheit des menschlichen Handelns stets Rechnung zu tragen. Dabei gilt unverrückbar, dass Technik dem Menschen dient und nicht der Mensch der Technik unterworfen wird. Dieses **Verständnis vom Menschen** liegt unserer Verfassungsordnung zugrunde und steht in der Tradition der europäischen Kultur- und Geistesgeschichte.

Durch digitale Technologien hat sich unser ethischer Ordnungsrahmen im Sinne der grundlegenden Werte, Rechte und Freiheiten, wie sie in der deutschen Verfassung und in der europäischen Charta der Grundrechte verankert sind, nicht verändert. Diese Werte, Rechte und Freiheiten erfordern angesichts neuer Herausforderungen jedoch eine erneute Vergewisserung und neue Abwägungen. Die folgenden ethischen und rechtlichen Grundsätze und Prinzipien hält die DEK vor diesem Hintergrund für gesellschaftlich anerkannte und unverzichtbare Handlungsmaßstäbe:

Die Würde des Menschen

Die Würde des Menschen, die für den unbedingten Wert jedes menschlichen Lebewesens steht, verbietet etwa die digitale Totalvermessung des Individuums ebenso wie seine Herabwürdigung durch Täuschung, Manipulation oder Ausgrenzung.

Selbstbestimmung

Die Selbstbestimmung ist elementarer Ausdruck von Freiheit und schließt die informationelle Selbstbestimmung mit ein. Wird der Mensch selbstbestimmter Akteur in der Datengesellschaft, kann von „digitaler Selbstbestimmung“ gesprochen werden.

Privatheit

Das Recht auf Privatheit dient der Wahrung der Freiheit und der Integrität der persönlichen Identität. Sie kann durch umfassende Erhebung und Auswertung von Daten bis hin in die intimsten Bereiche bedroht sein.

Sicherheit

Die körperliche und emotionale Sicherheit des Menschen und die Sicherheit der Umwelt schützen hochrangige Güter. Sicherheit zu gewährleisten stellt hohe Anforderungen beispielsweise in der Mensch-Maschine-Interaktion oder bezüglich der Resilienz von Systemen gegenüber Angriffen und missbräuchlicher Verwendung.

Demokratie

Digitale Technologien sind systemrelevant für die Entfaltung der Demokratie. Sie ermöglichen neue Formen der politischen Beteiligung, können aber auch Gefahren im Hinblick auf Manipulation und Radikalisierung mit sich bringen.

Gerechtigkeit und Solidarität

Angesichts der massiven daten- und technologieinduzierten Anhäufung von Macht und neuen Gefahren von Ausgrenzung und Diskriminierung ist die Gewährleistung von Zugangs- und Verteilungsgerechtigkeit eine dringliche Aufgabe. Digitalisierung sollte gesellschaftliche Teilhabe unterstützen und damit den sozialen Zusammenhalt fördern.

Nachhaltigkeit

Digitale Entwicklung steht auch im Dienste nachhaltiger Entwicklung. Digitale Technologien sollten dazu beitragen, ökonomische, ökologische und soziale Nachhaltigkeitsziele zu verwirklichen.

Ethik geht nicht im Recht auf, d. h. nicht alles, was ethisch relevant ist, kann und sollte rechtlich reguliert werden, und umgekehrt gibt es Aspekte rechtlicher Regulierung, die rein pragmatisch motiviert sind. Das Recht muss aber mögliche ethische Implikationen stets reflektieren und ethischen Ansprüchen genügen. Die DEK ist der Ansicht, dass **ethische Grundsätze und Prinzipien rechtliche Regulierung nicht entbehrlich machen können**. Dies ist insbesondere dort der Fall, wo angesichts der Grundrechtsrelevanz eine Entscheidung des demokratisch legitimierten Gesetzgebers notwendig ist. Dies legt zudem die Grundlage dafür, dass Bürger, Unternehmen und Institutionen auf eine ethisch ausgerichtete gesellschaftliche Transformation vertrauen können. **Regulierung soll gleichwohl technologische und soziale Innovationen sowie eine dynamische Marktentwicklung nicht blockieren.** Allzu starre und detaillierte Gesetze können Handlungsspielräume einschränken und bürokratischen Aufwand auf eine Weise erhöhen, dass innovative Prozesse in Deutschland der Geschwindigkeit der internationalen technologischen Entwicklungen nicht mehr folgen können.

Das Recht ist allerdings nur eines von mehreren Formaten, um ethische Prinzipien zu implementieren. Die Komplexität und Dynamik von Datenökosystemen erfordert das **Zusammenwirken verschiedener Governance-Instrumente** auf unterschiedlichen Ebenen (Mehr-Ebenen-Governance). Diese Instrumente umfassen neben rechtlicher Regulierung und Standardisierung verschiedene Formen der Ko- oder Selbstregulierung. Ferner kann Technik und ihr Design selbst als Governance-Instrument genutzt werden. Das Gleiche gilt für Geschäftsmodelle und Möglichkeiten ökonomischer Lenkung. In einem weiteren Sinne gehören zur Governance auch bildungs- und forschungspolitische Entscheidungen. Jedes der genannten Governance-Instrumente muss nicht nur national, sondern gerade auch **europäisch und international** gedacht werden.

Aus Sicht der DEK sind die Leitfragen der Bundesregierung aus zwei verschiedenen Perspektiven formuliert, einer primär auf Daten fokussierten Perspektive **(„Daten-Perspektive“)** und einer primär auf algorithmische Systeme fokussierten Perspektive **(„Algorithmen-Perspektive“)**. Bei den beiden Perspektiven handelt es sich weder um miteinander konkurrierende Sichtweisen noch um verschiedene Seiten ein- und derselben Medaille, sondern um **sich wechselseitig ergänzende und bedingende ethische Diskurse**, welche sich typischerweise auch in unterschiedlichen Governance-Instrumenten, einschließlich unterschiedlicher Rechtsakte, widerspiegeln.

2

Daten

Die **Daten-Perspektive** richtet die Sicht auf die digitalen Daten, die zum Maschinellen Lernen, als Datenbasis für algorithmisch geprägte Entscheidungen und für eine Fülle weiterer Zwecke verwendet werden. Sie betrachtet Daten vor allem im Hinblick auf deren Herkunft sowie auf die möglichen Auswirkungen der Datenverarbeitung auf bestimmte Akteure, die mit Kontext und Bedeutungsgehalt der Daten zu tun haben, sowie auf die Gesellschaft. Aus ethischer wie aus rechtlicher Sicht geht es einerseits um **objektive Anforderungen** an den Umgang mit Daten, noch mehr aber typischerweise um **subjektive Rechte**, welche Akteure gegenüber einem bestimmten anderen Akteur oder auch gegenüber jedermann geltend machen können. Eine zentrale Unterscheidung ist diejenige zwischen personenbezogenen und nicht personenbezogenen Daten, welche über die Anwendbarkeit des Datenschutzrechts entscheidet.

Allgemeine Anforderungen an den Umgang mit Daten

Zu den objektiven Anforderungen an jede verantwortungsvolle Nutzung von Daten gehören nach Auffassung der DEK die folgenden datenethischen Grundsätze:

- **Vorausschauende Verantwortung:** Bei der Sammlung, Verarbeitung und Weitergabe von Daten müssen mögliche Auswirkungen auf Einzelne oder die Allgemeinheit unter Berücksichtigung künftiger Akkumulations-, Netzwerk- und Skaleneffekte, technologischer Möglichkeiten und Akteurskonstellationen abgeschätzt werden.

- **Achtung der Rechte beteiligter Personen:** Akteure, die an der Generierung von Daten beteiligt waren – sei es als Subjekt der Information, sei es in einer anderen Rolle –, können Rechte in Bezug auf diese Daten zustehen, die zu achten sind.

- **Wohlfahrt durch Nutzen und Teilen von Daten:** Daten können als nicht-rivales Gut vervielfältigt und parallel von vielen Akteuren zu vielen verschiedenen Zwecken genutzt werden und damit das Gemeinwohl fördern.

- **Zweckadäquate Datenqualität:** Ein verantwortungsvoller Umgang mit Daten setzt die Sicherstellung einer dem jeweiligen Zweck angemessenen Datenqualität voraus.

- **Risikoadäquate Informationssicherheit:** Daten sind anfällig gegenüber Ausspähung und Verfälschung von außen und können, in andere Hände gelangt, nur schwer zurückgeholt werden. Es bedarf daher eines dem jeweiligen Risikopotenzial angemessenen Maßes an Informationssicherheit.

- **Interessenadäquate Transparenz:** Derjenige, der Daten als Verantwortlicher verarbeitet, muss bereit und in der Lage sein, dafür Rechenschaft abzulegen. Dies erfordert ein angemessenes Maß an Transparenz und Dokumentation des Handelns und ggf. auch entsprechende Haftungsregelungen.

Datenrechte und korrespondierende Datenpflichten

Um sich als Akteure in der Datengesellschaft selbstbestimmt bewegen zu können, bedürfen Personen subjektiver Rechte, die ihnen gegenüber anderen Akteuren zustehen. Dies betrifft in erster Linie die Rechte eines jeden Menschen in Bezug auf seine **personenbezogenen Daten**, die sich aus dem grundrechtlich verbürgten Recht auf informationelle Selbstbestimmung ableiten und durch das geltende Datenschutzrecht gewährleistet werden. Digitale Selbstbestimmung umfasst darüber hinaus auch die selbstbestimmte wirtschaftliche Verwertung der eigenen Datenbestände sowie den selbstbestimmten Umgang mit **nicht-personenbezogenen Daten**, die etwa durch den Wirkbetrieb eigener Geräte generiert werden. Nach Auffassung der DEK gilt ein Recht auf digitale Selbstbestimmung im Grundsatz auch für Unternehmen und **juristische Personen** und – zumindest in Ansätzen – für Gruppen von Personen (Kollektive).

Vielfach tragen unterschiedliche Akteure in unterschiedlichen Rollen zur Generierung von Daten bei – sei es als Subjekt der Information, sei es als Eigentümer einer datengenerierenden Vorrichtung, sei es in einer anderen Rolle. Ein solcher Beitrag zur Generierung von Daten sollte nach Auffassung der DEK aber nicht zu exklusiven Eigentumsrechten an Daten führen, sondern vielmehr gegebenenfalls zu Datenrechten in der Form spezieller **Mitsprache- und Teilhaberechte** eines Akteurs, mit denen korrespondierende Pflichten anderer Akteure einhergehen. Anerkennung und Ausgestaltung solcher Datenrechte eines Akteurs hängen von den folgenden allgemeinen Faktoren ab:

a) Umfang und Art des **Beitrags dieses Akteurs zur Datengenerierung**;

b) **Gewicht seines Individualinteresses** an der Gewährung des Datenrechts;

c) Gewicht der ggf. **konfligierenden Individualinteressen** desjenigen Akteurs, dem gegenüber das Datenrecht geltend gemacht wird, oder Dritter, unter Berücksichtigung von Ausgleichsmöglichkeiten (z. B. Schutzmaßnahmen, Vergütung);

d) **Interessen der Allgemeinheit**; und

e) **Machtverteilung** zwischen den Akteuren.

In ihrer Zielrichtung können Datenrechte insbesondere gerichtet sein auf

- eine **Unterlassung** der Datennutzung (bis hin zur Löschungspflicht);
- eine **Korrektur** von Daten;
- **Zugang** zu Daten (bis hin zu Portabilität); oder
- wirtschaftliche **Teilhabe**.

Für jede dieser Ausprägungen gelten jeweils eigene **Konkretisierungen**. Dabei kommt es nach Auffassung der DEK etwa bei Unterlassungs-Verlangen maßgeblich auf das Schädigungspotenzial einer Datennutzung sowie auf die Umstände an, unter denen der Beitrag zur Datengenerierung geleistet wurde. Auch für Korrektur-Verlangen kann das Schädigungspotenzial relevant sein, doch sind die Anforderungen geringer. Bei Zugangs-Verlangen eines Akteurs gilt ein abgestuftes Spektrum berechtigter Zugangsinteressen, die insbesondere in bestehenden Wertschöpfungssystemen zum Tragen kommen. Eigenständige Rechte einer Person auf wirtschaftliche Teilhabe an der Wertschöpfung, die andere mit Daten betreiben, kommen dagegen nur unter extrem engen Voraussetzungen in Betracht. Die **Betroffenenrechte** der Datenschutz-Grundverordnung (DSGVO) sind eine besonders wichtige und – weil einheitlich an der Qualifikation von Daten als personenbezogen anknüpfend – in gewisser Weise typisierte Ausprägung dieser Grundsätze speziell zum Schutz derjenigen natürlichen Person, auf die sich die Information bezieht.

Unter Berücksichtigung dieser Grundsätze gelangt die DEK zusammenfassend zu den folgenden zentralen Handlungsempfehlungen:

Anforderungen an die Nutzung personenbezogener Daten

1

Die DEK empfiehlt **Maßnahmen gegen ethisch nicht-vertretbare Datennutzungen**. Dazu gehören etwa Totalüberwachung, die Integrität der Persönlichkeit verletzende Profilbildung, gezielte Ausnutzung von Vulnerabilitäten, sog. Addictive Designs und Dark Patterns, dem Demokratieprinzip zuwiderlaufende Beeinflussung politischer Wahlen, Lock-in und systematische Schädigung von Verbrauchern sowie viele Formen des Handels mit personenbezogenen Daten.

2

Sowohl das Datenschutzrecht als auch die übrige Rechtsordnung (u. a. Zivilrecht, Lauterkeitsrecht) enthalten bereits eine Fülle von Instrumenten, die gegen derartige Datennutzungen eingesetzt werden können. Gemessen an Breitenwirkung und Schädigungspotenzial werden diese Instrumente indessen bislang nicht in ausreichender Weise genutzt – insbesondere gegenüber marktmächtigen Unternehmen. Dieses **Vollzugsdefizit** hat verschiedene Ursachen, die es systematisch anzugehen gilt.

3

Neben der Schärfung des Bewusstseins bei handelnden Akteuren (z. B. Aufsichtsbehörden) für die bereits bestehenden Möglichkeiten ist dringend eine **Konkretisierung und punktuelle Verschärfung des geltenden Rechtsrahmens** angezeigt. Dazu gehören etwa eine spezielle Normierung von datenspezifischen Klauselverboten, Schutz- und Treuepflichten, Deliktstatbeständen und unlauteren Geschäftspraktiken sowie die Schaffung eines weitaus konkreteren Rechtsrahmens für Profilbildungen und Scoring wie auch für den Datenhandel.

4

Um die Wirkungskraft der Aufsichtsbehörden zu erhöhen, bedürfen diese einer weitaus besseren personellen und sachlichen Ausstattung. Sofern es nicht gelingt, die Abstimmung unter den deutschen Datenschutzaufsichtsbehörden zu verstärken und zu formalisieren und so die einheitliche und kohärente Anwendung des Datenschutzrechts zu gewährleisten, ist eine **Zentralisierung der Datenschutzaufsicht für den Markt** in einer – mit einem weiten Mandat ausgestatteten und eng mit anderen Fachaufsichtsbehörden kooperierenden – Behörde auf Bundesebene zu erwägen. Die Zuständigkeit der Landesdatenschutzbehörden für den öffentlichen Bereich soll hingegen unangetastet bleiben.

5

Die Anerkennung von **„Dateneigentum“** im Sinne eines dem Sacheigentum oder dem geistigen Eigentum nachgebildeten Ausschließlichkeitsrechts an Daten würde nach Auffassung der DEK bestehende Probleme nicht lösen und stattdessen eine Reihe neuer Probleme schaffen. Sie wird daher **nicht empfohlen**. Die DEK empfiehlt auch nicht die Anerkennung genereller wirtschaftlicher Verwertungsrechte an personenbezogenen Daten, wie sie etwa durch Verwertungsgesellschaften geltend gemacht werden könnten.

6

Wenngleich die plakative Bezeichnung zur allgemeinen Bewusstseinsbildung beigetragen hat, plädiert die DEK dafür, **von der Bezeichnung von Daten als „Gegenleistung“ abzusehen**. Unabhängig von der künftigen Auslegung des sog. Koppelungsverbots durch die Aufsichtsbehörden und den EuGH fordert die DEK, dass Verbrauchern jeweils **zumutbare Alternativen** gegenüber der Freigabe von Daten zur auch kommerziellen Nutzung angeboten werden müssen (z. B. entsprechend ausgestaltete **Bezahlmodelle**).

7

Die Verwendung von Daten zur **personalisierten Risikoeinschätzung** (z. B. im Rahmen von Telematiktarifen bei bestimmten Versicherungen) sollte an **enge Voraussetzungen** geknüpft werden. So darf die Datenverarbeitung beispielsweise nicht den Kern privater Lebensführung betreffen, es muss ein klarer ursächlicher Zusammenhang zwischen Daten und Risiko vorliegen, und die Preisdifferenz zwischen personalisiertem und nicht personalisiertem Tarif sollte im Einzelnen noch festzulegende Prozentwerte nicht überschreiten. Weitere Anforderungen betreffen Transparenz, Nichtdiskriminierung und den Schutz dritter Personen.

8

Die DEK empfiehlt der Bundesregierung, Fragen rund um den **„digitalen Nachlass“** mit dem Urteil des BGH von 2018 nicht als erledigt anzusehen. Die praktisch lückenlose Aufzeichnung von digital geführter Kommunikation, die in vielen Fällen an die Stelle des flüchtig gesprochenen Wortes tritt, und ihre Aushändigung an Erben bedeutet eine neue Dimension von Gefährdung für die Privatheit. Ihr sollte mit einer Reihe von Maßnahmen begegnet werden, welche neue Pflichten von Diensteanbietern, Qualitätssicherung bei Angeboten digitaler Nachlassplanung sowie nationale Regelungen zum postmortalen Datenschutz umfassen.

9

Die DEK empfiehlt der Bundesregierung, die Sozialpartner einzuladen, ausgehend von den bereits in Tarifverträgen bestehenden Beispielen guter Übung eine gemeinsame Linie für gesetzliche Konkretisierungen des **Beschäftigtendatenschutzes** zu entwickeln. Dabei sollten auch die Belange von Personen in unüblichen Beschäftigungsformen berücksichtigt werden.

10

Mit Blick auf die Vorteile eines **digitalisierten Gesundheitswesens** spricht sich die DEK für einen raschen Ausbau digitaler Infrastrukturen innerhalb des Gesundheitssektors aus. Der qualitative und quantitative Ausbau digitalisierter Versorgungsmaßnahmen sollte die informationelle Selbstbestimmung des Patienten stärken. Hierzu gehört der partizipative Auf- und Ausbau der elektronischen Patientenakte (ePA) sowie die Weiterentwicklung von Verfahren zur Prüfung und Bewertung digitaler Gesundheitsanwendungen im ersten und zweiten Gesundheitsmarkt.

11

Die DEK fordert, dem erheblichen Vollzugsdefizit des geltenden Rechts betreffend den **Schutz von Kindern und Jugendlichen** im digitalen Raum abzuhelfen. Insbesondere sollten Technologien – einschließlich eines effektiven Identitätenmanagements – sowie Standardoptionen entwickelt und verpflichtend vorgesehen werden, welche einen zuverlässigen Schutz der Kinder und Jugendlichen gewährleisten und zugleich familienadäquat sind, indem sie Erziehungsberechtigte weder überfordern noch eine übermäßige Überwachung im privaten Bereich ermöglichen oder gar hierzu animieren.

12

Was den Umgang mit Daten **pflege- und schutzbedürftiger Menschen** betrifft, sollte für professionelle Akteure im Pflegebereich durch Standards und Leitlinien mehr Rechtssicherheit geschaffen werden. Zugleich ist eine gesetzliche Klarstellung zu erwägen, dass – soweit eine Datenverarbeitung auf die Einwilligung des pflege- und schutzbedürftigen Menschen gestützt werden muss – in Patientenverfügungen auch bestimmte Dispositionen in Bezug auf die Datenverarbeitung (z. B. für den Fall der dauernden Einwilligungsunfähigkeit infolge von Demenz) getroffen werden können.

13

Die DEK empfiehlt, eine Reihe verbindlicher Vorgaben für **datenschutzfreundliches Design von Produkten und Dienstleistungen** einzuführen und damit die an Verantwortliche im Sinne der DSGVO gerichteten Vorgaben von Datenschutz „by design" und „by default" bereits auf der Ebene der Hersteller wie auch der Diensteanbieter wirksam werden zu lassen. Dies betrifft insbesondere Vorgaben für Verbraucherendgeräte. In diesem Zusammenhang sind auch einheitliche Bildsymbole (Piktogramme) einzuführen, die dem Verbraucher eine informierte Kaufentscheidung ermöglichen.

14

Ferner bedarf es einer Reihe weiterer Maßnahmen auf verschiedenen Ebenen, um für Hersteller effektive **Anreize zur Implementierung eines datenschutzfreundlichen Designs** zu schaffen. Neben wirksamen Rechtsbehelfen entlang der Vertriebskette, mit deren Hilfe Hersteller mit in die Verantwortung für unzureichenden Datenschutz „by design" und „by default" genommen werden können, ist insbesondere an Vorgaben in Ausschreibungsbedingungen und Beschaffungsrichtlinien für die öffentliche Hand sowie an Bedingungen bei Förderprogrammen zu denken. Das Gleiche gilt für datenschutzfreundliche **Methoden der Produktentwicklung**, einschließlich des Trainierens algorithmischer Systeme.

15

Trotz des berechtigten Fokus auf Datenschutz natürlicher Personen darf der **Schutzbedarf von Unternehmen und juristischen Personen** nicht in den Hintergrund treten. Durch die umfassende Verknüpfbarkeit von Einzeldaten kann ein lückenloses Bild interner Betriebsabläufe entstehen und in die Hände von Konkurrenten, Verhandlungspartnern, Übernahmeinteressenten usw. gelangen. Dies stellt aufgrund umfangreicher Datenflüsse in Drittstaaten u. a. eine Gefährdung der digitalen Souveränität Deutschlands und Europas dar. Viele Handlungsempfehlungen sind daher sinngemäß auch auf die Daten juristischer Personen zu übertragen. Die DEK fordert die Bundesregierung auf, Schritte zu unternehmen, um den **datenbezogenen Schutz von Unternehmen zu verbessern**.

Verbesserung des kontrollierten Zugangs zu personenbezogenen Daten

16

Die DEK sieht in einer Datennutzung für gemeinwohlorientierte Forschungszwecke (z. B. zur Verbesserung der Gesundheitsfürsorge) enormes Potenzial, das es zum Wohle des Einzelnen und der Allgemeinheit zu nutzen gilt. Das geltende Datenschutzrecht erkennt dieses Potenzial durch eine Reihe weitreichender Privilegierungen prinzipiell an. Allerdings bestehen auch Unsicherheiten, insbesondere mit Blick auf die Reichweite des sog. Weiterverarbeitungsprivilegs sowie des Forschungsbegriffs im Zusammenhang mit der Entwicklung von Produkten. Dem muss aus Sicht der DEK durch entsprechende **gesetzliche Klarstellungen** begegnet werden.

17

Die Zersplitterung der Rechtslage, sowohl innerhalb Deutschlands als auch der EU Mitgliedstaaten untereinander, kann ein Hindernis für datengetriebene Forschung darstellen. Empfohlen wird daher eine **Harmonisierung der forschungsspezifischen Regelungen** sowohl auf Bundes- und Landesebene als auch der verschiedenen nationalen Regelungen innerhalb der EU. Auch die Einführung eines Notifizierungsverfahrens für mitgliedstaatliche Regelungen zum Forschungsdatenschutz sowie die Einrichtung einer europäischen Clearing-Stelle für grenzüberschreitende Forschungsprojekte könnte eine Erleichterung bringen.

18

Bei Forschung mit besonders sensiblen Kategorien personenbezogener Daten (z. B. Gesundheitsdaten) sollten Forschende durch **Handreichungen** zur rechtssicheren Einholung von Einwilligungen sowie durch die Förderung und gesetzliche **Anerkennung innovativer Einwilligungsmodelle** unterstützt werden. Zusätzlich zu den weiteren Entwicklungen zur Reichweite des sog. Weiterverarbeitungsprivilegs für die Forschung könnten dazu auch digitale Einwilligungsassistenten oder ein sog. Meta Consent gehören.

19

Die DEK unterstützt prinzipiell die Entwicklung in Richtung eines **„lernenden Gesundheitssystems“**, in dem die Daten aus der alltäglichen Gesundheitsversorgung systematisch und qualitätsgestützt im Sinne der evidenzbasierten Medizin genutzt werden, um die Versorgung kontinuierlich zu verbessern. Allerdings sollte flankierend, beispielsweise durch **Verwertungsverbote**, mehr Schutz vor dem erheblichen Diskriminierungspotenzial sensibler Datenkategorien geschaffen werden.

20

Im Zentrum aller Bemühungen um eine Verbesserung des kontrollierten Zugangs zu (ursprünglich) personenbezogenen Daten steht die Entwicklung von Verfahren und Standards der **Anonymisierung** und **Pseudonymisierung**. Durch rechtliche Vermutungen, dass bei Einhaltung des Standards kein Personenbezug mehr gegeben ist bzw. dass „geeignete Garantien“ für die Rechte betroffener Personen vorliegen, könnte die Rechtssicherheit deutlich verbessert werden. Diese Maßnahmen sollten flankiert werden durch strafbewehrte Verbote einer De-Anonymisierung (für den Fall, dass bei bisher anonymen Daten, etwa durch die Entwicklung der Technik, ein Personenbezug hergestellt werden kann) bzw. der Aufhebung der Pseudonymisierung jenseits eng definierter Rechtfertigungsgründe. Auch die Forschung im Bereich **synthetischer Daten** ist vielversprechend und sollte weiter gefördert werden.

21

Großes Potenzial sieht die DEK grundsätzlich auch in **innovativen Datenmanagement- und Datentreuhandsystemen**, sofern diese praxisgerecht, robust und datenschutzkonform ausgestaltet sind. Solche Modelle rangieren von rein technischen Dashboards (**Privacy Management Tools,** PMT) bis hin zu umfassenden Dienstleistungen der Daten- und Einwilligungsverwaltung (**Personal Information Management Services,** PIMS). Ziel ist die Befähigung des Einzelnen zur Kontrolle über seine personenbezogenen Daten sowie die Entlastung des Einzelnen von Entscheidungen, die ihn überfordern. Die DEK empfiehlt, Forschung und Entwicklung im Bereich von Datenmanagement- und Datentreuhandsystemen intensiv zu fördern, mahnt aber auch an, dass eine die Rechte und Interessen aller Beteiligten wahrende Entwicklung ohne eine **begleitende europäische Regulierung** nicht zu erwarten ist. Diese Regulierung müsste zentrale Funktionen absichern, ohne die Betreiber solcher Systeme nur sehr eingeschränkt tätig werden können. Andererseits geht es um den Schutz des Einzelnen vor vermeintlichen Interessenwaltern, die in Wahrheit vorrangig wirtschaftliche Eigeninteressen oder Interessen Dritter vertreten. Sofern dieser Schutz auch in der Praxis garantiert werden kann, kann Datentreuhandmodellen die Funktion einer wichtigen Schnittstelle zwischen Belangen des Datenschutzes und der Datenwirtschaft zukommen.

22

In Bezug auf das Recht auf **Datenportabilität** aus Art. 20 DSGVO empfiehlt die DEK die Erarbeitung branchenbezogener Verhaltensregeln und Standards betreffend Datenformate. Soweit Art. 20 DSGVO nicht nur Anbieterwechsel erleichtern, sondern auch den Datenzugang für andere Anbieter verbessern soll, empfiehlt sich eine sorgfältige Evaluierung, wie sich das bestehende Portabilitätsrecht auf den Markt auswirkt und wie eine zunehmende Stärkung der Marktmacht weniger Anbieter verhindert werden kann. Bevor die Ergebnisse einer solchen Evaluierung vorliegen, sollte von einer vorschnellen Erweiterung des Portabilitätsrechts, etwa auf andere als bereitgestellte Daten oder auf Portierung in Echtzeit, abgesehen werden.

23

Eine **Pflicht zur Interoperabilität bzw. Interkonnektivität** in bestimmten Sektoren – etwa bei Messenger-Diensten und sozialen Netzwerken – könnte dazu beitragen, Markteintrittsbarrieren für neue Anbieter zu senken. Für eine solche Pflicht würde sich eine asymmetrische, d. h. nach Marktmacht gestaffelte Regulierung empfehlen. Dies wäre auch eine Voraussetzung dafür, bestimmte Basisdienstleistungen der Informationsgesellschaft in Europa neu aufzubauen bzw. zu stärken.

Datenzugangsdebatten jenseits des Personenbezugs

24

Für die Entwicklung der europäischen Datenwirtschaft sieht die DEK einen zentralen Faktor im Zugang europäischer Unternehmen zu geeigneten nicht-personenbezogenen Daten in geeigneter Qualität. **Datenzugang** nutzt allerdings nur Akteuren, die ein entsprechendes Bewusstsein für die Bedeutung von Daten haben und über entsprechende Datenkompetenz verfügen, und in ganz überproportionalem Ausmaß denjenigen, bei denen bereits der größte Ausgangsbestand an Daten und die besten Dateninfrastrukturen vorhanden sind. Die DEK empfiehlt daher, bei der Diskussion um eine Verbesserung des Datenzugangs stets die genannten Faktoren gemäß dem **ASISA-Prinzip** *(Awareness – Skills – Infrastructures – Stocks – Access)* mit zu berücksichtigen.

25

Daher unterstützt die DEK die bereits auf europäischer Ebene begonnenen Maßnahmen zur Förderung von **Dateninfrastrukturen** im weitesten Sinne (z. B. Plattformen, Standards für Programmierschnittstellen und weitere Elemente, Modellverträge, EU-Unterstützungszentrum) und empfiehlt der Bundesregierung, diese weiterhin durch entsprechende Bemühungen auf nationaler Ebene zu flankieren. In diesem Zusammenhang bietet sich die Einrichtung einer Ombudsstelle auf Bundesebene an, welche bei Aushandlung von Datenzugangsvereinbarungen und bei Streitigkeiten hilft und vermittelt.

26

Die DEK sieht einen Schlüsselfaktor in einer holistisch gedachten, nachhaltigen und strategischen **Wirtschaftspolitik**, welche der Abwanderung innovativer europäischer Unternehmen bzw. deren Aufkauf durch Akteure aus Drittstaaten ebenso effektiv entgegenwirkt wie der übermäßigen Abhängigkeit von Infrastrukturen (z. B. Serverkapazitäten) in Drittstaaten. Dabei ist die richtige Balance zu finden zwischen gewollter internationaler Kooperation und Vernetzung einerseits und andererseits der entschlossenen Übernahme von Verantwortung für nachhaltige Sicherheit und Wohlfahrt in Europa vor dem Hintergrund sich wandelnder globaler Machtverhältnisse.

27

Die DEK sieht auch unter dem Blickwinkel einer Förderung der Datenwirtschaft keinen Bedarf nach der Einführung neuer Ausschließlichkeitsrechte („Dateneigentum", „Datenerzeugerrecht"), sondern empfiehlt stattdessen eine **beschränkte Drittwirkung vertraglicher Vereinbarungen** (z. B. betreffend Beschränkungen der Nutzung und Weitergabe von Daten) nach dem Vorbild des neuen europäischen Regimes zum Schutz von Geschäftsgeheimnissen. Ferner wäre es wünschenswert, wenn gesetzlich Wege aufgezeigt würden, wie europäische Unternehmen – etwa unter Einschaltung von Treuhändern – unter voller Wahrung kartellrechtlicher Belange bei der Datennutzung kooperieren können **(„Datenpartnerschaften")**.

28

In bestehenden Wertschöpfungssystemen (z. B. Produktions- und Vertriebsketten) fallen vielfach Daten an, die innerhalb wie außerhalb des Wertschöpfungssystems von enormer wirtschaftlicher Bedeutung sind. Die zwischen den einzelnen Teilnehmern eines Wertschöpfungssystems bestehenden Verträge enthalten aber häufig entweder keine bzw. eine unfaire und/oder ineffiziente Regelung des Datenzugangs, oder es fehlt ganz an einer vertraglichen Vereinbarung. Weit über die klassische „Datenwirtschaft" hinaus ist daher **Bewusstseinsbildung bei Wirtschaftstreibenden** erforderlich, die durch praktische Hilfestellungen (z. B. Modellverträge) ergänzt werden sollte.

29

Darüber hinaus regt die DEK eine **behutsame Ergänzung des geltenden Rechtsrahmens** an. Dabei sollte ein erster Schritt darin liegen, die Sonderbeziehung zwischen einer Partei, welche zur Generierung von Daten in einem Wertschöpfungssystem beigetragen hat, und der Partei, welche die Daten faktisch kontrolliert, in § 311 BGB explizit anzuführen. Unter anderem sollte die Aufnahme von Vertragsverhandlungen über ein faires und effizientes Datenzugangsregime Bestandteil einer solchen allgemeinen Treuepflicht sein. Im Übrigen sollte geprüft werden, ob darüber hinaus Maßnahmen erforderlich sind, welche von punktuellen Klauselverboten in B2B-Geschäften über ein dispositives Datenschuldrecht bis zu sektorspezifischen Datenzugangsrechten rangieren könnten.

30

Die DEK sieht großes Potenzial in **Konzepten offener Daten des öffentlichen Sektors** (Open Government Data, OGD) und empfiehlt, solche Konzepte auszubauen und zu fördern. Sie empfiehlt eine Reihe von Maßnahmen, die einen teilweise noch nicht ganz vollzogenen **Bewusstseinswandel öffentlicher Stellen** befördern und das Teilen von Daten im Rahmen von OGD-Konzepten praktisch erleichtern könnten. Dazu gehört neben der Etablierung entsprechender **Infrastrukturen** (z. B. Plattformen) auch eine Harmonisierung und punktuelle Ergänzung des derzeit zersplitterten und nicht in jeder Hinsicht konsistenten **Rechtsrahmens**.

31

Allerdings sieht die DEK auch ein schwer zu lösendes Spannungsverhältnis zwischen der Diskussion um OGD (mit Prinzipien wie „offen by default" und „offen für alle Zwecke") einerseits und um besseren Schutz von Geschäftsgeheimnissen und personenbezogenen Daten (mit gesetzlichen Vorgaben wie „Datenschutz by default") andererseits. Sie plädiert dafür, in Zweifelsfällen zugunsten des staatlichen Schutzauftrags zu entscheiden, der in Bezug auf Daten, welche Einzelne oder Unternehmen dem Staat – oft nicht freiwillig – anvertraut haben (z. B. Steuerdaten), besteht. Diesem **staatlichen Schutzauftrag** ist durch eine Reihe von Maßnahmen nachzukommen, die auch technische und rechtliche Schutzvorkehrungen gegen Missbrauch umfassen.

32

In diesem Zusammenhang wird insbesondere empfohlen, für das Teilen von Daten durch den öffentlichen Sektor **Standardlizenzen und Modellkonditionen** zu entwickeln und – mindestens sektorspezifisch – deren Verwendung bindend vorzuschreiben. Diese sollten klar definierte Garantien für die Rechte betroffener Dritter enthalten. Ferner sollten sie Mechanismen vorsehen, die geeignet sind, eine gemeinwohlschädigende Nutzung der Daten ebenso zu verhindern wie eine wettbewerbsrechtlich unerwünschte Verstärkung bestehender Marktmacht oder eine Doppelbelastung des Steuerzahlers.

33

Betreffend **Konzepte offener Daten im privaten Sektor** sollte in erster Linie auf die **Ermutigung und Förderung eines freiwilligen Teilens** von Daten gesetzt werden. Dabei ist nicht nur an Infrastrukturen (z. B. Plattformen) zu denken, sondern auch an eine breite Palette möglicher Anreizstrukturen, etwa bei der Besteuerung, bei öffentlichen Ausschreibungen, bei Förderprogrammen oder bei Genehmigungsverfahren. Gesetzliche Datenzugangsrechte und korrespondierende Zugangsgewährungspflichten sollten dagegen erst in zweiter Linie in Betracht gezogen werden.

34

Insgesamt rät die DEK bei allgemeinen gesetzlichen Datenzugangsrechten zu einem behutsamen Vorgehen, idealerweise **zunächst in ausgewählten Sektoren**. Beispielsweise könnte ein Bedarf im Nachrichten-, Mobilitäts- oder Energiesektor geprüft werden. Dabei sind jeweils alle möglichen Konsequenzen einer Zugangsgewährungs- oder gar Offenlegungspflicht sorgsam zu bedenken und gegeneinander abzuwägen, angefangen von möglichen Implikationen für den Datenschutz und Schutz von Geschäftsgeheimnissen, über Folgen für Investitionsentscheidungen und die Verteilung von Marktmacht bis hin zu den strategischen Interessen deutscher und europäischer Unternehmen im Verhältnis zu Unternehmen in Drittstaaten.

35

Die DEK empfiehlt, Zugangsgewährungspflichten privater Unternehmen **zugunsten gemeinwohlorientierter Zwecke und des öffentlichen Sektors** (Business-to-Government, B2G) in Erwägung zu ziehen. Auch diesbezüglich dürfte indessen ein behutsames und sektorspezifisches Vorgehen anzuraten sein.

3

Algorithmische Systeme

Die primär auf algorithmische Systeme ausgerichtete Perspektive **(Algorithmen-Perspektive)** richtet den Blick auf die Architektur und Dynamik des datenverarbeitenden algorithmischen Systems und seine Auswirkungen auf Einzelne und die Gesellschaft. Der ethische und rechtliche Diskurs fokussiert dabei typischerweise auf die Beziehung von Mensch und Maschine und mit Blick auf Künstliche Intelligenz (KI) insbesondere auf die Automatisierung sowie auf die Verlagerung auch komplexer Handlungs- und Entscheidungsprozesse auf sog. autonome Systeme. In Abgrenzung zur Daten-Perspektive müssen die vom System betroffenen Personen nicht notwendig auch etwas mit den Daten zu tun haben, die das System verarbeitet – insbesondere können sich ethisch nicht vertretbare Auswirkungen auf Einzelne auch dann ergeben, wenn ausschließlich nicht-personenbezogene Daten genutzt wurden (z. B. für das Training eines algorithmischen Systems). Eine zentrale aktuelle Debatte, die hier zu verorten ist, ist diejenige um eine „Algorithmenkontrolle“ oder um die Haftung für KI.

Allgemeine Anforderungen an algorithmische Systeme

Die DEK unterscheidet je nach der konkreten Aufgabenverteilung zwischen menschlichem Akteur und Maschine drei unterschiedliche Stufen des Einbezugs von algorithmischen Systemen in menschliche Entscheidungen:

a) **algorithmenbasierte** Entscheidungen sind menschliche Entscheidungen, die sich auf algorithmisch berechnete (Teil-)Informationen stützen;

b) **algorithmengetriebene** Entscheidungen sind menschliche Entscheidungen, die durch die Ergebnisse algorithmischer Systeme in einer Weise geprägt werden, dass der tatsächliche Entscheidungsspielraum und damit die Selbstbestimmung des Menschen eingeschränkt werden;

c) **algorithmendeterminierte** Entscheidungen führen automatisiert zu Konsequenzen, so dass im Einzelfall keine menschliche Entscheidung mehr vorgesehen ist.

Ein verantwortungsvoller Umgang mit algorithmischen Systemen sollte sich nach Auffassung der DEK an folgenden Grundsätzen orientieren:

- **Menschenzentriertes Design:** Systeme müssen den Menschen, der die Systeme anwendet oder von ihren Entscheidungen betroffen ist, seine grundlegenden Rechte und Freiheiten, sein körperliches und emotionales Wohlbefinden, seine Kompetenzentwicklung und seine Grundbedürfnisse in den Mittelpunkt stellen.

- **Vereinbarkeit mit gesellschaftlichen Grundwerten:** Bei der Gestaltung von Systemen sind Auswirkungen gesamtgesellschaftlicher Relevanz zu berücksichtigen, insbesondere auf die demokratische Willensbildung, die Bürgernähe staatlichen Handelns, den Wettbewerb, die Zukunft der Arbeit und die digitale Souveränität Deutschlands und Europas.

- **Nachhaltigkeit:** Bei der Gestaltung und dem Einsatz algorithmischer Systeme erhalten Aspekte der Verfügbarkeit menschlicher Kompetenzen, der Partizipation, des Umweltschutzes und der nachhaltigen Ressourcenbewirtschaftung sowie des nachhaltigen wirtschaftlichen Handelns wachsende Bedeutung.

- **Qualität und Leistungsfähigkeit:** Algorithmische Systeme müssen korrekt und zuverlässig funktionieren, um die mit ihrer Hilfe verfolgten Zwecke zu erreichen.
- **Robustheit und Sicherheit:** Robuste und sichere Systemgestaltung umfasst sowohl die Sicherheit des Systems gegen Einflüsse von außen als auch den Schutz der Menschen und der Umwelt vor negativen Einflüssen durch das System.
- **Minimierung von Verzerrungen und Diskriminierung:** Die Entscheidungsmuster, die algorithmischen Systemen zugrunde liegen, dürfen keine systematischen Verzerrungen (Biases) aufweisen oder zu diskriminierenden Entscheidungen führen.
- **Transparenz, Erklärbarkeit und Nachvollziehbarkeit:** Es ist essenziell, dass sowohl die Anwender der algorithmischen Systeme deren Funktionsweise verstehen, erklären und kontrollieren können, als auch, dass die von einer Entscheidung Betroffenen genügend Informationen erhalten, um ihre Rechte angemessen wahrnehmen und die Entscheidung infrage stellen zu können.
- **Klare Rechenschaftsstrukturen:** Der Einsatz algorithmischer Systeme verlangt eine klare Zuordnung von Verantwortung und Rechenschaftspflichten einschließlich einer möglichen Haftung.

Systemkritikalität

Die konkret an ein algorithmisches System zu stellenden Anforderungen – insbesondere auch im Hinblick auf Transparenz und Kontrolle – sind abhängig von der **Systemkritikalität**. Die Systemkritikalität setzt am Schädigungspotenzial des algorithmischen Systems an. Dabei bedeutet Schädigungspotenzial die Kombination aus der **Wahrscheinlichkeit eines Schadenseintritts** und der **Schwere des zu befürchtenden Schadens**.

Die **Schwere** zu befürchtender Schäden, etwa im Falle einer Fehlentscheidung, bezieht sich auf die Wertigkeit der betroffenen Rechtsgüter und Interessen (z. B. Recht auf Privatheit, Grundrecht auf Leben und körperliche Unversehrtheit, Diskriminierungsverbot), die Höhe eines möglichen Schadens für Einzelne (einschließlich immaterieller Schäden bzw. monetär schwer zu beziffernder Nutzeneinbußen), die Zahl der Betroffenen, die Summe der potenziellen Schäden und den gesamtgesellschaftlichen Schaden, der über eine reine Summierung von Einzelschäden weit hinausgehen kann. Die **Wahrscheinlichkeit** eines Schadenseintritts hängt auch von den konkreten Systemeigenschaften ab – insbesondere von der Rolle algorithmischer Systemkomponenten im Entscheidungsprozess, der Komplexität der Entscheidung, den Wirkungen der Entscheidung und der Reversibilität der Wirkungen. Schwere und Wahrscheinlichkeit zu befürchtender Schäden können zudem abhängig sein vom staatlichen oder privaten Charakter des Handelns und – gerade in wirtschaftlichen Zusammenhängen – von der Marktmacht desjenigen Akteurs, der sich des algorithmischen Systems bedient.

Unter Berücksichtigung dieser Grundsätze gelangt die DEK zusammenfassend zu den folgenden Handlungsempfehlungen:

Empfehlung eines risikoadaptierten Regulierungsansatzes

36

Die DEK empfiehlt einen **risikoadaptierten Regulierungsansatz** für algorithmische Systeme. Er sollte auf dem Grundsatz aufbauen, dass ein steigendes Schädigungspotenzial mit wachsenden Anforderungen und Eingriffstiefen der regulatorischen Instrumente einhergeht. Für die Beurteilung kommt es jeweils auf das **gesamte sozio-technische System** an, also alle Komponenten einer algorithmischen Anwendung einschließlich aller menschlichen Akteure, von der Entwicklungsphase (z. B. hinsichtlich der verwendeten Trainingsdaten) bis hin zur Implementierung in einer Anwendungsumgebung und zur Phase von Bewertung und Korrektur.

37

Die DEK empfiehlt, die Bestimmung des Schädigungspotenzials algorithmischer Systeme für Einzelne und/oder die Gesellschaft anhand eines **übergreifenden Modells** einheitlich vorzunehmen. Dafür sollte der Gesetzgeber mit Hilfe von **Kriterien** ein Prüfschema definieren, nach welchem die Kritikalität algorithmischer Systeme auf der Grundlage der von der DEK vorgestellten allgemeinen ethischen und rechtlichen Grundsätze und Prinzipien zu bestimmen ist.

38

Regulatorische Instrumente und **Anforderungen** an algorithmische Systeme sollten u. a. Korrektur- und Kontrollinstrumente, Vorgaben für die Transparenz, die Erklärbarkeit und die Nachvollziehbarkeit der Ergebnisse sowie Regelungen zur Zuordnung von Verantwortlichkeit und Haftung für den Einsatz umfassen.

39

Die DEK erachtet es als sinnvoll, mit Blick auf das Schädigungspotenzial algorithmischer Systeme in einem ersten Schritt **fünf Kritikalitäts-Stufen** zu unterscheiden. Auf der untersten Stufe (Stufe 1) von Anwendungen ohne oder mit geringem Schädigungspotenzial besteht keine Notwendigkeit einer besonderen Kontrolle oder von Anforderungen, die über die allgemeinen Qualitätsanforderungen, welche auch für Produkte ohne algorithmische Elemente gelten, hinausgehen.

40

Bei Anwendungen mit einem **gewissen Schädigungspotenzial** (Stufe 2) kann und soll bedarfsgerechte Regulierung einsetzen, wie etwa Ex-post-Kontrollen, die Pflicht zur Erstellung und Veröffentlichung einer angemessenen Risikofolgenabschätzung, Offenlegungspflichten gegenüber Aufsichtsinstitutionen oder auch gesteigerte Transparenzpflichten sowie Auskunftsrechte für Betroffene.

41

Bei Anwendungen mit **regelmäßigem** oder **deutlichem Schädigungspotenzial** (Stufe 3) können zusätzlich Zulassungsverfahren gerechtfertigt sein. Bei Anwendungen mit **erheblichem Schädigungspotenzial** (Stufe 4) fordert die DEK darüber hinaus verschärfte Kontroll- und Transparenzpflichten bis hin zu einer Veröffentlichung der in die algorithmische Berechnung einfließenden Faktoren und deren Gewichtung, der Datengrundlage und des algorithmischen Entscheidungsmodells sowie die Möglichkeit einer kontinuierlichen behördlichen Kontrolle über eine Live-Schnittstelle zum System.

42

Bei **Anwendungen mit unvertretbarem Schädigungspotenzial** (Stufe 5) ist schließlich ein vollständiges oder teilweises **Verbot** auszusprechen.

43

Zur Umsetzung der durch die DEK vorgeschlagenen Maßnahmen empfiehlt die DEK eine Regulierung algorithmischer Systeme durch allgemeine **horizontale Vorgaben im Recht** der Europäischen Union **(Verordnung für Algorithmische Systeme, EUVAS)**. Dieser horizontale Rechtsakt sollte die zentralen Grundprinzipien für algorithmische Systeme enthalten, wie sie die DEK als Anforderungen an algorithmische Systeme entwickelt hat. Insbesondere sollte er im Lichte der Systemkritikalität allgemeine materielle Regelungen zur Zulässigkeit und Gestaltung algorithmischer Systeme, zur Transparenz, zu Betroffenenrechten, zu organisatorischen und technischen Absicherungen und zu den Institutionen und Strukturen der Aufsicht bündeln. Der horizontale Rechtsakt sollte auf der Ebene der EU und der Mitgliedstaaten eine **sektorale Konkretisierung erfahren**, die wiederum am Gedanken der Systemkritikalität orientiert ist.

44

Im Zuge der hier empfohlenen Entwicklung einer EUVAS sollte die Aufgabenverteilung zwischen dieser Regulierung und der **DSGVO** überdacht werden. Dabei ist zum einen zu berücksichtigen, dass sich spezifische Risiken

algorithmischer Systeme für den Einzelnen und für Gruppen auch dann manifestieren können, wenn keine personenbezogenen Daten verarbeitet werden, und dass die Risiken nicht unbedingt solche des Datenschutzes sind, wenn sie etwa das Vermögen, Eigentum, körperliche Integrität oder Diskriminierung betreffen. Zum anderen ist zu bedenken, dass für eine künftige horizontale Regulierung algorithmischer Systeme ein flexibleres, stärker risikoadaptiertes Regulierungsregime als für den Datenschutz in Betracht gezogen werden sollte.

Instrumente

45

Die DEK empfiehlt bei algorithmischen Systemen erhöhter Systemkritikalität (ab Stufe 2) eine **Kennzeichnungspflicht**: Eine solche Pflicht trägt Betreibern auf, deutlich zu machen, wann und in welchem Umfang algorithmische Systeme zum Einsatz kommen (Information über das „Ob"). Eine Kennzeichnungspflicht sollte unabhängig von der Systemkritikalität stets im Falle einer ethisch relevanten Verwechselungsgefahr zwischen Mensch und algorithmischem System bestehen.

46

Das Recht einer betroffenen Person auf aussagekräftige **Informationen** über die „involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen" eines algorithmischen Systems (vgl. DSGVO) sollte nicht nur für vollständig automatisierte Systeme, sondern bereits für **Profilbildungen als solche** und unabhängig von einer nachgelagerten Entscheidungssituation bestehen. Es sollte – abgestuft nach der Systemkritikalität – künftig auch bereits für algorithmenbasierte Entscheidungen greifen. Dazu sollte teilweise eine gesetzliche Klarstellung und teilweise eine Erweiterung der Regelung auf europäischer Ebene erfolgen.

47

In bestimmten Bereichen kann es sachgerecht sein, dem Betreiber algorithmischer Systeme zusätzlich zur allgemeinen Erläuterung der Logik (Vorgehensweise) und Tragweite des Systems eine **individuelle Erklärung** der getroffenen Entscheidung abzuverlangen. Wesentlich ist dabei, dass betroffene Personen verständlich, relevant und konkret informiert werden. Die DEK begrüßt daher die technischen Bemühungen, die Erklärbarkeit algorithmischer (insbesondere selbstlernender) Systeme zu stärken („Explainable AI"), und empfiehlt der Bundesregierung, die weitere Forschung und Entwicklung in diesem Bereich zu fördern.

48

In bestimmten Sektoren, in denen nicht nur individuelle, sondern in besonderem Maße auch gesellschaftliche Interessen berührt sind, sollten auch **nicht unmittelbar betroffene Personen** ein Recht auf Zugang zu bestimmten Informationen über die algorithmischen Systeme erhalten. Entsprechende Rechte werden in erster Linie für journalistische und Forschungszwecke infrage kommen und sind zudem mit Blick auf die betroffenen Interessen der Betreiber durch hinreichende Schutzmaßnahmen zu flankieren. Unter Umständen, insbesondere beim staatlichen Einsatz von algorithmischen Systemen mit einem erheblichen Schädigungspotenzial (Stufe 4), kommen nach Ansicht der DEK darüber hinaus auch voraussetzungslose Informationszugangsansprüche in Frage.

49

Bei algorithmischen Systemen ab einem gewissen Schädigungspotenzial (ab Stufe 2) ist es sachgerecht und zumutbar, dem Betreiber gesetzlich die Erstellung und Veröffentlichung einer angemessenen **Risikofolgenabschätzung** abzuverlangen, die auch bei der Verarbeitung nicht-personenbezogener Daten greift und Risiken außerhalb des Datenschutzes berücksichtigt. Sie sollte insbesondere auch eine Abschätzung der Risiken für Selbstbestimmung, Privatheit, körperliche Unversehrtheit, persönliche Integrität sowie Vermögen, Eigentum und Diskriminierung umfassen. Außerdem sollte sie neben den zugrundeliegenden Daten und der Logik des Modells auch Qualitätsmaße und Fairnessmaße zu den Daten und zur Modellgüte berücksichtigen, etwa zu Bias oder (statistischen) Fehlerquoten (insgesamt oder für bestimmte Teilgruppen), die ein System bei der Vorhersage/Kategorienbildung aufweist.

50

Die Anforderungen an **Dokumentation und Protokollierung** in Bezug auf die verwendeten Datensätze und Modelle, die Granularität, die Aufbewahrungszeiten und die Verwendungszwecke sollten konkretisiert werden, damit die Verantwortlichen und Auftragsverarbeiter Rechtsklarheit erhalten. Zum anderen sollte für sensible Anwendungen künftig eine Pflicht etabliert werden, die Programmabläufe einer Software, die nachhaltige Schäden verursachen können, zu dokumentieren und zu protokollieren. Die verwendeten Datensätze und Modelle sind so zu beschreiben, dass diese für Aufsichtsinstitutionen im Falle einer Kontrolle nachvollziehbar sind (etwa hinsichtlich der Herkunft und Aufbereitung von Datensätzen oder der Optimierungsziele der Modelle).

51

Der Normgeber sollte Betreibern ein Mindestmaß an **technischen und mathematisch-prozeduralen Qualitätsgarantien** abverlangen, welche die Korrektheit und Rechtmäßigkeit der algorithmisch ermittelten Ergebnisse durch Verfahrensvorgaben absichern. Dazu können insbesondere Vorgaben für Korrektur- und Kontrollmechanismen oder für die Datenqualität sowie die Sicherheit des Systems gehören. So wäre es beispielsweise sachgerecht, qualitative Anforderungen an das Verhältnis zwischen der Datengrundlage und dem Ergebnis des algorithmischen Datenverarbeitungsprozesses vorzugeben.

52

Beim Einsatz algorithmischer Systeme im Kontext menschlicher Entscheidungen sieht die DEK zunächst Klarstellungs- und Konkretisierungsbedarf betreffend die Anwendungsvoraussetzungen und Rechtsfolgen von Art. 22 DSGVO. Darüber hinaus empfiehlt die DEK, **Schutzmechanismen auch für algorithmenbasierte und -getriebene Entscheidungssysteme** vorzusehen, da sich der Einfluss dieser Systeme in der Praxis nahezu ebenso stark auswirken kann wie bei algorithmendeterminierten Anwendungen. Diesbezüglich empfiehlt sich anstelle des von Art. 22 DSGVO bislang verfolgten Verbotsprinzips ein flexibleres, risikoadaptiertes Regulierungsregime, das dem Einzelnen angemessene Schutzgarantien (insbesondere im Falle von Profiling) und Verteidigungsmöglichkeiten gegen Fehler und Bedrohungen seiner Rechte vermittelt.

53

Es ist erwägenswert, den **Anwendungsbereich des Antidiskriminierungsrechts** in situativer Hinsicht auf Diskriminierungen auszudehnen, die auf einer automatisierten Datenauswertung oder einem automatisierten Entscheidungsverfahren beruhen. Der Gesetzgeber sollte darüber hinaus Maßnahmen eines wirksamen Schutzes gegen **Diskriminierungen aufgrund von Gruppenmerkmalen** etablieren, die an sich nicht zu den gesetzlich geschützten Diskriminierungsmerkmalen zählen, und bei denen Diskriminierungen derzeit vielfach auch nicht als mittelbare Diskriminierung aufgrund eines geschützten Merkmals qualifiziert werden können.

54

Zusätzlich zu bereits bestehender Regulierung ist es für algorithmische Systeme mit deutlichem oder regelmäßigem (Stufe 3) oder sogar erheblichem Schädigungspotenzial (Stufe 4) sinnvoll, **Zulassungsverfahren oder Vorabprüfungen** von algorithmischen Systemen durch Aufsichtsinstitutionen zu etablieren, um Schäden für einzelne Betroffene, Bevölkerungsgruppen oder die Gesellschaft als Ganzes abzuwenden.

Institutionen

55

Die DEK empfiehlt der Bundesregierung, die bestehenden Aufsichtsinstitutionen und -strukturen im Rahmen ihrer Zuständigkeit zu stärken, neu auszurichten und, wo erforderlich, auch neue Institutionen und Strukturen zu schaffen. Dabei sollten die behördlichen Aufsichtsaufgaben und Kontrollbefugnisse primär jeweils denjenigen **sektoralen Aufsichtsbehörden** zugewiesen werden, die bereits sektorspezifische Sachkompetenzen ausgebildet haben. Von großer Bedeutung ist es dabei, dass die zuständigen Behörden mit den erforderlichen finanziellen, personellen und technischen **Ressourcen** ausgestattet werden.

56

Darüber hinaus empfiehlt die DEK der Bundesregierung die Schaffung eines **bundesweiten Kompetenzzentrums Algorithmische Systeme**, welches die sektoralen Aufsichtsbehörden durch technischen und regulatorischen Sachverstand in ihrer Aufgabe unterstützt, algorithmische Systeme im Hinblick auf die Einhaltung von Recht und Gesetz zu kontrollieren.

57

Aus Sicht der DEK sollten Initiativen unterstützt werden, die – ggf. differenziert nach kritischen Anwendungsbereichen – technisch-statistische **Standards für die Qualität von Testverfahren und Audits** festlegen. Für die Überprüfbarkeit algorithmischer Systeme können derartige Testverfahren künftig eine zentrale Rolle spielen, wenn sie hinreichend aussagekräftig, verlässlich und sicher ausgestaltet sind.

58

Innovative Formen der **Ko- und Selbstregulierung** verdienen aus Sicht der DEK neben und in Ergänzung zu staatlichen Formen der Regulierung besondere Aufmerksamkeit. Die DEK empfiehlt der Bundesregierung die Prüfung verschiedener Modelle der Ko- und Selbstregulierung, die für bestimmte Konstellationen adäquate Antworten liefern können.

59

Die DEK hält es für erwägenswert, den Betreibern – nach dem Regulierungsmodell „Comply or Explain" – die gesetzliche Pflicht aufzuerlegen, sich zu den Regeln eines **Algorithmic Accountability Codex** zu bekennen. Die Erarbeitung eines solchen bindenden Codex für die Betreiber von algorithmischen Systemen könnte dabei durch eine unabhängige, paritätisch besetzte Kommission erfolgen, die nicht unter staatlichem Einfluss stehen dürfte. Vertreter der Zivilgesellschaft sollten bei der Erarbeitung eines solches Codex in angemessener Weise beteiligt werden.

60

Auch ein spezifisches **Gütesiegel** als freiwilliges oder verpflichtendes Schutzzeichen kann Verbrauchern Orientierung über vertrauenswürdige algorithmische Systeme geben und gleichzeitig marktwirtschaftliche Anreize für Entwickler und Betreiber setzen, vertrauenswürdige Systeme zu entwickeln und zu verwenden.

61

Ähnlich wie schon heute Unternehmen ab einer bestimmten Größe einen Datenschutzbeauftragten benennen müssen, sollten nach Auffassung der DEK künftig auch solche Unternehmen und Behörden, die kritische algorithmische Systeme betreiben, einen **Ansprechpartner** benennen müssen. Er soll für die Kommunikation mit Behörden zur Verfügung stehen und zu einer Mitwirkung verpflichtet sein.

62

Um sicherzustellen, dass bei der behördlichen Überprüfung algorithmischer Systeme auch die Interessen der Zivilgesellschaft und betroffener Unternehmen angemessen berücksichtigt werden, sollten geeignete **Beiräte bei den sektoralen Aufsichtsbehörden** gebildet werden.

63

Die DEK stuft technische Standards **akkreditierter Normungsorganisationen** als ein grundsätzlich sinnvolles Instrument zwischen staatlicher Regulierung und rein privater Selbstregulierung an. Sie empfiehlt daher der Bundesregierung, in geeigneter Weise auf die Entwicklung und Verabschiedung technischer Standards hinzuwirken.

64

Die in Deutschland bewährten **Klagerechte von Wettbewerbern** und von **Wettbewerbs- und Verbraucherverbänden** sind ein zentraler Baustein für eine zivilgesellschaftliche Kontrolle des Einsatzes von algorithmischen Systemen. Besonders legitimierte zivilgesellschaftliche Akteure können durch solche privaten Klagerechte die Einhaltung von Rechtsvorschriften im Bereich des Vertragsrechts, des Lauterkeitsrechts oder des Antidiskriminierungsrechts sicherstellen, ohne hierbei auf das Tätigwerden von Behörden oder die Mandatierung durch einzelne Betroffene angewiesen zu sein.

Besonderes Augenmerk: Algorithmische Systeme bei Medienintermediären

65

Vor dem Hintergrund der besonderen Gefahren von Medienintermediären mit **Torwächterfunktion für die Demokratie** empfiehlt die DEK, auch mit Blick auf eine Einwirkung auf den EU-Gesetzgeber (→ siehe oben Empfehlung Nr. 43) zu prüfen, wie den mit einer solchen Torwächterfunktion verbundenen Gefahren begegnet werden kann. Dabei sollte ein ganzes Spektrum gefahrenabwehrender Maßnahmen erwogen werden, das bis hin zu einer Ex-ante-Kontrolle (z. B. in Form eines Lizenzierungsverfahrens) reichen kann.

66

Den nationalen Gesetzgeber trifft die verfassungsrechtliche Pflicht, die Demokratie vor den Gefahren für die freie demokratische und plurale Meinungsbildung, die von Anbietern mit Torwächterfunktion ausgehen, durch **Etablierung einer positiven Medienordnung** zu schützen. Die DEK empfiehlt, die Anbieter in diesem engen Bereich zum Einsatz solcher algorithmischer Systeme zu verpflichten, die den Nutzern zumindest als zusätzliches Angebot auch einen Zugriff auf eine tendenzfreie, ausgewogene und die plurale Meinungsvielfalt abbildende Zusammenstellung von Beiträgen und Informationen verschaffen.

67

Für alle Medienintermediäre und auch bei Anbietern ohne Torwächterfunktion oder bei geringerem Schädigungspotenzial für die demokratische Meinungsbildung sollte die Bundesregierung Maßnahmen prüfen, die den charakteristischen Gefahren des Mediensektors Rechnung tragen. Dies könnte Mechanismen zur **Transparenzsteigerung** (z. B. Einblick in technische Verfahren der Nachrichtenauswahl und -priorisierung, **Kennzeichnungspflichten für Social Bots**) und ein Recht auf Gegendarstellung in Timelines umfassen.

Der Einsatz von algorithmischen Systemen durch staatliche Stellen

68

Der Staat ist im Interesse seiner Bürger zur Nutzung der besten verfügbaren Technik – einschließlich algorithmischer Systeme – verpflichtet, muss dabei jedoch im Lichte seiner Grundrechtsbindung sowie der Vorbildfunktion allen staatlichen Handelns besondere Sorgfalt walten lassen. Der Einsatz algorithmischer Systeme durch Hoheitsträger ist daher **im Allgemeinen als besonders sensibel im Sinne des Kritikalitätsmodells** einzustufen und erfordert mindestens eine umfassende Risikofolgenabschätzung.

69

Aufgaben in der **Rechtsetzung** und der **Rechtsprechung** dürfen algorithmischen Systemen allenfalls in Randbereichen übertragen werden. Insbesondere dürfen algorithmische Systeme nicht genutzt werden, um die freie Willensbildung im demokratischen Prozess und die sachliche Unabhängigkeit der Gerichte zu unterminieren. Große Potenziale für den Einsatz algorithmischer Systeme bestehen hingegen in der **Verwaltung**, vor allem in der Leistungsverwaltung. Um dem Rechnung zu tragen, sollte der Gesetzgeber verstärkt teil- und vollautomatisierte Verwaltungsverfahren zulassen. Dazu bedarf es auch einer vorsichtigen Fortentwicklung des zu engen § 35a VwVfG sowie der entsprechenden einfachrechtlichen Normen. Bei alledem gilt es, hinreichende Schutzmaßnahmen für die Bürger vorzusehen.

70

Staatliche Entscheidungen, die unter Nutzung algorithmischer Systeme zustande kommen, müssen **transparent und begründbar** bleiben. Dazu bedarf es ggf. Klarstellungen bzw. Erweiterungen der bestehenden Informationsfreiheits- und Transparenzgesetze. Ferner entbindet der Einsatz algorithmischer Systeme nicht vom Grundsatz, dass hoheitliche Entscheidungen regelmäßig im Einzelfall begründet werden müssen; im Gegenteil kann dieser Grundsatz dem Einsatz allzu komplexer algorithmischer Systeme Grenzen setzen. Schließlich trägt die Nutzung von Open-Source-Lösungen wesentlich zur Transparenz staatlichen Handelns bei und sollte daher verstärkt angestrebt werden.

71

Zwar ist aus ethischer Sicht ein generelles Recht auf Freiheit zur Nichtbefolgung von Normen nicht anzuerkennen. Gleichzeitig wirft ein automatisierter Totalvollzug des Rechts eine Reihe ethischer Bedenken auf. Daher ist regelmäßig ein technisches Design zu fordern, bei dem der Mensch im Einzelfall den **technischen Vollzug** außer Kraft setzen kann. Ferner muss stets die Verhältnismäßigkeit zwischen der potenziellen Normübertretung und der automatisierten (ggf. präventiven) Vollzugsmaßnahme gewahrt sein.

Haftung für algorithmische Systeme

72

Neben strafrechtlicher Verantwortlichkeit und Verwaltungssanktionen ist auch die Haftung auf Schadensersatz unverzichtbarer Bestandteil eines ethisch vertretbaren Ordnungsrahmens. Es ist bereits jetzt erkennbar, dass algorithmische Systeme – u. a. aufgrund der Komplexität und Dynamik der Systeme sowie aufgrund ihrer wachsenden „Autonomie" – das bestehende Haftungsrecht vor Herausforderungen stellen. Die DEK empfiehlt daher eine umfassende Prüfung und, soweit erforderlich, **Anpassung des geltenden Haftungsrechts**. Der Blick sollte sich dabei nicht allein auf bestimmte technologische Merkmale – wie etwa auf das Merkmal Maschinellen Lernens oder Künstlicher Intelligenz – verengen.

73

Der Gedanke, algorithmischen Systemen hoher Autonomie künftig Rechtspersönlichkeit zuzuerkennen und sie selbst für Schäden haften zu lassen **(„elektronische Person")**, sollte **nicht weiterverfolgt** werden. Soweit dieser Gedanke auf eine Analogie zwischen Mensch und Maschine gestützt wird, ist er schon ethisch nicht vertretbar, und soweit es schlicht um die Anerkennung einer neuen Gesellschaftsform im Sinne des Gesellschaftsrechts geht, löst er keine Probleme.

74

Dagegen ist es geboten, für den Einsatz sog. autonomer Systeme – abhängig von der Natur der dem System übertragenen Aufgaben – auch eine Zurechnung schädigender Vorgänge entsprechend den Regelungen über die Haftung für **Gehilfen** (vgl. insbes. § 278 BGB) vorzunehmen. Beispielsweise sollte eine Bank, die sich für die Prüfung der Kreditwürdigkeit eines autonomen Systems bedient, gegenüber ihrem Kunden mindestens in gleichem Maße haften, wie wenn sie sich eines menschlichen Mitarbeiters bedient hätte.

75

Daneben erscheint es nach derzeitigem Stand der Diskussion sehr wahrscheinlich, dass zusätzlich zu einer sachgerechten Anpassung der aus den 1980er Jahren stammenden **Produkthaftungsrichtlinie** und Verknüpfung mit neuen Standards der Produktsicherheit auch punktuelle Modifikationen der **Verschuldenshaftung** und/oder neue Tatbestände der **Gefährdungshaftung** erforderlich sein werden. Dabei wird jeweils zu klären sein, für welche Produkte, digitalen Inhalte und digitalen Dienstleistungen welches Haftungsregime sachgerecht und wie dieses konkret auszugestalten ist, wobei es wiederum wesentlich u. a. auf die Kritikalität des betreffenden algorithmischen Systems ankommen wird. Dabei sollten auch innovative Haftungskonzepte, wie sie derzeit auf europäischer Ebene entwickelt werden, in Betracht gezogen werden.

4

Für einen europäischen Weg

Die Fülle an Fragen, die sich der DEK im Rahmen ihrer Arbeit gestellt haben und deren Diskussion jeweils wieder neue Fragen aufgeworfen hat, lässt deutlich werden, dass dieses Gutachten lediglich einen weiteren Grundstein für einen **andauernden Zukunftsdiskurs über Ethik, Recht und Technologie** legen kann. Die DEK betont dabei, dass Ethik, Recht und Demokratie auch in der technischen Welt ihre gestaltende Kraft entfalten müssen. Dazu bedarf es eines interdisziplinären Diskurses in Politik und Gesellschaft sowie einer Gesetzgebung und Regulierung, die so offen gestaltet ist, dass sie auch bei schneller Entwicklung von Technik und Geschäftsmodellen ihre Regelungskraft und Reaktionsfähigkeit behält. Es bedarf zusätzlich der Instrumente, Verfahren und Strukturen, um die Regulierung effektiv durchzusetzen und bei Verstößen oder Fehlentwicklungen rechtzeitig einschreiten zu können.

Deutschland und Europa sehen sich allerdings im globalen Wettlauf um Zukunftstechnologien mit Wertesystemen, Gesellschaftsmodellen und Kulturen konfrontiert, die sich von unseren unterscheiden. Die DEK unterstützt den bislang eingeschlagenen **„europäischen Weg“**: Europäische Technologien sollten sich durch konsequente Ausrichtung an europäischen Werten und Grundrechten, wie sie insbesondere auch in der Charta der Grundrechte der Europäischen Union und in der Konvention zum Schutz der Menschenrechte und Grundfreiheiten des Europarats zum Ausdruck kommen, auszeichnen.

Die DEK sieht den Staat in besonderer Verantwortung, im Einklang mit unserer Werteordnung ethische Maßstäbe auch für den digitalen Raum zu formulieren und diese durchzusetzen. Um diese Garantie gegenüber den Bürgern auch einhalten zu können, bedarf es international einer Position politischer und ökonomischer Stärke: Wer von anderen übermäßig abhängig ist, wird vom „rule maker“ zum „rule taker“ und setzt seine Bürger letztlich Vorgaben aus, die von Akteuren aus anderen Regionen der Welt formuliert werden, oder von privaten Akteuren, die demokratischer Legitimation und Kontrolle weitgehend entzogen sind. Bemühungen um die **langfristige Sicherung der digitalen Souveränität Deutschlands und Europas** sind daher nicht nur ein Gebot politischer Weitsicht, sondern auch Ausdruck ethischer Verantwortung.

Teil A

Einleitung

1. Arbeitsauftrag und Grundverständnis

Die Digitalisierung verändert unsere Gesellschaft tiefgreifend. Neuartige datenbasierte Technologien können für das Leben des Einzelnen und das gesellschaftliche Zusammenleben Nutzen stiften, die Produktivität der Wirtschaft steigern, zu mehr Nachhaltigkeit und zu grundlegenden Fortschritten in der Wissenschaft beitragen, und tun dies zum Teil schon heute. Die digitale Transformation bietet gerade für Deutschland als eng vernetzte und hoch technologisierte Volkswirtschaft enorme Chancen, übt jedoch auch zunehmenden internationalen Konkurrenzdruck auf deutsche Unternehmen aus. Gleichzeitig zeigen sich bereits jetzt die Risiken der Digitalisierung für grundlegende Rechte und Freiheiten. Es stellen sich damit zahlreiche ethische und rechtliche Fragen, in deren Mittelpunkt die gewünschte Rolle und die Gestaltung der neuen Technologien steht. Wenn der digitale Wandel zum Wohl für den Einzelnen und die gesamte Gesellschaft führen soll, müssen sich Gesellschaft und Politik mit der Gestaltung datenbasierter Technologien einschließlich der KI befassen.

Die Bundesregierung hat am 18. Juli 2018 die Datenethikkommission (DEK) eingesetzt und sechzehn Mitglieder (siehe Anhang, 2.) berufen. Christiane Wendehorst und Christiane Woopen wurden gemeinsam zu Vorsitzenden bestellt. Die DEK erhielt den Auftrag, innerhalb eines Jahres ethische Maßstäbe und Leitlinien für den Schutz des Einzelnen, die Wahrung des gesellschaftlichen Zusammenlebens und die Sicherung und Förderung des Wohlstands im Informationszeitalter zu entwickeln. Die Kommission soll auch konkrete Handlungs- und Regulierungsempfehlungen unterbreiten, wie diese ethischen Leitlinien beachtet, implementiert und beaufsichtigt werden können. Dazu hat die Bundesregierung der DEK Leitfragen (siehe Anhang, 1.) an die Hand gegeben, die sich auf die drei Themenfelder: (I.) Algorithmenbasierte Prognose- und Entscheidungsprozesse, (II.) Künstliche Intelligenz und (III.) Daten konzentrieren.

Die DEK versteht **KI** in diesem Zusammenhang als Sammelbegriff für diejenigen Technologien und ihre Anwendungen, die durch digitale Methoden auf der Grundlage potenziell sehr großer und heterogener Datensätze in einem komplexen und die menschliche Intelligenz gleichsam nachahmenden maschinellen Verarbeitungsprozess ein Ergebnis ermitteln, das ggf. automatisiert zur Anwendung gebracht wird. Die wichtigsten Grundlagen für KI als Teilgebiet der Informatik sind die subsymbolische Mustererkennung, das maschinelle Lernen, die computergerechte Wissensrepräsentation und die Wissensverarbeitung, welche Methoden der heuristischen Suche, der Inferenz und der Handlungsplanung umfasst.

Die DEK hält es indessen für unangemessen, ethische und rechtliche Betrachtungen ausschließlich auf KI zu konzentrieren. Sie ist lediglich eine besondere Ausprägung und damit ein Teilbereich algorithmischer Systeme. Einige Eigenschaften, die zu ethischen Problemen führen können, teilt sie mit anderen Arten algorithmischer Systeme, so dass eine auf KI beschränkte Regulierung nur einen Teil der Probleme erfassen würde. Die bei KI im Vordergrund stehende Eigenschaft des Selbstlernens bringt zwar spezifische Herausforderungen mit sich, die bei einer Risikobestimmung besonders zu berücksichtigen sind, sie ist aber nicht die einzige Eigenschaft, die besonderer Aufmerksamkeit bedarf. Insofern beziehen sich die folgenden Ausführungen auf **alle Arten algorithmischer Systeme**.

Anwendungen beruhen selten auf einem einzigen Algorithmus, und eine isolierte Betrachtung von Algorithmen hat selten Aussagekraft. Für die ethische Beurteilung kommt es jeweils auf das **gesamte sozio-technische System** an, also alle Komponenten einer algorithmischen Anwendung einschließlich aller menschlichen Akteure, von der Entwicklungsphase (z.B. hinsichtlich der verwendeten Trainingsdaten) bis hin zur Implementierung in einer Anwendungsumgebung und zur Phase von Bewertung und Korrektur.

2. Arbeitsweise

Die DEK trat zwischen September 2018 und September 2019 monatlich zusammen. Die DEK diskutierte exemplarische Anwendungsbeispiele neuer Technologien („Use Cases“) in verschiedenen Sektoren, die im Hinblick auf ihre technischen Grundlagen sowie unter ethischen und juristischen Gesichtspunkten analysiert wurden. Die daraus und aus grundlegenden Diskussionen gewonnenen Erkenntnisse erlaubten die Identifizierung von übergeordneten Themen und Fragestellungen, um Eckpunkte zur ethischen Einordnung sowie konkrete Empfehlungen für künftiges politisches und gesetzgeberisches Handeln zu entwickeln. Bereits im Oktober 2018 unterbreitete die DEK auf der Grundlage eines Eckpunktepapiers der Bundesregierung zwei konkrete Empfehlungen für die Ausgestaltung der „Strategie Künstliche Intelligenz“, die von der Bundesregierung aufgegriffen wurden. Im November 2018 gab die DEK noch eine weitere Empfehlung ab, in der sie sich für eine partizipative Entwicklung der elektronischen Patientenakte aussprach.[1]

Die Öffentlichkeit wurde im Rahmen von zwei öffentlichen Tagungen mit einbezogen. Die erste Tagung fand am 7. Februar 2019 im Bundesministerium der Justiz und für Verbraucherschutz (BMJV) zu dem Thema „Selbst- und Fremdbestimmung im Zeitalter künstlicher Intelligenz“ statt. Die zweite Tagung wurde am 9. Mai 2019 als International Round Table mit dem Titel „Für eine ethische Gestaltung unserer digitalen Zukunft“ („Towards Shaping of Our Digital Future“) im Bundesministerium des Innern, für Bau und Heimat (BMI) ausgerichtet. Beide Veranstaltungen erlaubten einen intensiven Austausch der DEK mit Experten, Stakeholdern sowie der Öffentlichkeit und interessierten Bürgerinnen und Bürgern.[2]

Am 14. November 2018 fand anlässlich der Digitalklausur der Bundesregierung ein Austausch zwischen der Bundeskanzlerin sowie allen Mitgliedern der Bundesregierung und den beiden Vorsitzenden der DEK statt. Darüber hinaus wurden anlassbezogen Gespräche mit einzelnen Mitgliedern der Bundesregierung geführt. Zudem wurden Experten angehört und Konsultationstreffen mit anderen Institutionen und Gremien durchgeführt, die sich verwandten Themen widmen – darunter etwa die Enquete-Kommission „Künstliche Intelligenz“, die Kommission Wettbewerbsrecht 4.0, der Digitalrat der Bundesregierung, der Sachverständigenrat für Verbraucherfragen, u.v.m.

Es gehörte zu den wesentlichen Merkmalen der DEK, dass sie in völliger Unabhängigkeit und frei von jeglicher externen politischen Einflussnahme beraten und arbeiten konnte. Die in diesem Gutachten niedergelegten Standpunkte geben ausschließlich die persönliche Überzeugung der *ad personam* berufenen Mitglieder sowie die interne Meinungsbildung der institutionellen Mitglieder wieder. Die DEK hat alle Empfehlungen dieses Gutachtens im Konsens verabschiedet.

1 Beide Dokumente stehen auf der Internetseite der DEK (abrufbar unter: *www.datenethikkommission.de*).
2 Weitergehende Informationen zu den öffentlichen Tagungen, inklusive der Videoaufnahmen, auf der Internetseite der DEK (abrufbar unter: *www.datenethikkommission.de*).

3. Ziele und Gegenstand des Gutachtens

Die DEK möchte mit diesem Gutachten einen Beitrag dazu leisten, unseren ethischen und rechtlichen **Ordnungsrahmen** angesichts der Herausforderungen durch digitale Technologien weiter zu entwickeln. Im Vordergrund stehen dabei die Gewährleistung der essenziellen Bedingungen für die freiheitlich-demokratische Grundordnung sowie die Nutzung der Potenziale für die Verwirklichung nachhaltigkeitsorientierter Ziele und das Gedeihen unserer sozialen Marktwirtschaft.

Angesichts der zunehmenden Erfassung personenbezogener Daten und ihrer automatisierten Verarbeitung zu unterschiedlichen Zwecken ist es ein wichtiges Anliegen der DEK, die **grundlegenden Rechte und Freiheiten des Individuums** einschließlich des Schutzes seiner Selbstbestimmung und Integrität mit dem Fortschritt, dem Wohlstand, der Sicherung der Demokratie und der Gestaltung einer zukunftsfähigen Gesellschaft zusammen zu denken. Es ist die Aufgabe des Rechtsstaats, den Einzelnen vor Datenmissbrauch und vor Diskriminierung zu schützen und für die Sicherheit aller Akteure zu sorgen. Dafür muss er wirksame Regularien und Institutionen schaffen. Gleichzeitig sollte er innovative Geschäftsmodelle ermöglichen, die den zukünftigen Wohlstand für alle sichern.

Die DEK sieht in der Digitalisierung – insbesondere in Form der zunehmenden Verfügbarkeit von Daten und des Einsatzes komplexer algorithmischer Systeme einschließlich Künstlicher Intelligenz – **große Potenziale** für technische und soziale Innovationen sowie für die Verwirklichung der Nachhaltigkeitsziele der Vereinten Nationen. Das betrifft unter anderem die Förderung der Gesundheit, die Humanisierung der Arbeitswelt, die Gestaltung nachhaltiger Städte und Gemeinden, angemessene Bildung sowie Maßnahmen für einen wirksamen Klimaschutz. Gleichzeitig sind **hohe Risiken** zu bedenken, die sich, getrieben durch den umfassenden Einsatz digitaler Technologien, für den Einzelnen, für die Gesellschaft und für die freiheitlich-demokratische Grundordnung ergeben können. Dazu gehört beispielsweise die Möglichkeit der Erstellung feingranularer Persönlichkeitsprofile (von Online Tracking über die Analyse der Stimme im Rahmen fernkommunikativer Bewerbungsgespräche bis hin zur Diagnostik von pathologischen psychischen Zuständen anhand der Beiträge in sozialen Medien), die Möglichkeit der Ausnutzung dieser zur Steuerung und Manipulation (von individueller Preissetzung bis hin zur Manipulation demokratischer Meinungsbildungsprozesse im Rahmen des sog. Microtargeting), die Diskriminierung gesellschaftlicher Gruppen sowie die Delegation menschlicher Verantwortung an Maschinen. Die DEK ruft in diesem Zusammenhang zu einer aktiven Mitgestaltung unserer Zukunft auf, die Potenziale verwirklicht und Risiken vermeidet.

Der Weg zur Verwirklichung dieser Ziele durchläuft aus Sicht der DEK mehrere Ebenen. Er beginnt mit der ethischen Reflektion über den Wert menschlichen Handelns in einem technologiegeprägten Umfeld und der Bekräftigung **zentraler ethischer Grundsätze und Prinzipien**, auf denen unsere Gesellschaft aufgebaut ist (→ Teil B). Die Leitfragen enthalten aus Sicht der DEK eine datenfokussierte Perspektive („Daten-Perspektive“) und eine auf algorithmische Systeme fokussierte Perspektive („Algorithmen-Perspektive“) als zwei sich wechselseitig ergänzende und bedingende ethische Diskurse, die sich auch in jeweils unterschiedlichen **Governance-Instrumenten** widerspiegeln (→ Teil D).

Unter der Daten-Perspektive (→ Teil E) entwickelt die DEK allgemeine ethische Prinzipien für den **Umgang mit Daten** (→ E 1) und vor allem ethische Grundsätze von **Datenrechten und Datenpflichten** (→ E 2), um diese in eine Reihe konkreter Handlungsempfehlungen betreffend die Nutzung von Daten und den Datenzugang münden zu lassen (→ E 3 bis 5). Unter der Algorithmen-Perspektive (→ Teil F) formuliert die DEK allgemeine ethische Anforderungen an das **Design algorithmischer Systeme** (→ F 2) und an deren **risikoadaptierte Regulierung** (→ F 3). Sie leuchtet sodann im Detail die Instrumente und Institutionen einer solchen Regulierung aus, wie sie dem Gesetzgeber als Empfehlung unterbreitet werden (→ F 4 bis 8). Voraussetzung für diese Überlegungen ist ein gemeinsames Grundverständnis technischer Gegebenheiten und Zusammenhänge (→ Teil C). Das Gutachten endet mit einem Plädoyer für einen „europäischen Weg“ (→ Teil G).

Die Empfehlungen der DEK richten sich dem Auftrag gemäß primär an die deutsche **Bundesregierung** und die mit ihr verbundenen Institutionen. An einigen Stellen sind indessen auch andere Akteure angesprochen, etwa Länder und Gemeinden, Forschungseinrichtungen oder Unternehmen. Solche Empfehlungen sind insofern immer auch an die Bundesregierung gerichtet, als der Bundesregierung empfohlen wird, die anderen Akteure in ihren Bemühungen zu ermutigen und zu unterstützen. Alle Empfehlungen sind ferner im Kontext der europäischen und internationalen Entwicklungen und dort bereits bestehenden oder zu entwickelnden Institutionen und Regulierungen zu verstehen. Soweit eine Empfehlung der DEK auf **europäischer oder internationaler Ebene** umgesetzt werden sollte, ist sie als Empfehlung an die deutsche Bundesregierung zu verstehen, sich kraftvoll und zukunftsorientiert in Europa und international einzubringen.

Teil B

Ethische und rechtliche Grundsätze und Prinzipien

1. Der grundsätzliche Wert menschlichen Handelns

Im Zuge der rasant fortschreitenden Entwicklung digitaler Technologien einschließlich selbstlernender algorithmischer Systeme („Künstliche Intelligenz"), die bestimmte Funktionen menschlichen Handelns in ihrer Leistungskraft übersteigen, stellt sich die **grundlegende Frage, ob das Handeln eines Menschen einen ethisch relevanten Wert an sich darstellt,** der sich jenseits von Effektivität und Effizienz verwirklicht, und der dem Funktionieren maschineller Systeme vorzuziehen ist. Die Frage stellt sich umso dringender, als der internationale Wettbewerb eine Dynamik und Eigengesetzlichkeit entfaltet, die strikt auf im Wesentlichen ökonomische Effizienz ausgerichtet ist.

Das menschliche Handeln bezieht seinen grundsätzlichen Wert aus seiner moralischen Bedeutung. Der Mensch kann Gründe für sein Handeln angeben, sich für oder gegen ein bestimmtes Handeln entscheiden, und er muss sein Handeln verantworten. Im Handeln verwirklicht und entfaltet sich der Mensch gemäß seinen Möglichkeiten, seinen Präferenzen und seinen Vorstellungen von einem sinnvollen Leben. Diese **Sinndimension des Handelns** macht es zu einem Wert, den das Funktionieren technischer Systeme niemals erhalten kann. Technik ist stets nur Mittel zu einem Zweck, den Menschen gesetzt haben. Auch wenn Menschen – hypothetisch gesprochen – entscheiden sollten, dass sich algorithmische Systeme selbst Zwecke setzen können, ist die Ermöglichung technischer Zwecksetzung ein menschlich gesetzter Zweck. Insofern kann der Einsatz technischer Systeme zwar ein Element menschlichen Handelns sein – das in bestimmten Fällen sogar ethisch geboten sein mag – technische Systeme können aber menschliches Handeln in seiner moralischen Dimension niemals vollständig ersetzen. Menschen handeln und entfalten sich als Lebewesen in mehreren Dimensionen. Auch wenn Menschenbilder in unterschiedlichen Kulturen und aufgrund unterschiedlicher religiöser Überzeugungen erhebliche Unterschiede aufweisen, enthalten sie doch alle die Dimension des Lebendigen und der moralischen Verantwortung, und sie umfassen bei aller Unterschiedlichkeit der jeweiligen Antworten die Frage nach dem Sinn des Lebens – wohingegen technische Systeme lediglich funktionieren.

Ist nun zu entscheiden, wo menschlichem Handeln der Vorzug zu geben ist vor dem Einsatz algorithmischer Systeme, so spielen viele Kriterien eine Rolle. Grundsätzlich gebührt der höheren Effektivität der Vorrang, wenn es um die Erfüllung bestimmter begrenzter Funktionen geht. **Effektivität ist aber nicht der höchste Wert.** Sie darf die Entfaltung des Menschen in seinem eigenen Handeln nicht substanziell einschränken, und sie muss hinter der grundlegenden ethischen Dimension des sinnvollen und gelingenden Lebens als Einzelner und in der Gemeinschaft zurückstehen. Selbst wenn also beispielsweise ein Roboter einen Menschen effektiver pflegen könnte, dürfte die menschliche Zuwendung und Sorge für den pflegebedürftigen Menschen dadurch nicht ersetzt werden. Gleichwohl kann der Einsatz von Robotern in der Pflege zusätzlich zur menschlichen Zuwendung geboten sein, wenn dadurch die Sicherheit der zu pflegenden Person wesentlich erhöht wird. Wenn aber etwa ein Arbeitnehmer durch technische Systeme dazu gezwungen wird, seine gesamten Arbeitsabläufe in den Dienst maximaler Effektivität zu stellen und dabei seine Privatsphäre oder seine persönliche Integrität verletzt werden, hat die Effektivität zurückzustehen. Menschen dürfen nicht zu Objekten von Maschinen werden, sondern müssen ihre Subjektivität erhalten können.

Der Mensch ist moralisch verantwortlich für sein Handeln – er kann der moralischen Dimension nicht entkommen. Welche Ziele er verfolgt, welche Gründe er dafür hat und welche Mittel er einsetzt, liegt in seiner Verantwortung. Bei der Gestaltung unserer technologisch geprägten Zukunft ist dieser Dimension stets Rechnung zu tragen. Dabei gilt unverrückbar, dass Technik dem Menschen dient und nicht der Mensch der Technik unterworfen wird. Dieses **Verständnis vom Menschen** liegt unserer Verfassungsordnung zugrunde und steht in der Tradition der europäischen Kultur- und Geistesgeschichte.

2. Verhältnis von Ethik und Recht

Das Leben jedes einzelnen Menschen und alle Bereiche des gesellschaftlichen Zusammenlebens werden durch die exponentielle technische Entwicklung bei der Erhebung und Verwendung digitaler Daten sowie dem Einsatz algorithmischer Systeme und Künstlicher Intelligenz zunehmend geprägt. Dadurch entstehen weit reichende und tief greifende Fragen, deren Beantwortung sich an den **rechtlichen und ethischen Grundsätzen**, auf die sich die Gesellschaft in einer Demokratie verpflichtet hat, orientieren muss.

Die Maßstäbe und leitenden Prinzipien, anhand derer die Gesellschaft ihre unterschiedlichen Bereiche wie etwa die Wirtschaft, die Bildung, die Gestaltung des öffentlichen Raums, das Gesundheitswesen, den Finanzsektor, den Verkehr und die Energieversorgung gestaltet und zu gestalten hat, sind grundlegend ethischer Natur. Bei allem moralischen Pluralismus, der für ein freiheitliches System charakteristisch ist, gibt es dennoch einen gemeinsamen ethischen Ordnungsrahmen, der rechtlich in der Verfassung, und, betreffend das Verhältnis zwischen Staat und Individuum, insbesondere in den Grundrechten niedergelegt ist. Für die Frage, was dieser ethische und rechtliche Ordnungsrahmen bezogen auf einen Einzelfall und im Falle eines Konfliktes zwischen unterschiedlichen Werten oder Grundrechten bedeutet, gibt es nicht immer eindeutige Antworten. Das relativiert aber nicht die verbindliche Funktion und **konstitutive Bedeutung der ethischen Fundierung unseres Gemeinwesens**. Es betont vielmehr einmal mehr die unverzichtbare Bedeutung einer offenen und kontinuierlichen Debatte über die Gestaltung unserer Gesellschaft und ist die Grundlage demokratischer Entscheidungsprozesse, die ja gerade anerkennen, dass unterschiedliche Antworten im Rahmen der Verfassung denkbar sind.

Ethik geht nicht im Recht auf. Nicht jedes Detail, das ethisch relevant ist, kann und sollte rechtlich reguliert werden. Umgekehrt gibt es Aspekte rechtlicher Regulierung, die pragmatischer Art und ethisch nicht zwingend sind. Rechtssetzung muss aber immer mögliche ethische Implikationen reflektieren und ethischen Ansprüchen genügen – den verfassungsrechtlichen Vorgaben ohnehin.

Die Datenethikkommission ist der Ansicht, dass ethische Prinzipien und Leitlinien rechtliche Regulierung nicht entbehrlich machen können, wo es der verfassungsgerichtlich entwickelte **Wesentlichkeitsgrundsatz** erforderlich macht, demokratisch legitimierte und gegenüber jedermann durchsetzbare Regeln im Wege parlamentarischer Gesetzgebung zu erlassen. Internetpolitik ist auch Gesellschaftspolitik. Mit zunehmender Allgegenwärtigkeit algorithmischer Systeme einschließlich der Künstlichen Intelligenz werden auch Regeln für das gesellschaftliche Zusammenleben zu gestalten und zu sichern sein. Dies erfordert nicht nur eine fortwährende öffentliche, sondern insbesondere dort, wo Grundrechte betroffen sind, auch eine parlamentarische Debatte und gesetzgeberische Initiative. Auf durchsetzbare Regeln zugunsten von Freiwilligkeit systematisch zu verzichten erscheint angesichts der Erfahrungen mit der Rechtsdurchsetzung im Internet und der Beobachtung, dass Märkte, die durch digitale Technologien gekennzeichnet sind, in bestimmten Bereichen stärker zu einer Machtkonzentration neigen, nicht sinnvoll.

Regulierung soll gleichwohl technologische und soziale Innovationen sowie eine dynamische Marktentwicklung nicht blockieren. Allzu starre und detaillierte Gesetze können Handlungsspielräume einschränken und bürokratischen Aufwand auf eine Weise erhöhen, dass innovative Prozesse in Deutschland der Geschwindigkeit der internationalen technologischen Entwicklungen nicht mehr folgen können. Andererseits können und müssen regulative Rahmenbedingungen wesentliche Rechte und Freiheiten schützen und Rechtssicherheit schaffen. Dies ist die Grundlage dafür, dass Bürgerinnen und Bürger, Unternehmen und Institutionen auf eine ethisch ausgerichtete gesellschaftliche Transformation vertrauen können. Zudem bietet das Rechtssystem mit der Möglichkeit von Regulierung auf unterschiedlichen Ebenen – vom Gesetz über Verordnungen bishin zu Kodizes, Selbstverwaltung und Selbstverpflichtung – einen Instrumentenkasten, um anpassungsfähige und dem technologischen Fortschritt gerecht werdende Rahmenbedingungen zu gestalten.

Der **Bedarf an Orientierung geht jedoch über Regulierung weit hinaus**. Vor diesem Hintergrund haben in einer Zeit vielgestaltiger Umbrüche viele unterschiedliche Akteure wie etwa Berufsgruppen, Unternehmen und beratende Gremien auf nationaler, regionaler und internationaler Ebene Ethik-Kodizes oder Sets an leitenden ethischen Prinzipien formuliert und teilweise zur öffentlichen Diskussion gestellt.

Die Datenethikkommission begrüßt die Vielfalt des Engagements und der Diskussion über eine ethisch fundierte Gestaltung der Digitalisierung, die verdeutlicht, wie unverzichtbar die öffentliche Debatte und das **Einstehen aller für das Gelingen unseres Zusammenlebens** sind. In diesem Sinne orientiert sich die Datenethikkommission – wie im Koalitionsvertrag aufgetragen – bei ihren Empfehlungen für „einen Entwicklungsrahmen für Datenpolitik, den Umgang mit Algorithmen, künstlicher Intelligenz und digitalen Innovationen“ in Verbindung mit den verfassungsrechtlichen Grundsätzen an ethischen Querschnittsprinzipien, die in unterschiedlichen Gewichtungen in allen gesellschaftlichen Bereichen relevant sind und im Folgenden in aller Kürze skizziert werden.[1]

1 Die DEK bezieht sich mit diesem Ansatz auf dieselben Grundlagen, auf die sich auch die European Group on Ethics in Science and New Technologies (EGE) in ihrer Stellungnahme bezogen hat: EGE: Statement on Artificial Intelligence, Robotics and “Autonomous” Systems, 2018 (abrufbar unter: *http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf*).

3. Allgemeine ethische und rechtliche Grundsätze und Prinzipien

3.1 Die Würde des Menschen

Allem voraus und zugrunde liegt die Würde des Menschen, die in ethischer Hinsicht für den unbedingten Wert jedes menschlichen Lebewesens steht und als „tragendes Konstitutionsprinzip“ in der Verfassungsordnung verankert ist. Die Würde gebietet es anzuerkennen, dass jedem Menschen unabhängig von seinen Eigenschaften und Leistungen Respekt gebührt. Der **Schutz des dem Menschen inhärenten und nicht erst zu erwerbenden Wertes** beinhaltet, dass er nicht über alle seine Lebensbereiche und Tätigkeiten hinweg in ein klassifizierendes System eingeordnet („Super-Score“) oder wie ein Gegenstand mit einem Preis versehen und dementsprechend behandelt wird. Auch dort, wo menschliches Verhalten durch algorithmische Systeme gemessen und verarbeitet wird, ist stets zu berücksichtigen, dass jeder Mensch ein Individuum und kein Muster aus Datenpunkten ist. Algorithmische Systeme müssen daher stets so gestaltet sein, dass sie diesem Individualitätsanspruch des Einzelnen gerecht werden können.

Die Anerkennung der Menschenwürde erfordert, dass der Mensch immer „über der Technik“ steht, d.h. technischen Systemen nicht vollständig oder unwiderruflich unterworfen werden darf. Im konkreten Anwendungsfall kann sich dies auf unterschiedliche Ebenen der Gestaltungs- und Eingriffsmöglichkeiten beziehen, gleichwohl muss der **Grundsatz der menschlichen Gestaltungshoheit** gewahrt bleiben. Der Mensch ist in der Mensch-Maschine-Interaktion verantwortlicher Akteur und darf nicht als fehlerhaftes Wesen betrachtet werden, das von der Maschine optimiert oder perfektioniert werden muss. Vielmehr nutzt der Mensch algorithmische Systeme, um seine Ideen und Ziele besser, schneller und weniger fehlerbehaftet zu erreichen.

Der Würdeschutz umfasst darüber hinaus, dass der **Mensch als Beziehungswesen** technologisch nicht über die Art der Beziehung in die Irre geführt wird, wie es etwa der Fall sein könnte, wenn er mit einem Bot spricht und ihm systematisch vorgetäuscht wird, er spreche mit einem Menschen. Insbesondere schützt die Menschenwürde auch die **psychische Integrität der Einzelnen**. Untersagt ist daher die Nutzung von datengetriebenen Systemen zu manipulativen Zwecken, insbesondere wenn dies auf der Basis umfassender und feingranularer Persönlichkeitsprofile beruht. Gleiches gilt, wo algorithmische Systeme **Einzelne oder Gruppen systematisch diskriminieren,** also etwa herabstufen oder aus ethisch unvertretbaren Gründen von der Inanspruchnahme bestimmter Leistungen ausschließen oder bei der Beteiligung am demokratischen Diskurs sytematisch täuschen.

3.2 Selbstbestimmung

Mit der Würde des Menschen ist die Möglichkeit der Selbstbestimmung eng verbunden. Die Bestimmung seiner Lebensziele und seiner Lebensweise und damit die Bestimmung, Entfaltung und Darstellung seines Selbst sind **Ausdruck der Freiheit** des Menschen. Eine Gesellschaft, die Freiheit ernst meint, schafft Rahmenbedingungen, in denen sich die Bürgerinnen und Bürger in allen ihren Unterschiedlichkeiten frei entfalten können und gegenseitig die Freiheit des oder der jeweils Anderen respektieren. Entfaltungsbedingungen für ein selbstbestimmtes Leben in Freiheit bedeuten beispielsweise, dass technische Systeme den Handlungsspielraum des Menschen nicht ohne einen ethisch bedeutsamen Grund einschränken und beherrschen dürfen. Selbstbestimmung ist nicht ausschließlich individualistisch auszurichten. Der Mensch ist ein Beziehungswesen und entfaltet sein Leben in einem sozialen Miteinander mit vielfältigen wechselseitigen Verbindungen und Einflussnahmen.

Die Regeln dieses Miteinanders werden durch **kulturelle und sozialnormative Rahmenbedingungen** im gesellschaftlichen Zusammenleben über die Zeit geprägt. Zudem werden sie durch Recht gestaltet, in einer demokratischen Gesellschaft vor allem dort, wo Macht- und Informationsungleichgewichte herrschen.

Je mehr Informationen Dritte über den Einzelnen gesammelt haben, desto schwieriger wird es, in sozialen Situationen unbefangen zu agieren oder sich gar als Individuum ganz neu zu erfinden. Verhindert werden muss, dass Praktiken der Datensammlung und -auswertung persönliche und soziale Profile routinemässig an vielen Stellen erstellen und dauerhaft „zementieren". Insofern umfasst Selbstbestimmung auch ein **Recht, die eigene Identität auszubilden und zu ändern** und damit auch die Möglichkeit eines neuen Anfangs. Auch die Entscheidung darüber, wie ein Individuum in der Öffentlichkeit auftritt sowie der Schutz gegen eine falsche Darstellung in der Öffentlichkeit sind daher vom Recht auf Selbstbestimmung umfasst.

Selbstbestimmung bedeutet auch, dass Menschen nicht nur **Verantwortung** übernehmen dürfen, vielmehr müssen sie sie übernehmen und müssen ihr auch gerecht werden. Verantwortung liegt nie bei einer Maschine, sondern immer beim Menschen, gegebenenfalls im Rahmen von institutioneller Verantwortung. Auch wenn ein technisches System eingesetzt wird, um im Rahmen einer automatisierten Datenauswertung Schlussfolgerungen wie die Gewährung eines Kredites anzuwenden, ist es die Verantwortung des Menschen, dieses System in einer ethisch vertretbaren Weise zu entwickeln und einzusetzen.

Eine wichtige Ausprägung der Selbstbestimmung ist die **informationelle Selbstbestimmung**. Sie umfasst das Recht des Einzelnen zu bestimmen, wer wann und zu welchem Zweck welche personenbezogenen Daten erheben und verwenden darf. Durch die informationelle Selbstbestimmung kann der Einzelne seine Handlungsfreiheit und Privatheit in dem Umfang schützen, wie es ihm wichtig ist, und er kann auch im öffentlichen Raum bestimmen, als welche Persönlichkeit er wahrgenommen und behandelt werden möchte.

Im Zeitalter der Digitalisierung kommt dem Einzelnen über seine informationelle Selbstbestimmung hinaus als selbstbestimmtem Akteur in der Datengesellschaft eine besondere Bedeutung zu. Hierauf wird durch den Begriff der **digitalen Selbstbestimmung** Bezug genommen. Diese schließt die Kompetenz ein, selbst zu bestimmen, mit welchen Inhalten jemand in Beziehung zu seiner Umwelt tritt und wie jemand die eigene Persönlichkeit interaktiv entfaltet. Sie umfasst unter bestimmten Bedingungen etwa auch die selbstbestimmte wirtschaftliche Verwertung der eigenen Datenbestände sowie den selbstbestimmten Umgang mit nicht-personenbezogenen Daten, die durch den Betrieb eigener Geräte generiert werden. Digitale Selbstbestimmung geht dabei immer auch mit digitaler Selbstverantwortung einher.

Nach Auffassung der DEK gilt ein Recht auf digitale Selbstbestimmung auch für **Unternehmen und juristische Personen**. Juristische Personen können sich nicht auf die im Rahmen des allgemeinen Persönlichkeitsrechts über Art. 1 Abs. 1 GG geschützte Menschenwürde und somit nicht auf den damit verbundenen absolut geschützten Kernbereich der Persönlichkeitsentfaltung berufen. Allerdings verfügen juristische Personen über ein durch Art. 2 Abs. 1 GG in Verbindung mit Art. 19 Abs. 3 GG geschütztes Persönlichkeitsrecht, das auch ein informationelles Selbstbestimmungsrecht beinhaltet.

Im Hinblick auf Verbraucher sind ihre Selbstbestimmung und die Ermöglichung bewusster Konsumentscheidungen Voraussetzungen einer optimalen Ressourcenallokation und Wohlfahrtsmaximierung innerhalb der Volkswirtschaft. Ein Erodieren der für die Selbstbestimmung erforderlichen **Kompetenzen von Verbrauchern**, etwa durch einen übermäßigen Einsatz von Entscheidungsassistenten und damit verbundene Habituationseffekte, wirft ethische Fragen zur Fremdbestimmung und Entscheidungsfreiheit von Individuen, aber auch gesellschaftlicher Steuerung durch einzelne marktmächtige Akteure auf.

3.3 Privatheit

In enger Verbindung mit dem **Schutz der Menschenwürde und der Selbstbestimmung** steht wesentlich auch der Schutz der Privatheit. Die ethisch hochrangige Bedeutung, die eigene Privatsphäre bewahren zu können und sich in der Gewissheit der geschützten Privatheit auch in der Öffentlichkeit bewegen zu können, begründet das Recht des Einzelnen, darüber zu bestimmen, wer welche persönlichen Informationen zu welchem Zeitpunkt und zu welchem Zweck erhalten darf (→ informationelle Selbstbestimmung, oben 3.2). Die gesetzliche Regelung eines verantwortungsvollen Umgangs mit persönlichen Daten gehört zum Schutz der Würde des Menschen.

Privatheit umfasst darüber hinaus die **Wahrung der Integrität der persönlichen Identität**. Diese kann beispielsweise verletzt werden, wenn algorithmische Systeme anhand von Daten, die zu ganz anderen Zwecken entstanden sind, die Persönlichkeit eines Menschen, seine Präferenzen und Neigungen gleichsam ausrechnen, um dies unabhängig von oder sogar gegen seinen Willen zu eigenen Zwecken zu nutzen.

In einer Gesellschaft, deren unterschiedlichen Bereiche zunehmend durch datengetriebene Technologien geprägt werden, gilt es, die **Aufmerksamkeit zunehmend auf die Verwendung der Daten zu richten**. Viele Menschen geben auch personenbezogene Daten der Öffentlichkeit oder Teilen davon preis, da sie bestimmte Produkte und Services genießen oder einen Beitrag zum öffentlichen Wohl leisten wollen. Sie nur darauf zu verweisen, dass sie sparsam mit der Freigabe ihrer Daten umgehen sollen, hilft hier nicht weiter. Vielmehr müssen sie sich durch eine wirksame Regulierung darauf verlassen können, dass mit ihren Daten verantwortungsvoll umgegangen wird und ethisch unzulässige Verwendungen verboten sind.

3.4 Sicherheit

Algorithmische Systeme werfen zudem wichtige Fragen der Sicherheit auf. Je nach Anwendungskontext kann die Sicherheit der Nutzer gefördert oder gefährdet werden. Die ethische und rechtliche Relevanz von Sicherheit besteht in ihrer **Funktion, hochrangige Güter zu schützen**, wie etwa die körperliche und psychische Gesundheit und die Privatheit von Individuen oder auch die öffentliche Sicherheit, den Frieden sowie die Freiheit und Gleichheit demokratischer Wahlen.

Sicherheit kann sich auf die Datenerhebung und -verwendung beziehen und betrifft damit auch den **Schutz der Privatheit**. Datenskandale großen Ausmaßes, die in den letzten Jahren bekannt wurden, haben deutlich gemacht, dass sich die Verletzung der Privatheit und die Verwendung personenbezogener Daten zu manipulativen Zwecken bis in den Bereich der Politik mit weitreichenden Folgen auswirken kann.

Auch die **körperliche und emotionale Sicherheit** des Menschen bei der Bedienung und bei der Anwendung algorithmischer Systeme ist zu bedenken. Dies führt zu hohen Anforderungen beispielsweise in der Mensch-Maschine-Interaktion. So ist etwa beim Einsatz eines Pflegeroboters sicherzustellen, dass sowohl die zu pflegende als auch die pflegende Person dadurch in ihrer körperlichen sowie psychischen Integrität nicht geschädigt werden.

Darüber hinaus kann die **Sicherheit der Umwelt** berührt sein. Bei algorithmisch gesteuerten öffentlichen Infrastrukturen wie etwa dem Verkehr oder der Energie- und Wasserversorgung kann es bei Fehlfunktionen zu massiven Schädigungen kommen.

Zudem können algorithmische Systeme in sich unsicher sein und damit Fehlfunktionen verursachen oder sogar **Einfallstore für Angriffe und Manipulationen** in bösartiger Absicht bieten. Auch über eine solche systeminterne Anfälligkeit hinaus ist die Problematik des Missbrauchs eines algorithmischen Systems für schädliche Zwecke zu bedenken.

3.5 Demokratie

Digitale Technologien sind auf komplexe Art und Weise für die Entfaltung der Grundrechte (insbesondere die Meinungs- und Informationsfreiheit, das (informationelle) Selbstbestimmungsrecht und das Fernmeldegeheimnis, die Versammlungs- und Vereinigungsfreiheit sowie die Berufsfreiheit und das Eigentumsrecht), für die Demokratie, für die Sicherung von Vielfalt, für eine offene gesellschaftliche Debatte sowie für freie und gleiche Wahlen **systemrelevant**. Beispielsweise ermöglichen soziale Medien eine niedrigschwellige und grundsätzlich begrüßenswerte Beteiligung aller Bürgerinnen und Bürger an einer Debatte über die Gestaltung unserer Zukunft. Sie können aber auch Gefahren im Hinblick auf Manipulation und Radikalisierung mit sich bringen. Dem sollte der Staat durch Regeln und Institutionen, die Fehlentwicklungen und missbräuchliche Verwendung verhindern, entschieden entgegentreten.

Auch ist nicht zu verkennen, dass mit der Verbreitung des Internets der wirtschaftliche Niedergang des Journalismus und seiner privat finanzierten Pluralität einhergeht. Die elektronische Öffentlichkeit ersetzt aber in keiner Weise die für die Demokratie wichtige Funktion des Journalismus als „vierte Gewalt“ bzw. als „Wachhund der Demokratie“, also der Kontrolle von Macht und Wahrheitsanspruch durch systematische und unabhängige Nachforschung und Kritik. Die Gefahr des steuernden Einflusses machtvoller **Medienintermediäre mit Torwächterfunktion** für die demokratische Willensbildung kann unter bestimmten Umständen eine erhebliche Bedrohung für die Demokratie darstellen, der aus ethischen und verfassungsrechtlichen Gründen gesetzlich entgegenzuwirken ist.

Auch **Erziehung und Bildung** spielen bei der Sicherung einer freiheitlich-demokratischen Grundordnung eine herausragende Rolle, da sie auf vielfältige Weise die für eine Demokratie konstitutive, kritische Beteiligung der Bürgerinnen und Bürger an der Gestaltung der Gesellschaft, das Verständnis und die Einschätzung gesellschaftlich relevanter Zusammenhänge und Entwicklungen und damit auch letztlich das Vertrauen in eine wertebasierte, gestaltbare Zukunft beeinflusst. Durch Erziehung und Bildung zu vermittelnde Kompetenzen betreffen sowohl technische und mathematische als auch ethische, rechtliche, ökonomische und sozialwissenschaftliche Aspekte.

3.6 Gerechtigkeit und Solidarität

Für ein freiheitlich-demokratisches Zusammenleben in Frieden und Wohlstand ist unter anderem konstitutiv, dass die Gesellschaft und ihre Institutionen die Prinzipien der Gerechtigkeit umsetzen. Angesichts der massiven daten- und technologieinduzierten Anhäufung von wirtschaftlicher und damit auch gesellschaftlicher Gestaltungsmacht bei wenigen großen Unternehmen stellen sich neue Fragen einer gerechten Wirtschaftsordnung. Aber auch weitere Fragen von **Zugangs- und Verteilungsgerechtigkeit**, etwa bei Einkommen und im Gesundheitswesen, können durch die Verfügbarkeit großer Datenmengen und die Digitalisierung von Prozessen wie in der Arbeitswelt und der medizinischen Versorgung betroffen sein – im Sinne einer gerechteren Verteilung knapper Ressourcen, aber auch im Sinne einer Benachteiligung oder Diskriminierung bestimmter Personengruppen.

Gerechtigkeit ist zudem eng verbunden mit der Möglichkeit zur Beteiligung. Durch die – auch digital unterstützte – **Stärkung partizipativer Prozesse** kann bei technologieinduzierten sozialen Umbrüchen ein wichtiger Beitrag zur Beförderung sozialer Innovationen geleistet werden. Ein gerechtigkeitsrelevantes Problem besteht nicht zuletzt in einer Diskriminierung von Personen oder Personengruppen, die – insbesondere bei Anwendung selbstlernender algorithmischer Systeme – ohne rechtfertigende Gründe benachteiligt werden.

Die klare **Zuordnung von Verantwortung und Rechenschaftspflichten** ist in einem demokratischen Rechtsstaat unverzichtbar. Es bedarf einer ausreichenden Transparenz und Erklärbarkeit, um eine Überprüfbarkeit algorithmischer Systeme in Abhängigkeit von ihrem konkreten Schädigungspotenzial zu gewährleisten. Zudem muss es Möglichkeiten geben, unter bestimmten Voraussetzungen den Rechtsweg zu beschreiten und jemanden gegebenenfalls in die Verantwortung nehmen zu können, also haften zu lassen.

Der **Zugang zu digitalen Ressourcen** über das Internet ist heute eine elementare Voraussetzung digitaler und damit auch sozialer Teilhabe. Den Staat trifft als Teil seines Gewährleistungsauftrages die Pflicht, dafür zu sorgen, dass Bürgerinnen und Bürger flächendeckend sowohl stationär als auch mobil in angemessenem Umfang auf eine zeitgemäße Internetinfrastruktur zugreifen können. Sein Bildungsauftrag ist es, die Bürgerinnen und Bürger zu befähigen, sich selbstbestimmt in der digitalen Welt zu bewegen und Chancen sowie Risiken der Internetnutzung richtig einschätzen zu können.

Teilhabemöglichkeiten fördern auch den **sozialen Zusammenhalt**. Dieser basiert zudem auf einer Grundhaltung und einer institutionellen Verankerung von gesellschaftlicher Solidarität. Digitale Technologien können zu ihrer Stärkung beitragen, sie können sie aber auch schwächen oder zerstören. Bei der Anwendung von algorithmischen Systemen in bestimmten gesellschaftlichen Bereichen wie etwa im Versicherungswesen oder in der Vermittlung von sozialen Teilhabechancen ist auf zum Teil subtile Effekte mit der Folge einer systemischen Schwächung von Solidarität zu achten. So können nachvollziehbare und individuell gerechtfertigt erscheinende datengetriebene Differenzierungen und Ungleichbehandlungen durchaus in der Summe zu einer Entsolidarisierung mit bestimmten Personengruppen führen – auch solchen, die auf eine gesellschaftliche Unterstützung in besonderem Maße angewiesen sind.

3.7 Nachhaltigkeit

Digitale Technologien bringen große Chancen für eine effizientere Ressourcenbewirtschaftung und innovative Geschäftsmodelle mit sich. In der allgemeinen Diskussion steht dieser ökonomische Aspekt meist im Vordergrund. Weniger diskutiert wird aber bislang, ob die digitalen Technologien auch zu ökonomischer Nachhaltigkeit beitragen. Zusätzlich sind die Aspekte ökologischer und sozialer Nachhaltigkeit zu bedenken. Die Vereinten Nationen haben **17 Ziele für nachhaltige Entwicklung auf ökonomischer, sozialer und ökologischer Ebene** formuliert, die für alle Staaten gelten und bis 2030 umgesetzt sein sollen. Digitale Technologien können dazu beitragen, diese nachhaltigen Ziele zu verwirklichen, wie es etwa von der International Telecommunication Union (ITU) mit „AI for good“ verfolgt wird. So hat jüngst der Wissenschaftliche Beirat der Bundesregierung Globale Umweltveränderungen (WBGU) die Vision einer KI-basierten feingranularen Umweltsensorik entworfen, die ein bisher ungekanntes „umfassendes und echtzeitnahes Monitoring der natürlichen Erdsysteme, ihrer Zustände und ihrer Entwicklung“ ermöglichen und damit einen zentralen Baustein für eine künftige digitale Nachhaltigkeitspolitik bilden soll.

Digitale Technologien fördern allerdings nicht nur Ressourcen, sie erfordern auch Ressourcen etwa durch einen immer stärker anwachsenden Bedarf an elektrischer Energie und durch die Angewiesenheit digitaler Produkte auf bestimmte „seltene Erden“, die nur noch begrenzt und nur in bestimmten Staaten verfügbar sind. Zudem geht ihr Abbau mit massiven ökologischen Schäden einher. Das wirft Fragen in Bezug auf nachhaltige ökonomische und ökologische Entwicklung auf und berührt zusätzlich **Fragen der internationalen Gerechtigkeit** beim Umgang mit natürlichen Ressourcen sowie der globalen Verantwortung für künftige Generationen.

Nachhaltigkeit ist auch gefordert, was die Ressource menschlichen Wissens und menschlicher Kompetenz betrifft: In dem Maße, wie sich digitale Technologien entwickeln und dem Menschen Aufgaben abnehmen, werden nicht nur neue Kompetenzen hinzugewonnen, sondern es gehen auch **Kompetenzen des Menschen** verloren. Dies erfordert eine Diskussion, welche Verantwortlichkeit gegenüber der nächsten Generation besteht, und Maßnahmen, bestimmte Kompetenzen und Unabhängigkeiten zu bewahren und zu entwickeln.

Die umfassende und regelmäßige **Technikfolgenabschätzung**, wie sie dieses Gutachten an mehreren Stellen einfordert, wird auch die Aspekte der Nachhaltigkeit der neuen Technologien in ihren verschiedenen Ausformungen mit einbeziehen müssen. Der Gesetzgeber ist hier gefordert, Verantwortung für Nachhaltigkeit in die Regulierung der Datenwirtschaft und der algorithmischen Systeme einzubauen, beispielsweise durch die Einführung einer Pflicht zur Offenlegung der gesamten Energiebilanz eines energieintensiven Blockchain-Systems.

Zudem sollten **öffentliche Investitionen** in Datenwirtschaft und algorithmische Systeme insbesondere darauf ausgerichtet sein, Nachhaltigkeitsziele zu verfolgen, wie sie die Vereinten Nationen formuliert haben. Die Entwicklung von Daten und algorithmischen Systemen etwa zur Erfassung und Kontrolle von Umwelteinwirkungen und Entwicklungen in der Umwelt wie auch Systeme zur Optimierung und Reduzierung von Energie- und Ressourcenverbrauch sollten im Vergleich zu nur kurzfristigen wirtschaftlichen Gewinnen vorrangig Gegenstand öffentlicher Förderung sein. Auch sollten nachhaltigkeitsorientierte soziale Innovationen verstärkt gefördert werden, die gesellschaftliche Kreativität und Partizipation stärken.

Teil C

Technische Grundlagen

Datenintensive IT-Anwendungen haben unser Zusammenleben, unsere Arbeitswelt, Wirtschaft, Wissenschaft und Gesellschaft nachhaltig beeinflusst. Smartphones sind allgegenwärtige Begleiter, wir nutzen täglich Suchmaschinen, verlassen uns auf Empfehlungssoftware, schicken Text- oder Sprachnachrichten an Familie und Freunde, regulieren die Temperatur im Haus von der Ferne oder lassen uns durch Navigationsgeräte von einem Ort zum anderen leiten. Diese Möglichkeiten beruhen auf einer Reihe technologischer Entwicklungen der letzten Jahrzehnte. Im Folgenden sollen wichtige technologische Grundlagen beschrieben werden. Ziel ist hierbei nicht eine vollumfängliche Darstellung, sondern das Herausarbeiten zentraler Elemente, um resultierende Probleme und Ansatzpunkte für mögliche Governance identifizieren zu können.

1. Status Quo

Die Leistungssteigerung und Verkleinerung der physischen Komponenten von IT-Systemen (Hardware) zur Speicherung und Verarbeitung von Daten, zusammen mit einer sich stets verbessernden Konnektivität – sowohl kabelgebunden als auch über Funk – eröffnen die **Erschließung ganz neuer Anwendungsbereiche**. Smartphones, Tablets, Wearables durchdringen zusammen mit Sensoren, Aktuatoren und teilweise „autonom" agierenden Systemen, wie z. B. Robotern, die Arbeits- und Lebenswelt. So ist in weiten Teilen eine permanente und mobile Nutzung des Internets möglich, die beispielsweise in Kombination mit umfangreicher Sensorik in Smartphones (Geolokalisierung, Gyrosensoren, Kameras, Mikrofon usw.) neben Texteingaben auch Bild-, Video- und Audioaufnahmen fast von jedem Ort zu jedem Zeitpunkt im Internet bereitstellen kann. Neben der sozialen Vernetzung und Kommunikation ermöglicht diese technologische Durchdringung darüber hinaus die Vernetzung von Geräten zum sog. Internet der Dinge (engl. Internet of Things, IoT).

Die analoge Welt und die digitale Welt lassen sich nicht mehr genau trennen: Die analoge Welt enthält zunehmend Komponenten, die Informationen aus der analogen Welt in die digitale Welt weitergeben, doch ebenso werden auch digitale Informationen in der analogen Welt zur Verfügung gestellt, so dass beide Welten mehr und mehr zu einer **hybriden Welt** verschmelzen.

Bedingt durch umfangreiche Sensorik, durch das IoT und immer günstigere Datenspeicher ergibt sich ein **exponentiell ansteigendes Datenaufkommen**. Die Verarbeitung dieser großen Datenmengen erfordert spezialisierte Werkzeuge. Gleichzeitig hat dieses Datenaufkommen zusammen mit der leistungsfähigen Hardware die breitflächige Anwendung von Verfahren des maschinellen Lernens befördert. Diese Verfahren erzielen teilweise beeindruckende Ergebnisse, z. B. im Bereich des Sprach- und Bilderkennens.

Die Leistungssteigerung in der Spracherkennung und Videoverarbeitung reicht aber mittlerweile auch so weit, dass die **Grenzen zwischen der Wirklichkeit und computergenerierten Informationen** verschwimmen können. Dann ist es für Menschen nicht mehr klar, ob sie mit einem Sprach-Bot reden oder ob sie ein generiertes Video anschauen, in dem Menschen Worte in den Mund gelegt wurden, die diese nie gesagt haben (sog. Deep Fakes).

2. Systemelemente

2.1 Daten

2.1.1 Begriff und Eigenschaften von Daten

Aufgrund des Arbeitsauftrags der DEK liegt der Fokus des Gutachtens auf Daten, die **digital und maschinenlesbar** sind. Die Basis dieser Daten sind binäre, elektrische Impulse. Diese Impulse können nur für den Augenblick als Signal existieren, etwa als ein Steuerungsimpuls für ein technisches System, oder auch persistieren, d. h. auf einem Medium gespeichert sein.

Daten sind vielfältig. Der Begriff „Daten" versammelt eine immense Diversität von Erscheinungsformen unter einem Begriffsdach. So lassen sich Daten etwa anhand des Datentyps (z. B. binäre, nominale, ordinale, metrische und textuelle Daten), des datengenerierenden Prozesses (z. B. Umfragedaten, Sensordaten), des Erhebungsbereichs (z. B. Finanzdaten, Wetterdaten) oder ihrer Funktion in einem digitalen System (z. B. Log-in-Daten, Trainingsdaten) einteilen. Eine weitere Einteilung setzt am Grad der Verarbeitung (Veredelung) an: Ohne eine weitere Verarbeitung spricht man auch von „Rohdaten", je nach dem Grad der Strukturierung (Normalisierung) von „strukturierten" oder „unstrukturierten" Daten. Daten können der Input in ein System sein oder auch der Output, der wiederum der Input in das nächste System sein mag. Daten können zugleich digitale Vermögensgüter (digital assets) repräsentieren, wie multimediale Inhalte oder Einheiten von Kryptowährungen. Von erheblicher juristischer Bedeutung ist zudem die Differenzierung zwischen personenbezogenen und nicht-personenbezogenen Daten.

Daten sind nicht immer Information. Um die **binären, elektrischen Impulse**, welche die Basis für digitale Daten darstellen, zu verstehen, d. h. um aus Daten „Information" werden zu lassen, ist es erforderlich, den **Kontext** und die **Semantik** (Bedeutung) zu kennen. Der Kontext kann durch den Ursprung der Signalerzeugung gegeben sein, beispielsweise die Kenntnis, von welchem Sensor ein Signal gesendet wird. Die Semantik gibt an, welche Information in einer gewissen Abfolge von binären Signalen liegt, z. B. kann bei einer Umfrage die Ziffer 4 die Anzahl der Kinder im Haushalt oder genauso gut die Anzahl der im letzten Halbjahr gekauften Zahnpastatuben sein. Kontext und Semantik findet man in Metadaten, Domaintabellen, Ontologien, Identifikatoren und weiteren die Datenwerte ergänzenden technischen Spezifikationen. In diesem Gutachten ist mit dem Begriff Datum immer auch die Kenntnis von Kontext und Semantik gemeint.

Daten sind von unterschiedlicher Qualität. Die meisten Daten – oder besser gesagt die in ihnen enthaltene Information – sollen die Realität möglichst getreu abbilden, indem etwa den richtigen Entitäten (Informationsobjekten) diejenigen Attribute zugeordnet werden, die diese Entitäten auch in der Realität aufweisen. Dabei kann es zu Fehlern kommen. Es gibt auch viele Daten, die eine Wahrscheinlichkeit für die Realität, oder für eine künftige Realität, zum Ausdruck bringen sollen, oder auch Daten, die eine hypothetische Realität konstruieren sollen oder gar keinen Bezug zur Realität haben. In allen Fällen kann der Datenbestand **fehlerbehaftet** sein. Davon zu unterscheiden sind Situationen, in denen Daten das leisten, was sie vorgeben zu leisten, diese Leistung aber **ungeeignet** ist, um ein bestimmtes Ziel zu erreichen, etwa eine bestimmte Analyse durchzuführen (z. B. die Daten sind nicht granular genug oder zu alt oder nicht vollständig genug).

Entscheidend für datengetriebene Systeme ist die Qualität der verwendeten Daten: Selbst ein perfekter Algorithmus wird keine Qualität liefern, wenn er schlechte, d. h. ungenaue oder inadäquate Daten als Eingabe erhält. Datenqualität ist kein absoluter Wert, die relevanten Datenqualitätsdimensionen und deren Qualitätslevel sind abhängig von der spezifischen Verwendung (→ s. Abb. 1).

Abbildung 1: Beispiel für unterschiedliche Qualitätsanforderungen je nach Verwendung

2.1.2 Data Management

Daten sind nicht gegeben, sondern gemacht. Im Prozess der Erhebung, Aufbereitung und Verarbeitung von Daten treffen Menschen vielfältige Entscheidungen, welche Konsequenzen für die weitere Datennutzung haben. Fehlt beispielsweise zu einem Datensatz Kontext oder Semantik, kann das Potenzial dieser Daten unwiederbringlich verloren sein. Um dies zu vermeiden, ist sorgfältiges **Data Management** erforderlich.

Möchte man Daten aus unterschiedlichen Quellen zusammenführen, ist sicherzustellen, dass sowohl auf technischer als auch auf semantischer Ebene eine solche Zusammenführung möglich ist. Dies bezeichnet man als Interoperabilität. Es gilt, eine Abbildung der Daten aus diesen Quellen aufeinander zu finden, die der jeweiligen Semantik der Daten gerecht wird. Ist die Interoperabilität von besonderer Bedeutung, sollte eine **Standardisierung** bezüglich der technischen Spezifikationen (Formate, Metadaten zur Beschreibung usw.) angestrebt werden. Dabei spielen etwa Referenzdaten eine wichtige Rolle, also standardisierte Schemata oder Ontologien, die beispielsweise von nationalen oder internationalen Institutionen verantwortet werden (z. B. die von der WHO herausgegebene internationale Klassifikation von Krankheiten).

2.1.3 Big Data und Small Data

Nicht um einen eigenen Datentyp, sondern um einen neuen methodischen Ansatz zum Auffinden von Zusammenhängen handelt es sich bei **Big Data**. Eine besonders bekannte, frühe Definition von Big Data geht auf Laney[1] zurück, welcher Big Data durch drei Vs charakterisiert: *volume, velocity und variety* – eine große Menge an vielfältigen Daten aus potenziell unterschiedlichen Quellen, die mit hoher Geschwindigkeit (oft in Echtzeit) generiert werden. Um in der Lage zu sein, vielfältige, in ihrer Qualität variierende, sich schnell verändernde große Datenmengen zu bearbeiten, bedarf es besonderer Technologien. Besonders geeignet ist die Analyse großer Datensätze (Big Data), wenn aus einer Vielzahl möglicher Zusammenhänge Hinweise auf diejenigen ermittelt werden sollen, die vielversprechend sind. Beispielsweise ist es für die medizinische Forschung hilfreich, aus der Vielzahl von Umweltfaktoren, die eine Krankheit möglicherweise begünstigen, mit Hilfe von Big Data zunächst einige wahrscheinliche Kandidaten zu identifizieren und sodann nur für diese aufwändige und exakte Experimente oder Studien durchzuführen. Ein besonderes Problem dieses Ansatzes ist es, dass er zunächst einmal nur **Korrelationen**, aber keine Kausalitäten aufzeigt. Es können daher auch ganz falsche Kandidaten identifiziert werden.

1 Doug Laney: 3D Data Management. Controlling data Volume, Velocity and Variety, META Group Inc., 2001.

In vielen Bereichen werden niemals ausreichend große Datenmengen vorliegen, die mit den Big-Data-Methoden analysiert werden können (z. B. mag der Kundenstamm eines mittelständischen Unternehmens nie größer als 200 Kunden werden, die Anzahl an Parteien in einem Land ist selten dreistellig). Auch für den Bereich der **Small Data** kann mit geeigneten Analysemethoden viel Wissen und Information aus den Daten extrahiert werden. Nicht die Menge an Daten ist entscheidend. Es gilt vielmehr, Daten mit adäquater Qualität und in für die Fragestellung ausreichender Menge mit geeigneten Werkzeugen zu kombinieren, um gute Datenanalysen zu ermöglichen.

2.2 Datenverarbeitung

2.2.1 Algorithmen

Während im Datenschutz unter **Verarbeitung** die Gesamtheit des Prozesses von der Datengenerierung über die Extraktion zur Speicherung und jedwede Transformation der Daten selbst bezeichnet wird (Art. 4 Nr. 2 DSGVO), verwenden die mathematisch-technischen Disziplinen den Begriff in erster Linie, um die Nutzung der Daten zu bezeichnen. Dieses Verständnis liegt auch den folgenden Ausführungen zu Grunde.

Jede digitale Datenverarbeitung folgt dem **EVA (Eingabe-Verarbeitung-Ausgabe)-Prinzip**: Daten gehen als Eingabe (Input) ein, werden verarbeitet und als Ergebnis ausgegeben. Jede interne Verarbeitung in einem EVA-System beruht auf einem Algorithmus: einer operativen Verarbeitungsvorschrift, die einen Ablaufplan als eine Folge von Verarbeitungsschritten spezifiziert, um ein angestrebtes Ergebnis durch die schrittweise Transformation der Eingangsdaten zu erzielen. Schon Euklid spezifizierte einen Algorithmus als Rechenanleitung zum einfachen Auffinden des größten gemeinsamen Teilers zweier natürlicher Zahlen. Der Begriff leitet sich aus dem Namen des arabischen Mathematikers al-Chawarizmi (latinisiert Algorismi) ab, der um 830 n. Chr. eine Sammlung von Rechenvorschriften für das Lösen algebraischer Gleichungen veröffentlicht hat.

Insbesondere in der modernen Informatik ist der Begriff „Algorithmus“ von fundamentaler Bedeutung. Wenn man eine gegebene Fragestellung mittels Datenverarbeitung beantworten möchte, muss man einen Algorithmus sowohl korrekt implementieren als auch produktiv einsetzen. Dies setzt die Kenntnis des Algorithmus voraus. In vielen Situationen ist der zum Ziel führende Algorithmus allerdings noch nicht bekannt und die Kernaufgabe besteht zunächst darin, **einen geeigneten Algorithmus zu finden**. Für viele praktisch relevante Situationen können die Verarbeitungsvorschriften aus Fachwissen, bekannten Modellen oder auch Rechtsvorschriften direkt abgeleitet, d. h. deduziert werden. In anderen Situationen ist unser Verständnis des Zusammenhangs noch nicht weit genug ausgereift, um diesen in mehr oder weniger einfachen mathematischen Formeln beschreiben zu können.

Fehlt dieser Kenntnisrahmen, gibt es diverse Strategien, um einen Algorithmus zu finden, wie z. B. per Zufall, Versuch und Irrtum oder durch **Schließen** (auch Ableiten) aus Daten. Letztere Herangehensweise folgt dem Prinzip der Induktion: Aus Einzelfällen wird versucht, auf eine übergeordnete Regel zu schließen. Hierbei dienen die Daten als Einzelfälle. Findet man eine übergeordnete Regel, die zur Lösung der Fragestellung geeignet ist, hat man damit einen geeigneten Algorithmus gefunden. Dabei ist zu beachten, dass es durchaus mehrere zur Lösung geeignete Regeln geben kann. Des Weiteren ist eine Induktion nicht zwingend korrekt. Es ist möglich, aus vorliegenden Einzelfällen zu einem Schluss zu kommen, der teilweise oder vollständig falsch ist.

2.2.2 Statistisches Schließen

Das Schließen aus Daten ist das Kerngebiet der Statistik. Die **Verfahren des statistischen Schließens** aus Daten können sowohl zur Bearbeitung von Fragestellungen angewandt werden, deren inhärente Logik unbekannt ist, als auch insbesondere in Situationen, in denen der Zufall ein Teil des zu modellierenden Prozesses ist, beispielsweise um die Regenwahrscheinlichkeit für den Folgetag abzuschätzen oder Personen zu identifizieren, die mit einer hohen Wahrscheinlichkeit ein Produkt kaufen. Es gibt viele Methoden des statistischen Schließens: Die Spanne reicht von diversen Formen der Regression (lineare, logistische oder regularisierte Ridge Regression) über Support Vector Machines (SVM), Bayesian Networks und Regellernern (z. B. Aprioiri, CART und Random Forest) bis hin zu Neuronalen Netzen (NN). Alle diese Verfahren eignen sich zur Extraktion von Information aus vorliegenden Daten. Einige sind darauf spezialisiert, Regressionsfragestellungen zu lösen, etwa die Frage nach der erwarteten Körpergröße eines Kindes in Anbetracht der Größe der Eltern. Andere – zum Beispiel SVM, CART, NN - werden eingesetzt, wenn es sich um Klassifikationsfragestellungen handelt: z. B. schwanger oder nicht schwanger, Hund oder Katze. Ihre Eignung für den Einsatz zur Bearbeitung einer Fragestellung hängt dabei von vielen Faktoren ab, unter anderem dem Umfang und der Art der Daten.

Neben den Methoden zur Induktion verfügt die Statistik über ein breites Spektrum an **Verfahren und Maßen, um die Qualität des Ergebnisses zu bewerten**. Mithilfe dieser Maßzahlen ist es möglich, etwaige Fehler sowohl abzuschätzen als auch während des Einsatzes zu kontrollieren. Die Frage nach der erwarteten Körpergröße eines Kindes kann so mit 175 cm +/-4 cm beantwortet werden. Die Sicherheit, dass das Ergebnis eines Schwangerschaftstests positiv ist, kann bei 93 % liegen. Kontrollieren lässt sich – am Beispiel des Schwangerschaftstests – die Anzahl von Fehlalarmen, bei denen die Frau nicht schwanger ist, und fehlenden Alarmen, bei denen eine schwangere Frau ein negatives Ergebnis erhält. Ein perfektes statistisches Verfahren würde zu keinem dieser Fehler führen. In der Praxis muss abgewogen werden, welcher Fehler schwerwiegender und daher exakt zu kontrollieren ist. Ist es kritischer, eine tatsächlich schwangere Frau zu spät zu informieren, oder ist es kritischer, Frauen fälschlicherweise eine Schwangerschaft anzuzeigen? Es ist nicht möglich, beide Fehlerarten gleichzeitig zu minimieren: je kleiner der eine, desto größer in aller Regel der andere. Die „Balance“ wird je nach Kontext unterschiedlich gewählt werden.

Die Bewertung der Qualität der Ergebnisse leitet sich aus Qualitätsmerkmalen der Methoden selbst ab. Für einige Methoden können sogar **Qualitätsgarantien** gegeben werden. So gibt es eine Gruppe von Schätzverfahren, die als UMVUE (Uniformly Minimum Variance Unbiased Estimator, gleichmäßig bester erwartungstreuer Schätzer) bezeichnet werden. Verwendet man einen solchen Schätzer, stellt man sicher, bei gegebener Datenlage das bestmögliche Ergebnis zu erhalten. Liefert eine Regression, deren Parameter durch Verwendung eines UMVUE-Schätzers ermittelt wurde, die erwartete Körpergröße des Kindes 175 cm +/-4 cm, gibt es keinen anderen Schätzer, der einen kleineren Fehlerbereich liefert. Ein anderes Beispiel ist die Garantie bei einer Support Vector Machine, dass es sich bei einem aus den Daten ermittelten Modell um das bestmögliche für die Methode handelt, sofern sich ein solches finden lässt. Für manche Verfahren liegen zur Zeit weder für das Modell an sich noch für die Schätzungen, die mit dem Modell erzeugt werden, fundierte Verfahren zur Bewertung der Qualität vor. Dies gilt insbesondere für die Methodenklasse der NN. Aber auch für NN können Qualitätsangaben gemacht werden. Besonders wichtig sind Maßzahlen, die angeben, wie gut ein Modell auf Grundlage bisher unbekannter Daten funktioniert. Das Modell wird auf einem Datensatz (Trainingsdaten) gelernt und auf einem anderen in seiner Güte bewertet (Testdaten). Mit dieser Herangehensweise können Modelle identifiziert werden, die nicht die übergeordnete Regel abbilden, sondern ihre Trainingsdaten auswendig gelernt haben. Man spricht in einem solchen Fall von overfitting. Ein überangepasstes Modell wird auf den Trainingsdaten wesentlich bessere Qualitätswerte erzielen als auf den Testdaten.

Viele Verfahren der Statistik können analytisch gelöst werden. Das bedeutet, dass die Fragestellung als eine mathematische Gleichung oder ein Gleichungssystem formuliert und durch – häufig geschicktes – Transformieren aufgelöst werden kann. Eine Vielzahl von Methoden lassen sich dagegen nicht direkt analytisch auflösen (bspw. wenn Zusatzbedingungen wie ein Regularisierungsterm hinzukommen, s.u.). In diesen Fällen kann man auf **Optimierungsverfahren** zurückgreifen, die sich in vielen kleinen Schritten an die Lösung herantasten. Optimierungsverfahren sind nicht notwendigerweise optimal, so kann es sich bei dem berechneten Ergebnis gegebenenfalls nur um ein lokales Optimum und nicht das (oder eines der) globale/n Optimum/a handeln.

Verschiedene Lösungsansätze: Analytische Verfahren und Optimierungsverfahren

Eine direkte analytische Lösung ist für Aufgaben möglich wie „Gesucht ist der Wert von y für die Gleichung mit $y=4 \cdot x+3$ mit $x=3$“.

Dies ist nicht möglich für die Aufgabenstellung: „Gesucht ist die Lösung für die lineare Gleichung $a \cdot x_1 + b \cdot x_2 + \cdots + h \cdot x_8 = y$, bei der möglichst viele Parameter $a, b, \ldots, g, h$ gleich 0 sind“.

Dafür wird ein Regularisierungsterm hinzugenommen: $\min((a \cdot x_1 + b \cdot x_2 + \cdots + h \cdot x_8 - y) + (\text{Anzahl Parameter} \neq 0))$.

Um Lösungen zu finden, verwendet man Optimierungsverfahren.

2.2.3 Maschinelles Lernen

Die Abgrenzung zwischen klassischer Statistik und dem erstmals durch Mitchell[2] definierten **Maschinellen Lernen** ist schwierig. Spätestens wenn Optimierungsverfahren (→ dazu soeben unter) für das Lösen des induktiven Schließens verwendet werden, bietet es sich an, von Maschinellem Lernen zu sprechen.

Die Ansätze für **Schätz- bzw. „Lern"strategien** des Maschinellen Lernens unterscheiden sich durch die Formulierung des zu lösenden Optimierungsproblems. Es wird zwischen verschiedenen Lernverfahren differenziert:

- Überwachtes Lernen: Für überwachtes Lernen ist es erforderlich, zu jeder Eingabeinformation (das „E" im EVA-Prinzip) die korrekte Ausgabe (das „A") zu kennen. Ein klassisches Beispiel ist die Körpergröße: Möchte man von der Körpergröße der Eltern (Eingabe) auf die Körpergröße der Kinder (Ausgabe) schließen, so muss für jedes Kind die Größe vorab bekannt sein. Bei Schwangerschaftstests muss das korrekte Ergebnis, bei Wettervorhersagen das Wetter, bei Bodenanalysen die Bodenbeschaffenheit usw. gegeben sein.
  In der Praxis besteht die Herausforderung oft in der Beschaffung und Qualitätssicherung dieser korrekten Ausgabeinformation. Diese Ausgabeinformation wird häufig als **Label** bezeichnet. Aktuell wird die Mehrzahl aller im Einsatz befindlicher durch Maschinelles Lernen spezifizierter Algorithmen mittels überwachten Lernens trainiert.
- Es ist entscheidend für diese Lernverfahren, wie das eigentliche Optimierungsproblem formuliert ist, welche Regularisierungen verwendet werden und wie die Verlustfunktion definiert ist (d. h. werden alle Fehler gleich behandelt oder gibt es unterschiedliche Gewichte, Schweregrade, z. B. im Vergleich von als False Negatives nicht erkannten Krebserkrankungen zu als False Positives fälschlich ausgewiesenen Krebserkrankungen?).

Qualität von Labeln

Label können ebenfalls fehlerbehaftet sein.
Man kann mehrere Komplexitätsstufen für die Feststellung der Labels definieren:

1. Label, deren Korrektheit zum Erhebungszeitraum überprüfbar ist. – Beispiel: Bei physikalischen Systemen oder Eigenschaften, der Geschwindigkeit eines Objektes, der Raumtemperatur, aber auch dem Geburtsdatum jedes Menschen, existiert nur ein richtiger zugehöriger Wert. Diese Werte können daher prinzipiell als Label durch einen Algorithmus ermittelt werden.

2. Label, deren Korrektheit zum Erhebungszeitpunkt und gegebenenfalls auch später nicht überprüfbar ist.

3. Label mit einem konstruierten und nicht überprüfbaren Bezug zur realen Welt. – Beispiel: Um Menschen und ihr Verhalten besser verstehen und analysieren zu können, wurden etwa Konzepte wie soziale Milieus oder Charaktertypen entwickelt. Diese Konzepte sind Abstraktionen, die eine Wahrheit – sofern diese existiert – nicht notwendigerweise korrekt abbilden.

2 Tom Mitchell: Machine Learning, McGraw-Hill, 1997.

Festsetzung des Optimierungsziels

Ein ÖPNV-Unternehmen plant, die Linienführung von Bussen der Stadtentwicklung anzupassen, weil viele Einwohner in die Randbereiche abgewandert sind, große innerstädtische Brachflächen erschlossen wurden und sich die Bevölkerungszusammensetzung in den Bezirken durch Gentrifizierung stark verändert hat. Der Projektleiter hat Daten zu Fahrgast- und Nutzungszahlen erhoben und arbeitet an einer Optimierung, um eine bedarfsgerechte Linienführung der Buslinien zu erreichen, ohne dass der Einsatz zusätzlicher Busse nötig wird. Für die Optimierung kommen verschiedene Ziele oder Nebenbedingungen infrage, z. B. Einsparung von Bussen, Einsatz von weniger Fahrerinnen und Fahrern, kein Schaffen neuer Routen. Abhängig von der Formulierung des Optimierungsproblems könnte es sein, dass beispielsweise dichtbesiedelte Stadtteile im Vergleich besser versorgt werden, dafür aber alle Bewohner von Außenbezirken etwas längere Fahrtzeiten oder geringere Frequenzen in Kauf nehmen. Da der Projektleiter selbst im Speckgürtel wohnt, ist ihm die Optimierung lieber, bei der die längste Fahrzeit minimiert wird. Dies führt dazu, dass schnellere Verbindungen auch in die Randbezirke entstehen. Seinem Vorgesetzten gefallen beide Modelle nicht. Er möchte, dass so viele Passagiere wie möglich transportiert werden. Dies führt dazu, dass gute Kurzstreckenverbindungen verstärkt werden, längere Fahrten über mehr als vier Stationen jedoch im Nachteil sind. Hier zeigt sich, dass die Entscheidung über die Optimierungsfunktion gesellschaftliche Auswirkungen haben kann. Es stellen sich u. a. folgende Fragen: Wer entscheidet über das Ziel der Optimierung? Wer sollte (mit)entscheiden? Wie kann der notwendige/sinnvolle gesellschaftliche Diskurs geführt werden? Welchen Rechtsschutz (z. B. Klagemöglichkeiten) genießen Gruppierungen/Stadtteile, die sich anderen gegenüber benachteiligt fühlen?

- Beim **Reinforcement Learning** werden Handlungen eines Agenten durch eine Bestrafung (Penalty) bzw. durch eine Belohnung (Bonus) bewertet. Ein Agent kann aus einer Menge von Handlungen auswählen und eine Handlung durchführen. Sein Handeln verändert den Zustand des Systems. Als Eingabe zur Optimierung dient die Handlung des Agenten. Zusammen mit dem Zustand bzw. der Zustandsänderung des Systems, die von der Handlung des Agenten herbeigeführt wird, muss eine eindeutige Bonusfunktion existieren. Während beim überwachten Lernen zu jeder Eingabe die korrekte, optimale Lösung vorliegt, ist dies bei Reinforcement Learning nicht unbedingt gegeben. Die Optimierung hat dabei die Aufgabe, diejenigen Handlungsstrategien zu finden, die zum besten Endzustand bezogen auf das Optimierungsproblem führen. Dabei kann es erforderlich sein, auf Handlungen, die eine kurzfristige Verbesserung liefern, verzichten zu müssen. Für diese Lernstrategie spielt neben dem eigentlichen Optimierungsproblem und der relevanten Verlustfunktion insbesondere die Bonusfunktion eine entscheidende Rolle.
- **Unüberwachtes Lernen** nutzt eine Menge von Eingabedaten und sucht in diesen Daten nach Strukturen. Die Kenntnis von korrekten Strukturen oder einer Bonusfunktion ist nicht erforderlich. Dagegen ist es notwendig, genau zu definieren, nach welcher Struktur gesucht werden soll. Beispielsweise kann nach Clustern, d. h. Gruppen in den Daten, gesucht werden, wobei die Anforderung darin besteht, dass alle Datenpunkte eines Clusters sich so stark wie möglich ähneln, wohingegen der Unterschied zwischen den Clustern maximiert werden soll. Daraus ergibt sich das Optimierungsproblem für das unüberwachte Lernen. Das unüberwachte Lernen wird auch als **Data Mining** bezeichnet.

Neben den Lernverfahren spielt die Bereitstellung der Daten eine entscheidende Rolle, da diese in hinreichendem Umfang, einer guten Qualität und angemessenen Breite zur Verfügung stehen müssen, um eine gute Näherung des Optimierungsziels zu erreichen. In vielen Fällen stehen Daten leider nicht im erforderlichen Umfang, der Breite oder Qualität zur Verfügung, so dass andere Wege eingeschlagen werden müssen, um dennoch gute Ergebnisse durch Maschinelles Lernen zu erzielen.

So ist es möglich, **synthetische Daten**, d. h. Daten, die künstlich generiert und nicht unmittelbar in der echten Welt erhoben wurden, zu verwenden. Sie haben mehrere Vorteile gegenüber den echten Daten:[3] Synthetische Daten können in beliebiger Menge produziert werden; dies ist besonders wichtig für Simulationen, wenn die echten Daten noch gar nicht angefallen sein können. Bei der Erzeugung kann dafür gesorgt werden, das gesamte Wertespektrum möglichst vollständig abzubilden, z. B. um das Verhalten eines technischen Systems auch bei ungewöhnlichen Datenkonstellationen zu testen. Die Qualität der synthetischen Daten ist messbar; je nach Bedarf kann im Einzelfall gewährleistet werden, dass die Eigenschaften eines Referenzdatenbestands aus der echten Welt erhalten bleiben, oder es können gezielt Verzerrungen, die in Echtbeständen vorkommen können, im Sinne einer Diskriminierungsvermeidung herausgenommen werden. Solange der synthetische Datenbestand keinen Personenbezug aufweist, ist er anonym, und die DSGVO ist nicht anwendbar. Synthetische Daten können für ein Training von Algorithmen oder ein Test von Systemen hilfreich sein. Jedoch besteht das Risiko, dass Eigenschaften der generierten Daten den Algorithmus beeinflussen, die keine Entsprechung in der Realität haben. Daher sind vor dem praktischen Einsatz gesonderte Funktionstests erforderlich.

Ein häufig genutzer Mittelweg ist die sogenannte **Augmentation**. Hierbei werden echte Daten so erweitert, dass im Training eine größere Menge an Konstellationen abgedeckt werden kann. Somit bleibt einerseits der Bezug zu echten Daten erhalten und andererseits wird eine Verbreiterung der Datenbasis erzielt. Augmentation beschreibt den Prozess, neue Daten zu generieren, die leicht von den Ursprungsdaten abweichen. Beispielsweise zeichnet sich ein augmentiertes Bild dadurch aus, dass es verschoben, rotiert oder verzerrt worden ist.

2.2.4 Künstliche Intelligenz

Im aktuellen Sprachgebrauch wird Maschinelles Lernen, weiter verengt auf Neuronale Netze, als **Künstliche Intelligenz (KI)** bezeichnet. Diese Bezeichnung kann durchaus für Verwirrung sorgen: Maschinelles Lernen ist nur ein spezielles Verfahren innerhalb der „schwachen KI", welche wohlspezifizierte Aufgaben löst. Im Gegensatz dazu wird von der „starken KI" erwartet, nicht nur eine Aufgabe, sondern ein breites Spektrum von Aufgaben, womöglich ohne Eingriffe eines Menschen, zu bewältigen. Dies leistet das Maschinelle Lernen entgegen der Erwartung, die der Begriff Künstliche Intelligenz weckt, nicht.

Historisch bezeichnet der Begriff Künstliche Intelligenz ein breites Forschungsgebiet innerhalb der Informatik, das bereits 1956 in den USA unter dem Namen **Artificial Intelligence** begründet wurde (Dartmouth Proposal).[4] Seit Gründung hat das Gebiet mehrfache Zyklen überzogener Erwartungen und folgender Ernüchterung erlebt. Der Sprung aus der Forschung in Wirtschaft und (Lebens-)Alltag gelang spätestens in den 1970er und 1980er Jahren durch die sogenannten Expertensysteme. In Deutschland setzte die verstärkte Forschung in den 1980er Jahren ein.

Neben dem Maschinellen Lernen hat das Forschungsgebiet der KI eine Vielzahl weiterer wichtiger Methoden hervorgebracht, beispielsweise Verfahren zur **Mustererkennung**, zur **Wissensrepräsentation, zur automatischen Inferenz und Handlungsplanung sowie zur Benutzermodellierung**. Diese Verfahren werden bspw. in Sprach-, Bild- und Dialogverstehen, in der Robotik und bei Multiagentensystemen eingesetzt.

3 Jörg Drechsler / Nicola Jentzsch: Synthetische Daten: Innovationspotential und gesellschaftliche Herausforderungen, Stiftung Neue Verantwortung, 2018 (abrufbar unter: *https://www.stiftung-nv.de/de/publikation/synthetische-daten-innovationspotential-und-gesellschaftliche-herausforderungen*).

4 John McCarthy / Marvin Minsky / Nathaniel Rochester / Claude Shannon: A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1955.

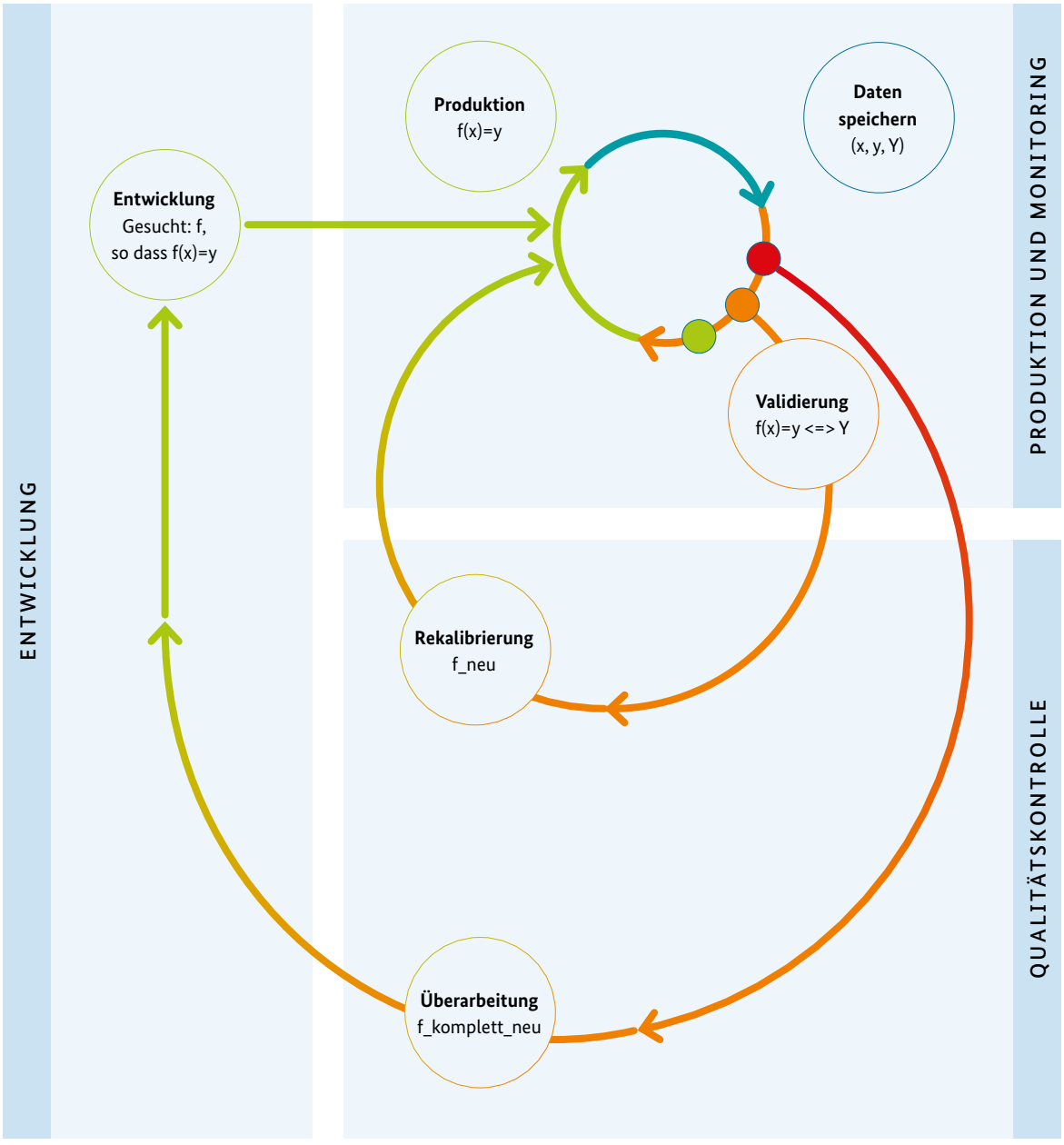
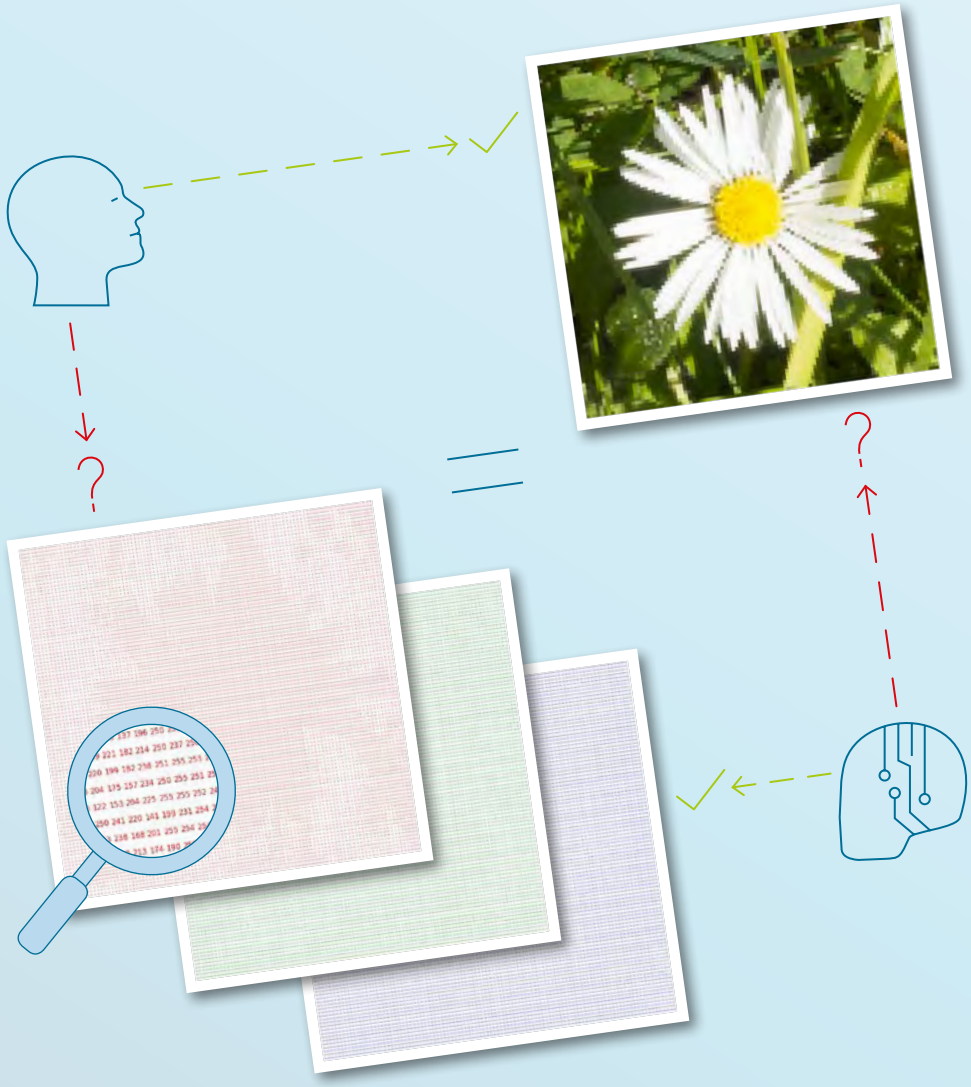


Abbildung 2: Prozessmodell eines auf Maschinellem Lernen basierenden Algorithmus: Fortlaufende Beobachtung und Bewertung. Der Prozess beginnt mit der Entwicklung des Algorithmus f auf den Trainingsdaten. Ist ein Algorithmus gefunden worden, der die angestrebten Qualitätsstandards erfüllt, wird dieser in Produktion gebracht. Um die Möglichkeit zum Monitoring und zur Qualitätskontrolle zu haben, muss in Produktion der Input x in den Algorithmus, der Output y des Algorithmus und der zugehörige korrekte Wert Y gesichert werden. Auf der Basis dieser Information kann eine Kontrolle des Algorithmus im produktiven Umfeld etabliert werden. Hierzu wird verglichen, in welchem Maß die Outputs y des Algorithmus den erwarteten Wert Y reflektieren. Sind die Abweichungen unkritisch, kann der Algorithmus ohne Veränderung weiterbetrieben werden. Bei signifikaten Abweichungen kann es erforderlich werden, die Parameter des Algorithmus neu zu schätzen, d.h. zu rekalibrieren. Bei kritischen Abweichungen empfiehlt sich eine Überarbeitung.

Verständlichkeit und Nachvollziehbarkeit als Problem

Ein intuitives Verstehen von mathematisch oder technisch abgebildeten Methoden ist für Menschen häufig schwierig oder unmöglich. Dies betrifft selbst Experten im Bereich der Modellierung. Sogar bei mathematisch gut verstandenen, verhältnismäßig einfachen Klassifikationsmethoden wie der logistischen Regression, weiß kaum jemand intuitiv, welches Ergebnis sie für welche Eingabewerte liefert.

Ein Beispiel sind Neuronale Netze (NN) in der Bilderkennung: Während ein Mensch auf einem Foto meist sofort erkennt, was abgebildet ist, ist dies kaum möglich bei dem Betrachten der Datenstrukturen desselben Fotos, die Eingabe für ein NN sind. Das bedeutet: Selbst wenn man alle digitalen Eingabewerte kennt und alle Schritte in einem NN nachvollziehen kann, folgt daraus nicht, dass Menschen den Erkennungsprozess verstehen und etwa im Fehlerfall feststellen können, warum eine Fehlerkennung passiert und wie diese behoben werden kann. Menschliche und maschinelle Objekt- und Mustererkennung funktioniert also nach unterschiedlichen Regeln, die nicht leicht übersetzbar sind.



2.2.5 Algorithmische Systeme

Ein algorithmisches System besteht in der Regel nicht aus einem einzigen Algorithmus, sondern aus einer Vielzahl von Algorithmen, die zusammenarbeiten können. Eine Komponente beschreibt einen ausführbaren Teil eines solchen Systems. Ein Algorithmus kann in unterschiedlichen Komponenten technisch verschieden umgesetzt sein, beispielsweise im Bereich der Microservice-Architekturen. Es ist zu berücksichtigen, dass die einzelnen Komponenten eines solchen Systems in Produktion unterschiedlichen Anforderungen in Bezug auf rechtliche Vorgaben oder Schutzzielen unterliegen können. Hinzu kommt, dass in einem algorithmischen System verschiedene Akteure, beispielsweise in Form von Zulieferern, Betreibern oder Herstellern, für verschiedene Komponenten des Systems verantwortlich sein können. Dabei ist zu beachten, dass an die einzelnen Komponenten unterschiedliche Anforderungen gestellt werden bzw. dass unterschiedliche Regularien greifen, z. B. für Datenqualität, Diskriminierungsfreiheit oder Vertragsfreiheit.

2.3 Software

Wird ein Algorithmus nicht in natürlicher Sprache, sondern in einer Programmiersprache (formale Sprache) formuliert, so wird er als **Programm** (oder **Software**) automatisch auf einem Rechner ausführbar. Die Funktionsweise einer Software ist nicht allein abhängig von den Daten, die sie verarbeitet, sondern auch vom Ausführungskontext (vgl. u.a. Technologie-Stack, der sämtliche Hard- und Softwarekomponenten beinhaltet, die für die Ausführung verwendet werden) und von ihrer Parametrisierung. Mit Hilfe von Parametern kann eine Software quasi von außen eingestellt werden. Auf diese Weise können einfache Informationen wie beispielsweise Darstellungsoptionen oder Pfadangaben bis hin zu komplexen Modellen an die Software übergeben werden. Je stärker eine Software parametrisierbar ist, desto flexibler ist sie einsetzbar, desto komplexer ist sie in der Regel zu entwickeln und desto relevanter werden die Parameter. So kann parametrisierbare Software relativ leicht an verschiedene Kontexte adaptiert werden, ohne erneut den Quelltext, d. h. die eigentliche Umsetzung, verändern zu müssen. Eine spezielle Variante sind adaptive Systeme, die sich über die Zeit automatisch an ihren Kontext, beispielsweise die Person des Nutzers oder die Einsatzumgebung, anpassen.

Um die Entwicklung qualitativ hochwertiger Software in einem gleichzeitig immer komplexer werdenden Umfeld effizient(er) zu gestalten und Kommunikationsprobleme im Entwicklungsprozess zu reduzieren, werden seit vielen Jahren erfolgreich **modellgetriebene Entwicklungsansätze** verfolgt. Hierbei wird eine generische Softwarekomponente mit einem komplexen Modell auf eine den Anwendungskontext bezogene Sprache parametrisiert. Einen Spezialfall stellen dabei mathematisch-statistische Modelle dar, die sich von domänenspezifischen Sprachen dadurch abgrenzen, dass hier nicht ein Modell explizit spezifiziert oder programmiert, sondern das mathematisch-statistische Modell (implizit) auf Basis von Daten angelernt bzw. trainiert wird (→ siehe oben 2.2.3 zum Maschinellen Lernen).

2.4 Hardware

Die Software wird von Hardware und speziell von sog. **Prozessoren** ausgeführt, deren Leistungsfähigkeit in der Vergangenheit stets zunahm, während die Geräte selbst kontinuierlich kleiner wurden, so dass sich das Feld der Einsatzszenarien stets erweiterte. Die Steigerung der Leistungsfähigkeit nach der sog. Moore'schen Gesetzmäßigkeit (Hundertfache Leistungssteigerung in 10 Jahren) unterliegt jedoch physikalischen Grenzen. Mit der Annährung von Chipkomponenten an die Größe einzelner Atome wird es zunehmend kostspieliger und technisch aufwändiger, die Vorhersagen von Moore mit Silizium als Transistormaterial zu erfüllen. Es wird daher heute mit alternativen Materialien, wie Graphen, in Verbindung mit neuen Berechnungskonzepten, wie photonischen Quantencomputern, geforscht, wobei eine Alltagstauglichkeit noch offen ist. Bereits etablierte Lösungen, die stark auf Parallelität setzen, sind dagegen Multi- und Many-Core-Prozessoren oder der Einsatz von Grafikprozessoren (GPU = Graphic Processing Unit). Für die Beschleunigung des Maschinellen Lernens über Massendaten wurden auch anwendungsspezifische Chips wie die Tensor-Prozessoren (TPU = Tensor Processing Units) entwickelt, die auf das hochparallele Addieren und Multiplizieren von Matrizen für neuronale Netze optimiert sind.

Durch die immer stärkere Parallelisierung der Berechnungen entsteht das Problem, dass für Menschen Fehler in solchen Prozessoren sehr schwer zu finden und auch die durchgeführten Berechnungen auf der Hardwareebene **kaum reproduzierbar und nachvollziehbar** sind.

2.5 Systemarchitektur

Heute laufen Anwendungen selten auf einem einzelnen Rechner; es handelt sich dagegen um viele Software-Komponenten auf verschiedenen Rechnern, die miteinander interagieren, um eine Aufgabe zu erfüllen. Aufgrund der Verteilung auf unterschiedliche Hardware-Knoten spricht man von einem **verteilten System**. Ein verteiltes System setzt sich aus unterschiedlichen Software- und Hardware-Komponenten zusammen, die in einem Netz interagieren. Die Netzknoten kommunizieren miteinander über Funk oder Kabelverbindungen.

Für die Netzkommunikation existieren vielfältige **Protokolle und Standards**. Mittels dieser werden Daten auf den Netzknoten verarbeitet und über das Netz weitergeleitet bzw. zu anderen Knoten transportiert. Die Spezifikation für Anfragen, die an einen Server gestellt werden dürfen, wird beispielsweise in einer sog. Programmierschnittstelle (API = Application Programming Interface) veröffentlicht, wobei der Zugriff über diese Schnittstelle in der Regel gegen fehlerhafte Nutzung oder Angriffe abgesichert werden muss.

IT-Infrastrukturen, die über das Internet erreichbar sind, werden als **Cloud** bezeichnet. Cloud-Anwendungen können Milliarden von Nutzern erreichen. Bestimmte verwandte Cloud-Anwendungen werden oft als **Digitale Plattform** bezeichnet und haben einen hohen Bekanntheitsgrad wie beispielsweise die sog. Big Four oder GAFA (Google, Apple, Facebook, Amazon) bzw. GAFAM (wenn Microsoft mit dazu genommen wird).

Während zu Beginn des Internets der Dinge die meisten Daten direkt in die Cloud geschickt wurden, um dort auf großen digitalen Plattformen verarbeitet zu werden, werden derzeit vermehrt Lösungen entwickelt, bei denen die Daten direkt und möglichst nah an dem Erhebungspunkt, also gleichsam „am Rande“ (on the edge) des Internets, verarbeitet oder zumindest vorverarbeitet werden. Die Verarbeitung nah am Erhebungspunkt wird im Gegensatz zur Verarbeitung in der Cloud (Cloud-Computing) als **Edge-Computing** bezeichnet. Gerade die Vorverarbeitung von Daten ermöglicht es, die Kommunikationsaufwände zu minimieren, aber auch datenschutzfreundlichere Systeme zu erstellen, indem bereits an dieser Stelle, nämlich nahe des Erhebungspunktes, ein nicht-erforderlicher Personenbezug entfernt werden kann.

Die mittlerweile entstandene komplexe Systemlandschaft einschließlich Internet, Edge-Computing und IoT führt dazu, dass Einzelsysteme wegen einer starken Verschränkung nur schwer voneinander abgrenzbar sind.

Die Ausgestaltung der Architektur verteilter Systeme hat durch die Entscheidung, welche Technologie eingesetzt wird, auf welchen Netzwerkknoten die Software läuft, mit wem über welche Schnittstellen und Protokolle kommuniziert wird, auch signifikanten **Einfluss auf die Geschäftsprozesse**, die das System unterstützt. Wenn beispielsweise Hersteller von Hardware Daten, die ihre Geräte erfassen, nutzen wollen, um diese langfristig zu verbessern, so können sie eine eigene Kommunikationsinfrastruktur aufbauen, gegebenenfalls die Infrastruktur des Anwenders nutzen oder aber den Anwender bitten, die Daten über eine Schnittstelle zur Verfügung zu stellen. Der Umgang mit solchen Daten in kooperativen Prozessen sollte transparent gestaltet werden und muss gegebenenfalls vertraglich geregelt werden. Die vertragliche Gestaltung des Datenaustauschs kann dabei durch die technischen Gegebenheiten einschränkt sein.

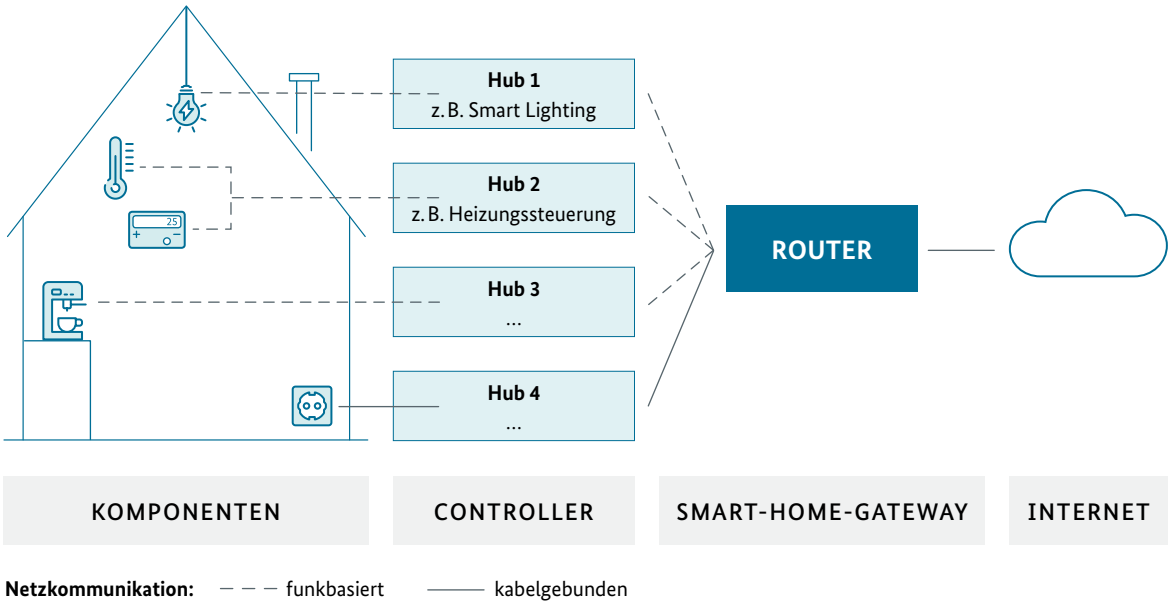


Abbildung 3: Beispiel der Systemarchitektur im Smart Home

Blockchain und andere Distributed Ledger Technologien

Die signifikanten Verbesserungen im Bereich der verteilten Systeme ermöglichen den Einsatz von **Distributed Ledger Technologien** (DLT), die auf dem Konzept verteilter Kassenbücher basieren. Statt einer zentralen Verwaltung werden dabei viele grundsätzlich gleichgestellte Kopien eines sog. Kassenbuchs von unterschiedlichen Partnern verwaltet. Neue Einträge werden in allen Kopien ergänzt und der aktuelle Stand durch einen Konsensus (eine Übereinkunft) geklärt. Die zugrundeliegende Architektur solcher Systeme variiert je nach Verwendungszweck und Transaktionsgestaltung von linearen Ansätzen zu verschiedensten graphbasierten Systemen. Auch eine Übereinkunft kann auf verschiedene Arten erzielt werden, die durch ein sog. Konsensusprotokoll festgelegt wird.

Einer der bekanntesten Vertreter einer DLT-Architektur ist die **Blockchain**, zu der es Implementierungen wie Bitcoin oder Ethereum gibt. In einer Blockchain werden Daten in einer Liste von Datensätzen („Blöcken“) gespeichert. Die Blöcke sind miteinander kryptographisch verknüpft, so dass eine Transaktion, die als Block gespeichert wird, implizit die Richtigkeit früherer Transaktionen, d. h. der gesamten Kette (Chain), bestätigt und somit Manipulationen wie Veränderungen oder Löschungen von Einträgen erschwert. Durch das dezentrale Konsensusprotokoll ist keine zusätzliche Instanz zur Integritätsbestätigung von Transaktion erforderlich.

Teil D

Mehr-Ebenen-Governance komplexer Datenökosysteme

Die konkrete Umsetzung des von der DEK zugrunde gelegten ethischen und rechtlichen Ordnungsrahmens stellt die Regulierung, die Steuerung und das Design von Datenökosystemen in Anbetracht der hohen Komplexität und Dynamik dieser Systeme vor neue Herausforderungen und erfordert das Zusammenwirken verschiedener Akteure und unterschiedlicher Governance-Instrumente auf mehreren Regulierungsebenen (Mehr-Ebenen-Governance). Der folgende Teil zeigt zunächst **relevante Governance-Instrumente und Akteure** auf. Weitere Spezifizierungen, auch zum Zusammenspiel verschiedener Instrumente und Akteure, finden sich in den beiden anschließenden Kapiteln zu Daten und algorithmischen Systemen.

1. Allgemeine Rolle des Staates

Diejenigen, die ethisch begründete Rechte wahrnehmen und korrespondierende Pflichten befolgen müssen – seien es etwa Bürger, Unternehmen oder staatliche Stellen – müssen dazu auch in der Lage sein. Hieraus ergeben sich zahlreiche Aufgaben für den Staat. Zunächst ist der Staat verantwortlich für die **rechtlichen Rahmenbedingungen,** in denen sich eine gemeinwohlorientierte Datengesellschaft entwickeln kann. Die Geschwindigkeit, in der sich algorithmische Systeme fortentwickeln und mit der sie in immer mehr Lebensbereiche vordringen, führt zu großen Herausforderungen für die Gesetzgebung und die konkretisierende Gerichtspraxis. Der Staat hat sicherzustellen, dass Regulierung in einem solchen Umfeld einerseits hinreichend Steuerungskraft entfaltet und andererseits die nötige Flexibilität aufweist, um ihre Aufgabe auch unter geänderten technologischen Bedingungen erfüllen zu können. Dafür bedarf es **technikneutraler Formulierung** von Rechtsnormen und **innovativer Regulierungsmodelle**.

Zudem bedarf es **angemessener infrastruktureller und technischer Voraussetzungen**, etwa befähigender Technologien, Institutionen und Intermediäre, unter Einschluss eines breiten Spektrums zivilgesellschaftlicher Akteure. Auch insofern kommt dem Staat nach Auffassung der DEK eine entscheidende Garantie- und Gewährleistungsfunktion im Rahmen der Daseinsvorsorge zu.

Durch die neuen Möglichkeiten der Datengesellschaft entsteht zudem eine umfassende **Bildungsaufgabe**. Hier stellt sich die Frage, welche Kompetenzen für den kreativen sowie gleichzeitig reflektierten Umgang mit digitalen Technologien notwendig sind und welche Rahmenbedingungen geschaffen werden sollten, um adäquate Bildungsangebote für vielfältige Zielgruppen umsetzen zu können. Die Bildungsaufgabe des Staates ist in einem umfassenden Sinn zu verstehen und schließt eine entsprechende Bewusstseinsbildung durch **Öffentlichkeitsarbeit** mit ein.

Ferner kommt dem Staat allgemein die Aufgabe zu, **Forschung und Entwicklung** zu fördern. Von besonderer Wichtigkeit ist hier, Forschung und Entwicklung ethisch fundierter Technologien bspw. zu Nachvollziehbarkeit, zu Transparenz oder zum Schutz vor Diskriminierung zu unterstützen. Die Berücksichtigung von ethischen und rechtlichen Grundsätzen und Prinzipien erfordert intensive Forschungs- und Entwicklungsarbeit, die verstärkt gefördert werden sollte.

Der Staat muss zwar nicht alle Mittel selbst oder durch staatsnahe Institutionen bereitstellen, aber er muss die **rechtlichen und sonstigen Rahmenbedingungen** für eine Datengesellschaft schaffen, in welcher Einzelne sich ebenso wie Unternehmen auf der Grundlage ethischer Werte und Prinzipien selbstbestimmt und zugleich ausreichend geschützt bewegen können und in welcher Potenziale von Daten und algorithmischen Systemen für die Gestaltung einer lebenswerten Zukunft genutzt werden.

Im Sinne einer ethisch fundierten Governance auf mehreren Ebenen sollte sich Deutschland auch kraftvoll in einen **europäischen und internationalen Diskurs** einbringen. Die globale Dimension der technologischen Entwicklung kann nicht von einem einzigen Nationalstaat und allein durch nationale Regulierung angemessen adressiert werden. Deswegen begrüßt die DEK die bereits bestehenden europäischen und internationalen Initiativen (z. B. der Europäischen Kommission und der OECD) für eine ethisch fundierte Gestaltung unserer Zukunft. Der Gewährleistung der **digitalen Souveränität** Deutschlands und Europas im internationalen Kontext (näher Teil G) kommt hierbei eine herausragende Bedeutung zu.

2. Unternehmerische Selbstverpflichtungen und Corporate Digital Responsibility

Sich um die Risiken der Digitalisierung zu sorgen und zu kümmern, aber auch deren erhebliche Potenziale wahrzunehmen, ist nicht nur eine Frage staatlicher Verantwortung und Regulierung. Ebenso tragen jene Akteure, die Technologien entwickeln, verbreiten und einsetzen, eine solche **Verantwortung**, und dies auch **jenseits gesetzlicher Vorgaben**. Auch wenn der Staat, nicht zuletzt infolge seiner Schutzpflichten zur Gewährleistung der Vertraulichkeit und Integrität informationstechnischer Systeme sowie weiterer Grundrechte, eine hervorgehobene Verantwortung hat, sind Instrumente der Selbstregulierung gerade im Kontext der Digitalen Transformation unverzichtbar.

Die Wahrnehmung einer jeweils eigenen Verantwortung für die Folgen der Digitalisierung wird, bezogen auf Unternehmen als Hersteller und Betreiber digitaler Technologien, unter dem Begriff der **Corporate Digital Responsibility (CDR)** diskutiert und praktiziert. CDR wird – in Anlehnung an Corporate Social Responsibility (CSR) – als Teilbereich der Unternehmensverantwortung verstanden, hier bezogen auf freiwillige unternehmerische Aktivitäten im digitalen Bereich, die über das heute gesetzlich Vorgeschriebene hinausgehen und die digitale Welt aktiv zum Vorteil der Gesellschaft im Allgemeinen und der Kunden und Mitarbeiter im Besonderen mitgestalten. In diesem Sinne hat das Bundesministerium der Justiz und für Verbraucherschutz (BMJV) im Mai 2018 eine Initiative zur Etablierung von Grundsätzen und Konzepten einer unternehmerischen digitalen Verantwortung gestartet (*www.bmjv.de/cdr*). CDR kann demnach viele Themenbereiche umfassen[1], unter anderem den Schutz personenbezogener Daten, die Sicherung der Inklusion in der digitalen Sphäre, die Transparenz etwa von Algorithmen oder im Datenschutz, die Entwicklung digitaler Innovationen, die zur Erreichung von Nachhaltigkeitszielen beitragen, den gemeinwohlorientierten Einsatz von Algorithmen, Open Data und die Gewährleistung von Informationssicherheit.

Eine verantwortungsvolle Entwicklung digitaler Produkte und Dienstleistungen muss bei allen unternehmerischen Entscheidungen und auf allen Ebenen des Unternehmens eine zentrale Rolle einnehmen. Ethische Fragestellungen dürfen dabei nicht allein den Rechtsabteilungen und Compliance-Beauftragten zugewiesen werden. Sie ist vielmehr als **Querschnittsaufgabe in sämtliche Prozesse zu integrieren**. Alle Beteiligten müssen sich verantwortlich fühlen, ethische Werte wie Teilhabe, Fairness, Gleichbehandlung, Selbstbestimmung und Transparenz zu berücksichtigen. So sollen die negativen sozialen und gesellschaftlichen Effekte der Digitalisierung und digitaler Geschäftsmodelle auf Mitarbeiter, Lieferanten, Kunden sowie die Gesellschaft und Umwelt insgesamt minimiert und die neuen Möglichkeiten der Digitalisierung zur Verwirklichung gesamtgesellschaftlicher Ziele genutzt werden. Richtig eingesetzt kann CDR zu Verbraucherschutz, digitaler Teilhabe und einer **nachhaltigen Entwicklung der Digitalwirtschaft** beitragen.

Im Kern ist CDR, genauso wie Corporate Social Responsibility (CSR), eine Selbstverpflichtung auf freiwilliger Basis. Dementsprechend kann die Umsetzung insbesondere durch interne Strategien wie unternehmensinterne oder branchenspezifische Wertekodizes abgesichert werden. Die DEK begrüßt insoweit die vermehrte Ausbildung professionsethischer Standards und Verhaltenskodizes (Codes of Conduct) durch Verbände und Unternehmen der datenverarbeitenden Wirtschaft, soweit diese zur Konkretisierung der Vorgaben beitragen. Maßnahmen im Rahmen von CDR dürfen nicht bloß ein „Feigenblatt" sein, um dem Unternehmen einen „Anstrich digitaler Ethik" zu geben.

1 Corporate Digital Responsibility-Initiative: Digitalisierung verantwortungsvoll gestalten – Eine gemeinsame Plattform, 2018 (abrufbar unter: *https://www.bmjv.de/SharedDocs/Downloads/DE/News/Artikel/100818_CDR-Initiative.pdf?__blob=publicationFile&v=4*).

Aus Sicht der DEK sollte schon in der Entwicklungsphase eines digitalen Produkts nicht nur gegebenenfalls eine Datenschutz-Folgenabschätzung nach der DSGVO, sondern im Sinne der Übernahme vorausschauender Verantwortung eine darüber hinausgehende, allgemeine **Risiko-Folgenabschätzung** für die Gesellschaft (einschließlich der von der Digitalen Transformation besonders betroffenen Mitarbeiter und Kunden eines Unternehmens) durchgeführt werden, die auch die gesellschaftlichen Langzeitwirkungen datengetriebener Geschäftsmodelle berücksichtigt. Dabei könnte für marktmächtige Unternehmen empfohlen werden, nach dem Vorbild von Verbraucher- oder Kundenbeiräten einen Beirat mit den Vertretern der je nach Geschäftsmodell spezifisch betroffenen Personengruppen einzurichten, welcher an einer solchen Folgenabschätzung beteiligt werden könnte.

3. Bildung: Stärkung digitaler Kompetenzen und kritischer Reflexion

Digitale Selbstbestimmung setzt digitale Kompetenz voraus. In diesem Zusammenhang sind die Bemühungen der Bundesregierung, von Verbraucherschutzverbänden, von juristischen Berufsvereinigungen sowie seitens anderer Stellen um eine **Sensibilisierung der Bevölkerung für den selbstbestimmten Umgang mit Daten und digitalen Technologien** – von Konfigurationsmöglichkeiten auf dem eigenen Smartphone bis hin zur digitalen Nachlassplanung – und um einfache und verständliche Informationen über die Gestaltungsmöglichkeiten nebst praktischen Hilfestellungen uneingeschränkt zu begrüßen. Das betrifft auch Bemühungen, bei Verbrauchern ein Bewusstsein für das Potenzial der Daten zu wecken und sie verstärkt über ihre Rechte und über die tatsächlichen Chancen und Risiken, ihre Daten wirtschaftlich zu nutzen, aufzuklären. Die DEK empfiehlt, all diese Bemühungen aufrecht zu erhalten und noch zu intensiveren.

Auch in den Schulen sollte möglichst früh ein Bewusstsein für die Digitalisierung geschaffen werden. Digitale Kompetenz sollte in die **Lehrpläne** integriert werden und Lehrkräfte müssen regelmäßig und umfassend geschult werden. Nur so können neue Generationen zu kompetenten „Digital Natives“ heranwachsen, die sowohl Chancen als auch Risiken neuer digitaler Anwendungen einschätzen, informierte Entscheidungen treffen und ihre Rechte effektiv einfordern können.

Daneben bedarf es einer **lebenslangen Bildung** zum Umgang mit Daten und digitalen Technologien, die für alle Altersstufen und gesellschaftlichen Gruppen gewährleistet werden muss. Hierbei ist zu berücksichtigen, dass digitale Kompetenz nicht nur grundlegende Kenntnisse technischer Aspekte voraussetzt, für die fortlaufend technisch-mathematische Kompetenzen vermittelt werden müssen, sondern auch ausreichende Kenntnisse ökonomischer, rechtlicher, ethischer und sozialwissenschaftlicher Art. Die Vielfalt an Kenntnissen ist erforderlich, um unterschiedliche Chancen und Risiken in ihrer Komplexität erfassen, diskutieren und bewerten zu können.

Von besonderer Relevanz ist hier die Ausbildung in Informatik, Softwareentwicklung und Datenwissenschaft (Data Science). Hier bedarf es einerseits grundlegender Lehrveranstaltungen zu ethischen und rechtlichen Fragen sowie andererseits weiterführende Ausbildung zu Statistik, Methodologie und Wissenschaftstheorie. Insbesondere die Verankerung daten- und forschungsethischer Fragestellungen in der fachspezifischen Methodenausbildung ist hier von zentraler Bedeutung und sollte deutlich vorangetrieben werden, damit diejenigen, die digitale Produkte und Dienstleistungen entwickeln oder über ihre Entwicklung entscheiden, ethische und rechtliche Gesichtspunkte frühzeitig in ihre Überlegungen mit einbeziehen.

Um diese Ziele zu verwirklichen, bedarf es zunächst des Zusammenwirkens einer **Vielzahl staatlicher, staatsnaher und privater Akteure** auf Bundes- wie auf Landesebene und in den Kommunen. Aufbau und langfristige Sicherstellung digitaler Kompetenz der Bevölkerung ist eine zu große Aufgabe, und die Herausforderungen in einzelnen Lebenszusammenhängen sind zu vielfältig, um dies zentralisiert in die Hände einer einzigen Stelle zu legen. Jedenfalls dürfte den Aufsichtsbehörden (Datenschutzbehörden und/oder jeweilige Fachaufsichtsbehörden), der Stiftung Datenschutz und den Verbraucherzentralen sowie den für die Bildung zuständigen Stellen eine zentrale Rolle zukommen. Auch den Medien und den Institutionen der Medienregulierung kommt in diesem Zusammenhang eine wichtige Funktion zu. Diese besteht nicht nur in der Aufklärung der Gesellschaft über neue Technologien und in der kritischen Begleitung des technischen Fortschritts, sondern auch in der Bereitstellung neuer Foren für Debatten.

Digitale Kompetenz der Bevölkerung lässt sich allerdings trotz der primären Verantwortlichkeit staatlicher Stellen nicht umfassend verwirklichen ohne den Aufbau entsprechender **zivilgesellschaftlicher Strukturen**, wie des digitalen Ehrenamts, des sog. Tech-Accountability-Journalismus und der verbraucherorientierten Marktbeobachtung. Die DEK empfiehlt daher der Bundesregierung, den Aufbau derartiger Strukturen nachhaltig zu fördern.

Auch innerhalb von **Unternehmen** ergeben sich Bildungsaufgaben. So kann ein Unternehmen nur dann hohen ethischen Standards genügen, wenn diejenigen, die im Unternehmen tätig sind, insbesondere im Management und in der Produktentwicklung, eine hinreichende Sensibilität für ethische und rechtliche Fragen aufweisen. Im Bereich der Aus- und Fortbildung sollten Fragestellungen rund um Datenethik und Datenrecht ferner bei einer **breiten Palette akademischer und beruflicher Ausbildungswege** sowie in der betrieblichen Fortbildung berücksichtigt werden. Dabei ist insbesondere an technische und betriebswirtschaftliche Berufsrichtungen zu denken, damit diejenigen, die digitale Produkte und Dienstleistungen entwickeln oder über ihre Entwicklung entscheiden, ethische und rechtliche Gesichtspunkte frühzeitig in ihre Überlegungen mit einbeziehen.

4. Technologieentwicklung und ethisch fundiertes Design

Die Bemühungen, digitale Kompetenzen in der Bevölkerung zu verbessern, dürfen keine Verschiebung von Verantwortung weg von Produzenten und digitalen Dienstleistern hin zu den Nutzern bedeuten, zumal die Nutzer nur begrenzte Möglichkeiten haben, alle Verarbeitungsschritte ihrer Daten und die dahinterliegenden Geschäftsmodelle nachzuvollziehen und zu verstehen. Die Übernahme von Verantwortung ist vielmehr zuvörderst auf der Seite derjenigen erforderlich, welche Einfluss auf die Entwicklung von Produkten und Dienstleistungen haben. Solche Verantwortung äußert sich insbesondere in ethisch fundiertem Design (sog. **Ethics by Design** bzw**. Ethics in Design**) und ist bspw. in Bezug auf Privatsphäre- und Datenschutz bereits in der DSGVO unter den Stichworten Datenschutz „by design" und Datenschutz „by default" enthalten. Die Orientierung der Entwicklung von Technologien und Produkten (einschließlich Diensten und Anwendungen) an den zuvor dargelegten ethischen Werten und Prinzipien ist zudem geeignet, Vertrauen und Akzeptanz der Bevölkerung in digitale Produkte zu stärken.

Dabei muss allerdings jedes Produktdesign **auf die adressierten Nutzergruppen abgestimmt** sein, wobei **partizipative Produktentwicklung**, die Nutzergruppen und ihre Bedürfnisse bereits im Stadium der Produktentwicklung mit einbezieht, hilfreich sein kann. Insbesondere dort, wo ein Produkt auch wenig digital affine und/oder vulnerable Nutzergruppen adressiert, sollte Design, einschließlich der datenschutzfreundlichen Voreinstellungen, **inklusiv gestaltet** sein, sodass auch diese Nutzergruppen in ihrer digitalen Selbstbestimmung geschützt sind. Damit können Hersteller und Betreiber der besonderen grundrechtlichen Verankerung der informationellen Selbstbestimmung in Art. 1 Abs. 1 GG (Menschenwürde) gerecht werden, die es verbietet, den Schutz von den individuellen Fähigkeiten und der individuellen Lebenssituation des Einzelnen abhängig zu machen.

Typische Technikentwicklungsmethoden und -plattformen, weitverbreitete Bibliotheken oder andere Code-Komponenten unterstützen bisher kaum die Anforderungen von Ethics by Design. Gleichzeitig führen Komponenten, die mit einer aus ethischer oder datenschutzrechtlicher Sicht besseren Gestaltung aufwarten, allenfalls ein Nischendasein. In diesen Bereichen sind Änderungen nötig, damit der Einbau ethischer Prinzipien im Allgemeinen und Datenschutzprinzipien im Speziellen die Regel wird, statt weiterhin eine Ausnahmeeigenschaft darzustellen. Ethics by Design erfordert einen Brückenschlag zwischen verschiedenen Gemeinschaften („Communities") und hat Auswirkungen auf die betroffenen Berufsbilder. Hilfreich für die Umsetzung wären neben Informationen zu Methoden und Katalogen **Best-Practice-Konzepte, unterstützende Werkzeuge, Entwicklungs-Frameworks** und **(Open-Source-) Code-Komponenten**. Über Plattformen mit Repositorien für solche Komponenten sowie verwendbare Datenbestände, die gegebenenfalls Überprüfungen erst möglich machen, könnten die besonderen Eigenschaften herausgestellt, nötige Dokumentationen gleich mitgeliefert und Möglichkeiten zum Austausch von Erfahrungswissen bereitgestellt werden.

Auch wenn Ethics by Design ein wichtiges Governance-Instrument ist, um Produkte, Prozesse und Dienstleistungen von Beginn an im Interesse des Individuums und des Gemeinwohls zu gestalten, ist es kein Garant für ethische Produkte und Dienstleistungen. Ethische Prinzipien können und sollen die Technologieentwicklung positiv beeinflussen, **Ethik lässt sich aber nicht an Technik delegieren**. Zudem sollten Entscheidungen darüber, welche ethischen Prinzipien wie umgesetzt werden, z. B. ob, und wenn ja, welche Fairnessmaße für algorithmische Systeme verwendet werden, nicht Entwicklern alleine überlassen werden, sondern kontextspezifisch und ggf. unter Einbeziehung Betroffener ausgehandelt werden.

5. Forschung

Während in der Forschung häufig Lösungsansätze für eine ethisch besser fundierte Gestaltung von datenverarbeitenden Systemen entwickelt und exemplarisch umgesetzt werden, besteht zwischen Wissenschaft und Praxis eine gewisse Kluft. Dies könnte darauf zurückzuführen sein, dass einige der technischen Lösungen, bspw. auf Basis von kryptographischen Mechanismen, kontraintuitive Eigenschaften aufweisen, die im Vergleich zu herkömmlichen Methoden für viele Menschen schwer verständlich sind (wie ein Ausweis, der bei jedem Vorzeigen anders aussieht und damit die Verknüpfbarkeit von beobachteten Aktionen verhindert). Die vorhandenen **mentalen Modelle**, die viele Menschen aus der (analogen) Welt haben, reichen nicht aus, um ein **Verständnis** für solche innovativen Technologien zu erzeugen oder den Mehrwert zu vermitteln. Solange aber Schwierigkeiten im Verständnis oder in der Verwendung bestehen, ist die Verbreitung solcher Technologien schwierig, selbst wenn sie Vorteile bezüglich ihrer ethischen oder datenschutzrechtlichen Eigenschaften mit sich bringen.

Vielfach erfordern das Verstehen von Implikationen neuer Entwicklungen und das ethisch fundierte Gestalten **eine übergreifende und damit auch interdisziplinäre Zusammenarbeit**, welche von den disziplinären Metriken für gute Wissenschaft und Forschung nicht erfasst wird. Hier ist ein Umdenken in unterschiedlichen Bereichen (z. B. Hochschulen, Publikationsbewertung, Gutachterwesen) erforderlich, damit interdisziplinäre Forschung eine angemessene Würdigung erhält. Forschungsförderungen sollten die interdisziplinäre Zusammenarbeit, die zu Ergebnissen führt, die in den Einzeldisziplinen gar nicht hätten erreicht werden können, besonders honorieren und langfristige Karrierepfade sowie geeignete instutionelle Rahmenbedingungen vorsehen.

Im Forschungsbereich sind bereits vielfach gute und vielversprechende technische Lösungen vorhanden, die jedoch noch zu wenig nachgefragt werden. Außerdem fehlt es an Methodiken oder Technologien, die es ermöglichen, vom jetzigen Realisierungsstand einen **Migrationspfad zu einem verbesserten Status** der Technologie zu erreichen. Auch dieser Aspekt verdient eine besondere **Entwicklungs- und Innovationsförderung,** um tatsächlich bessere Lösungen in die Realität zu bringen. Statt lediglich punktuell Spitzenleistungen zu fördern, muss auch ein Fortschritt in der Breite zum ethisch fundierten Design Anerkennung finden.

6. Standardisierung

Spätestens als vor 20 Jahren Lawrence Lessig „Code is Law“[2] postulierte und damit die Relevanz der technischen Realität heraushob, sollte klar geworden sein, dass technische Standardisierung essentiell für die Umsetzung rechtlicher und ethischer Vorgaben ist. Für **technische Standardisierung** im Bereich von Kommunikationsnetzen sind bspw. die weltweit aktiven Gremien ISO/IEC, IEEE, IETF, ITU, ETSI oder W3C zuständig, für Europa ferner CEN und in Deutschland neben weiteren spezifischen Standards für öffentliche Stellen v.a. DIN. Zwar hat ein technischer Standard allein keine Gesetzeskraft, und Anwender technischer Systeme müssen auch dann das geltende Recht einhalten, wenn dieses den Anforderungen eines globalen technischen Standards widerspricht. Dennoch beeinflusst die Standardisierung das Angebot auf dem Markt massiv, sodass möglichst vermieden werden muss, dass sich Standards etablieren, die gegen geltendes Recht verstoßen.

Der Prozess der Standardisierung steht häufig in der Kritik, weil ihm die demokratische Legitimation fehlt und faktisch **keine repräsentative Mitwirkung** der betroffenen Teile der Gesellschaft eröffnet ist. So sind Nichtregierungsorganisationen oder andere Vertreter der Zivilgesellschaft selten an der Standardisierung beteiligt. Auch Datenschutzbehörden können in der Regel nur in Einzelfällen an der Standardisierung technischer Systeme mitwirken. Dies kann im schlechtesten Fall dazu führen, dass der Betrieb von standardkonformen technischen Systemen nicht gleichzeitig rechtskonform wäre. Kritisiert wird auch, dass einige internationale Standards, an die sich Hersteller oder Betreiber halten sollen, **nicht öffentlich und kostenlos** zur Verfügung stehen, sondern erst erworben werden müssen.

Standardisierung in der Informationssicherheit hat in der Vergangenheit großteils dazu beigetragen, verstärkt Sicherheitsfunktionalität einzubauen und allmählich das Sicherheitsniveau zu erhöhen, bspw. beim Online-Banking. Allerdings haben die Snowden-Enthüllungen ans Tageslicht befördert, dass einige Geheimdienste und Regierungsbehörden gezielt Sicherheitslücken oder Hintertüren in Standards einzubringen versuchen, um sich zukünftige Zugriffsmöglichkeiten zu verschaffen. Es ist zu erwarten, dass die technische Standardisierung künftig einen größeren Stellenwert einnimmt, bspw. durch die Anforderung der DSGVO, den Stand der Technik zu berücksichtigen, oder als Konsequenz des IT-Sicherheitsgesetzes. Ebenso ist zu erwarten, dass die politische Einflussnahme aus vielen, auch außereuropäischen, Ländern zunehmen wird.

Eine **Folgenabschätzung** bezüglich existierender oder diskutierter Standards muss über rein technische und ökonomische Perspektiven hinausgehen und um ethische und gesellschaftliche Aspekte **erweitert** werden. Beim Standardisierungsprozess sollte der Staat Sorge dafür tragen, dass sich Akteure der Zivilgesellschaft, Datenschutzbehörden, Verbraucherschützer oder Vertreter von Betroffenenorganisationen ebenso in die Standardisierung einbringen können wie die bisher primär vertretenen Stakeholder.

2 Lawrence Lessig: Code and other Laws of Cyberspace, 1999.

7. Zwei Governance-Perspektiven: Daten- und Algorithmen-Perspektive

In den folgenden beiden Kapiteln werden die zuvor ausgeführten Überlegungen mittels zweier komplementärer Perspektiven auf datenbasierte, algorithmische Systeme angewendet. Die von der DEK zugrunde gelegten **allgemeinen ethischen Grundsätze und Prinzipien** (oben Teil B) müssen zum einen handlungsleitend sein für den Umgang mit Daten, insbesondere für die ethisch fundierte Gestaltung der Sammlung von Daten, des Zugangs zu Daten und der Datennutzung. Zum anderen müssen sie handlungsleitend sein für die Gestaltung datenverarbeitender, auf Algorithmen beruhender Systeme, einschließlich der vielfach so bezeichneten „Künstlichen Intelligenz". Bei der primär datenfokussierten Perspektive („Daten-Perspektive") und der primär auf algorithmische Systeme fokussierten Perspektive („Algorithmen-Perspektive") handelt es sich dabei weder um miteinander konkurrierende Sichtweisen noch um verschiedene Seiten ein und derselben Medaille, sondern um **sich wechselseitig ergänzende und bedingende ethische Diskurse**, welche sich typischerweise auch in unterschiedlichen Governance-Instrumenten, einschließlich unterschiedlichen Rechtsakten, widerspiegeln.

Die **Daten-Perspektive** richtet den Blick auf die Daten, welche zum Training algorithmischer Systeme, als Datenbasis für algorithmisch geprägte Entscheidungen oder auch für eine Fülle weiterer Zwecke verwendet werden, die in spezifischer Weise mit **Bedeutungskontext und Semantik von Daten** (→ Teil C, 2.1) verbunden sind. Sie betrachtet die Daten vor allem in Bezug auf deren Herkunft sowie auf die möglichen Konsequenzen der Datenverarbeitung für bestimmte Personen, welche mit Kontext und Semantik der Daten zu tun haben. Aus ethischer wie aus rechtlicher Perspektive geht es einerseits um objektive Anforderungen an den Umgang mit Daten, noch mehr aber typischerweise um **subjektive Rechte**, welche diese Personen gegenüber einer bestimmten anderen Person oder auch gegenüber jedermann geltend machen können. Eine zentrale Unterscheidung ist diejenige zwischen personenbezogenen und nicht-personenbezogenen Daten, welche über die Anwendbarkeit der datenschutzrechtlichen Betroffenenrechte entscheidet. Aktuelle Debatten, die hier zu verorten wären, sind etwa diejenigen um ein „Dateneigentum" oder um Open Data.

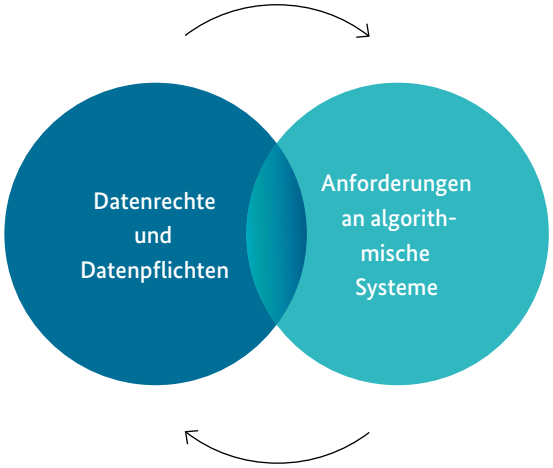


Abbildung 4:
Daten- und Algorithmenperspektive

Die **Algorithmen-Perspektive** richtet den Blick dagegen auf die Architektur und Dynamik des datenverarbeitenden algorithmischen Systems, seine Auswirkungen auf Einzelne und die Gesellschaft. Der ethische und rechtliche Diskurs fokussiert dabei typischerweise auf die Beziehung von **Mensch und Maschine** und mit Blick auf Künstliche Intelligenz, insbesondere auf die Automatisierung sowie auf die Verlagerung auch komplexer Handlungs- und Entscheidungsprozesse auf sog. autonome Systeme. In Abgrenzung zur Daten-Perspektive müssen die vom System betroffenen Personen nicht notwendig auch etwas mit den Daten zu tun haben, die das System verarbeitet, und wenn sie doch etwas mit diesen Daten zu tun haben, liegt dort bei der System-Perspektive nicht der Schwerpunkt der Betrachtung. Im Kern geht es um **objektive Anforderungen**, deren Beachtung möglicherweise eingefordert und an deren Missachtung Haftung und Sanktionen geknüpft werden können. Eine zentrale aktuelle Debatte, die hier zu verorten wäre, ist diejenige um eine sog. Algorithmenkontrolle.

Teil E

Daten

Daten bedeuten Zugang zu Information, Information kann zu Wissen führen, und Wissen verleiht Einfluss und Macht. Durch neue Möglichkeiten automatisierter Datenverarbeitung und den exponentiellen Anstieg von Speicherkapazitäten und Rechenleistung ist der mit dem Zugang zu Daten verbundene Zuwachs an Macht und Handlungsmöglichkeiten enorm. Dabei bringt die Verfügungsmacht über wichtige Ressourcen an sich schon ein besonderes Maß an Verantwortung mit sich. So dürfen Daten – wie andere Ressourcen auch – nur zu rechtmäßigen und ethisch vertretbaren Zwecken eingesetzt werden, und bei ihrer Nutzung sind – wie bei anderen Ressourcen auch – stets die Auswirkungen mit zu bedenken, welche die Nutzung für Einzelne oder die Allgemeinheit mit sich bringen kann. Daten weisen aber auch bestimmte Charakteristika auf, die sie von anderen Ressourcen unterscheiden.

Auf Grundlage der spezifischen Charakteristika von Daten konkretisiert die DEK daher zunächst – ohne Anspruch auf Vollständigkeit – die in Teil B genannten Grundsätze und Prinzipien zu allgemeinen Anforderungen an den Umgang mit Daten (→ unten 1.) sowie zu Datenrechten und korrespondierenden Datenpflichten (→ unten 2.). Sie entwickelt sodann konkrete Handlungsempfehlungen betreffend Anforderungen für die Nutzung personenbezogener Daten (→ unten 3.), für die Verbesserung eines kontrollierten Zugangs zu personenbezogenen Daten (→ unten 4.) und für den allgemeinen Zugang zu Daten, insbesondere nicht-personenbezogenen Daten (→ unten 5.).

1. Allgemeine Anforderungen an den Umgang mit Daten

Ausgangspunkt für die Formulierung spezieller Prinzipien für den Umgang mit Daten sind die Unterschiede zwischen Daten und klassischen Ressourcen wie z. B. Öl oder Waren. Die spezifischen Charakteristika von Daten kommen vor allem darin zum Ausdruck, dass:

- Daten in einem **verteilten, dynamischen und prinzipiell nie völlig abgeschlossenen Prozess** durch das Zusammenwirken mehrerer Personen – welche in sehr verschiedenen Rollen auftreten (z. B. als Subjekt, über das Informationen erhoben werden, als Betreiber eines datengenerierenden Systems, als Entwickler) – entstehen und weiterverarbeitet werden;
- Daten ein **nicht-rivales Gut** sind, d. h. beliebig vervielfältigt und von einer Vielzahl von Personen parallel und in verschiedener Weise genutzt werden können;
- Daten **multifunktional und quer über alle Lebensbereiche einsetzbar** sind, wobei die Potenziale und Risiken von Daten in außergewöhnlichem Maße abhängig sind von den konkreten Zielen und Möglichkeiten eines Akteurs, insbesondere von der Verknüpfbarkeit mit anderen Daten, auch unter Berücksichtigung von Skaleneffekten.

1.1 Vorausschauende Verantwortung

Die Charakteristika von Daten, wie die außergewöhnlich hohe Dynamik und Abhängigkeit der Chancen und Risiken von der konkreten Konstellation, führen bei der Abwägung über die Sammlung, Nutzung oder Weitergabe von Daten zu einem besonderen Bedarf an vorausschauender Verantwortung. Bei der Abschätzung der Folgen, einschließlich der Möglichkeit einer Verletzung von Rechten anderer, sind insbesondere die folgenden Punkte zu bedenken und zu berücksichtigen:

- **Umfang** der entstehenden Datensammlungen, mit besonderem Augenmerk auf etwaigen Akkumulations-, Netzwerk- und Skaleneffekten;
- **Technologische Mittel** der Datenverarbeitung, wobei besonders an die derzeit zur Verfügung stehenden und zukünftigen technologischen **Möglichkeiten** durch größere Unternehmen und staatliche Einheiten (v. a. auch in Bezug auf die Rekombination und Entschlüsselung von Daten) zu denken ist;
- **Zweck** der Datenverarbeitung, unter besonderer Berücksichtigung möglicher Änderungen des Anwendungskontexts und der Akteurskonstellationen (z. B. durch Zugriff staatlicher Stellen oder durch Konzernübernahmen).

Bei personenbezogenen Daten hat das Prinzip vorausschauender Verantwortung in den von der DSGVO betonten Grundsätzen der Datenminimierung und der Speicherbegrenzung in typisierter Weise Ausdruck gefunden. Aber auch eine Fülle von Pflichten, angefangen von der Pflicht zu einer Datenschutzfolgenabschätzung bis hin zu Anforderungen an Vereinbarungen mit Auftragsverarbeitern, sind unmittelbar Ausfluss dieses Prinzips.

1.2 Achtung der Rechte beteiligter Personen

Die Nutzung von Daten muss stets die Rechte anderer respektieren. Handlungen und Unterlassungen, die ganz allgemein ethisch nicht vertretbar oder rechtswidrig sind, weil sie **Rechte anderer** verletzen, bleiben ethisch nicht vertretbar oder rechtswidrig, wenn sie mit Hilfe von Daten begangen werden (Beispiel: Betrug ist mit oder ohne Nutzung von Daten strafbar). Die Tatsache, dass Daten in einem verteilten Prozess und durch das Zusammenwirken mehrerer Personen generiert werden, kann aber auch dazu führen, dass Personen, welche an der Generierung der Daten in irgendeiner Weise beteiligt waren – etwa als Subjekt der Information oder als Eigentümer einer datengenerierenden technischen Vorrichtung – aus ethischer und möglicherweise auch aus rechtlicher Sicht **genuin datenspezifische Rechte (Datenrechte**) in Bezug auf diese Daten zustehen (→ näher unten 2). Diese Datenrechte müssen bei jeder Nutzung von Daten geachtet werden.

Achtung von Datenrechten anderer heißt dabei deutlich mehr, als nicht in fremde Rechtssphären – etwa ein fremdes Urheberrecht – einzudringen. Verlangt ist vielmehr aus ethischer Sicht eine umfassende **Rücksichtnahme** auf die datenbezogenen Interessen von Personen, die in spezifischer Weise mit den Daten verbunden sind und denen daher ein Recht auf Mitsprache und Teilhabe zukommt. Diese Pflicht zur Rücksichtnahme kann auch eine Pflicht zu aktivem Handeln, etwa zur Gewährung von bestimmten Formen des Datenzugangs, beinhalten.

Bei personenbezogenen Daten hat das Prinzip der Achtung der Datenrechte anderer vor allem Ausdruck gefunden in den von der DSGVO betonten **Grundsätzen der Rechtmäßigkeit, der Verarbeitung nach Treu und Glauben sowie der Zweckbindung**. In der DSGVO selbst normierte Datenrechte sind die Betroffenenrechte, etwa auf Auskunft, Berichtigung, Einschränkung der Verarbeitung, Löschung oder Datenübertragung (Portabilität).

1.3 Wohlfahrt durch Nutzen und Teilen von Daten

Ressourcen, die zum Wohl wichtiger Rechtsgüter Einzelner (z. B. Gesundheit) oder zum Wohl der Allgemeinheit – insbesondere zur Förderung einer der 17 Ziele der Vereinten Nationen für nachhaltige Entwicklung auf ökonomischer, sozialer und ökologischer Ebene – eingesetzt werden können, sollten nicht brachliegen. Ihre Nutzung ist dort, wo dies der umfassend verstandenen Wohlfahrt dient und keine überwiegenden Interessen – insbesondere keine Datenrechte anderer – entgegenstehen, grundsätzlich **ethisch geboten**.

Daten zeichnen sich durch das besondere Charakteristikum aus, nicht-rivale Güter zu sein. Sie nutzen sich bei ihrer parallelen Nutzung durch viele verschiedene Akteure zu vielen verschiedenen Zwecken nicht ab und sind nahezu beliebig vervielfältigbar. Durch das **Teilen von Daten** kann ein Zustand eintreten, bei dem die teilende Partei zumindest nicht schlechter, alle anderen im weiteren Sinne Beteiligten dafür aber besser stehen als wenn das Teilen der Daten unterblieben wäre. Diesem Umstand sollte ein ethisch verantwortlicher Umgang mit Daten Rechnung tragen. Teilen von Daten hat zudem immense Bedeutung für die Sicherstellung eines **fairen und effizienten Wettbewerbs**.

Allerdings kann das Prinzip der Nutzung und des Teilens von Daten in einem Spannungsverhältnis mit dem Prinzip vorausschauender Verantwortung und dem Prinzip der Achtung der Datenrechte anderer stehen, ebenso wie mit Erwägungen zu einem angemessenen Leistungsschutz. Daher sollten Anreize zum **freiwilligen Teilen** stets Vorrang genießen und eine gesetzliche Pflicht zum Teilen die Ausnahme sein.

1.4 Zweckadäquate Datenqualität

Daten – zusammen genommen mit ihrem Kontext und der Semantik – sind gespeicherte Information. Information geht regelmäßig mit dem Anspruch einher, ein möglichst getreues Abbild der gegenwärtigen Realität oder eine möglichst treffsichere Voraussage einer künftigen Realität zu sein. Während es jenseits der automatisierten Verarbeitung von Daten durch algorithmische Systeme für alle offenkundig ist, dass Fehlinformationen nicht nur wertlos, sondern schädigend sein können, kommt es durch die Automatisierung oft zu einer verführerischen **Scheinobjektivität** und einer gefährlichen Bereitschaft, sich trotz einer falschen oder unvollständigen Datenbasis auf ein Berechnungsergebnis zu verlassen, das genauso schlecht ist wie seine Datenbasis *(„Garbage in, garbage out")*.

Ein verantwortungsvoller Umgang mit Daten in der Datengesellschaft setzt daher im Interesse aller auch das Bemühen um eine **dem Einsatzzweck angemessene Datenqualität** voraus (→ siehe oben Teil C, 2.1.1). Die Bestimmung dessen, was jeweils eine „angemessene" Datenqualität bedeutet, muss jedoch stets **kontextspezifisch** erfolgen. Beispielsweise ist zu berücksichtigen, dass Daten gesellschaftliche Vorannahmen, Stereotypen und Diskriminierungen abbilden können, welche in Folge die Funktionsweise eines algorithmischen Systems bestimmen, das mit Hilfe dieser Daten trainiert wird (→ näher unten Teil F, 2.6). Insofern kann es geboten sein, das getreue Abbild eines bestehenden Defizits, das etwa für statistische Zwecke qualitativ hochwertig sein kann, gerade nicht als Datengrundlage für andere Zwecke zugrunde zu legen.

In diesem Zusammenhang ist auch zu berücksichtigen, dass Daten über verschiedene Lebensbereiche hinweg und zu unterschiedlichen Zwecken einsetzbar sind. In diesem Zusammenhang kann daher auch das sog. **FAIR-Prinzip** *(Findable, Accessible, Interoperable, Reusable)* relevant sein, welches etwa die Modi der Speicherung und Kodierung von Daten betrifft. Danach sollten Daten möglichst so aufbereitet und gespeichert sein, dass sie auffindbar und zugänglich sind, dass sie in einem gängigen Format kodiert sind und in einer Weise, die kontextabhängig möglichst vielen Akteuren die weitere Nutzung der Daten ermöglicht.

Bei personenbezogenen Daten hat das Streben nach einem hohen Maß an Datenqualität in dem von der DSGVO betonten **Prinzip der Richtigkeit** Ausdruck gefunden.

1.5 Risikoadäquate Informationssicherheit

Daten können beliebig vervielfältigt werden. Sind sie einmal in andere Hände gelangt, können sie **kaum zurückgeholt** werden. Sie sind zudem aufgrund zahlreicher und vielfach unbemerkt bleibender **Angriffsmöglichkeiten** von außen besonders verletzlich gegenüber Verfälschung und Zerstörung. In unmittelbarem Zusammenhang mit dem Prinzip vorausschauender Verantwortung ebenso wie dem Prinzip der Achtung von Rechten beteiligter Personen steht daher in technischer Hinsicht ein hohes und dem jeweiligen Risikopotenzial angemessenes Maß an **Informationssicherheit**. Ausreichende Informationssicherheit, die eine breite Palette von Maßnahmen auf unterschiedlichen Ebenen umfasst, ist eine notwendige Voraussetzung für vertrauensvolles Handeln in der Datengesellschaft.

Bei personenbezogenen Daten hat Informationssicherheit in dem von der DSGVO betonten **Prinzip der Integrität und Vertraulichkeit** Ausdruck gefunden.

1.6 Interessenadäquate Transparenz

Die Tatsache, dass die faktische Kontrolle und die Nutzung von Daten auch Einfluss und Macht bedeuten kann, bedingt, dass derjenige, der Daten faktisch kontrolliert und nutzt, prinzipiell bereit und in der Lage sein muss, für sein Handeln **Rechenschaft** abzulegen. Das gilt auch und gerade zum Schutze derjenigen, deren Datenrechte potenziell betroffen oder gar verletzt sind. Damit diese Personen, oder aber auch Stellen, die zur Wahrnehmung der Datenrechte anderer berufen sind, überhaupt feststellen können, ob und inwieweit Datenrechte tatsächlich betroffen oder gar verletzt sind und wem gegenüber sie Ansprüche geltend machen können, bedarf es einer **den potenziell betroffenen Interessen angemessenen Transparenz**.

Bei personenbezogenen Daten ist die Transparenz im Sinne der **Nachvollziehbarkeit der Datenverarbeitung** für betroffene Personen ebenso ein festes Grundprinzip der DSGVO wie die **Rechenschaftspflicht**. Eine Vielzahl von Regelungen der DSGVO, etwa betreffend Information, Dokumentation oder Auskunftsrechten, sollen Transparenz gewährleisten.

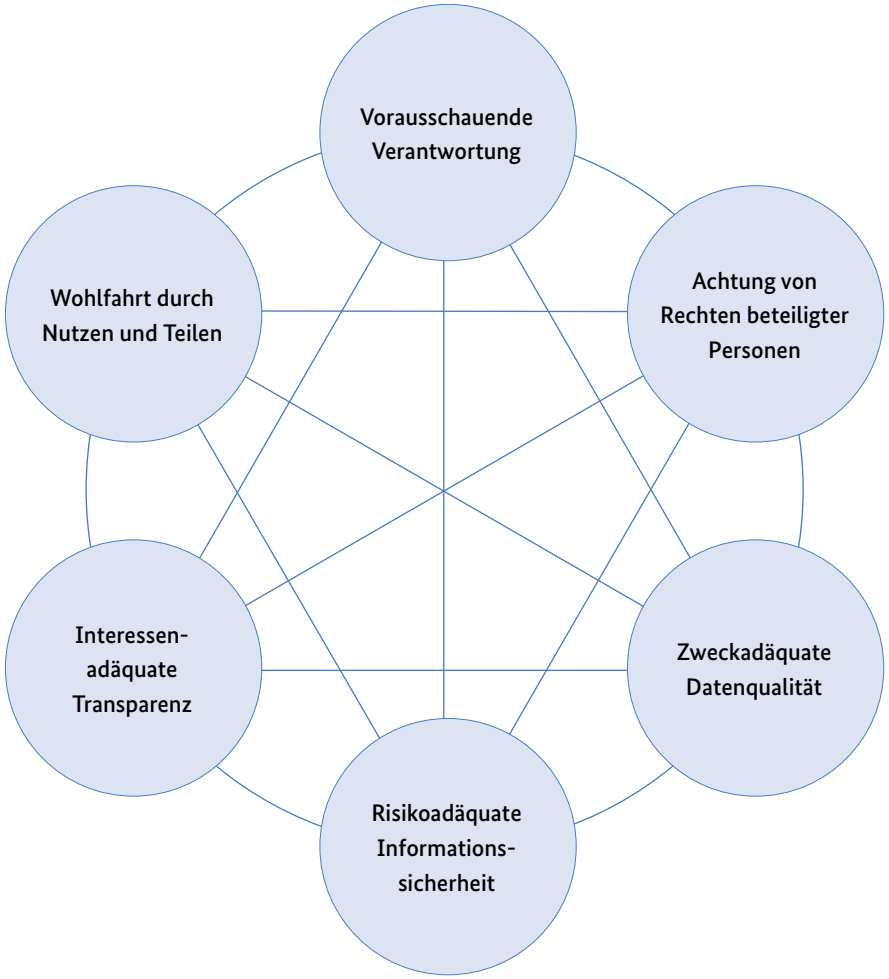


Abbildung 5: Anforderungen an den Umgang mit Daten

2. Datenrechte und korrespondierende Datenpflichten

Das ethische Prinzip digitaler Selbstbestimmung begreift den Einzelnen nicht nur als passiv, schutzbedürftig und aktuell oder potenziell bedroht, sondern als **selbstbestimmten Akteur in der Datengesellschaft**. Um sich als Akteur in der Datengesellschaft selbstbestimmt bewegen zu können, bedarf es subjektiver Rechte, die dem Einzelnen gegenüber anderen Akteuren zustehen. Dies betrifft in erster Linie die Rechte eines jeden Menschen in Bezug auf seine **personenbezogenen Daten**, die sich aus dem grundrechtlich verbürgten Recht auf informationelle Selbstbestimmung ableiten und durch das geltende Datenschutzrecht gewährleistet werden. Digitale Selbstbestimmung umfasst darüber hinaus auch die selbstbestimmte wirtschaftliche Verwertung der eigenen Datenbestände sowie den selbstbestimmten Umgang mit **nicht-personenbezogenen** Daten, die etwa durch den Wirkbetrieb eigener Geräte generiert werden. Nach Auffassung der DEK gilt ein Recht auf digitale Selbstbestimmung prinzipiell auch für Unternehmen und **juristische Personen** und – zumindest in Ansätzen – für Gruppen von Personen (Kollektive). Vor diesem Hintergrund sieht die DEK allgemeine, über Datenschutz hinausgehende, Grundsätze von Datenrechten und Datenpflichten.[1]

2.1 Allgemeine Grundsätze von Datenrechten und Datenpflichten

In komplexen Prozessen der Generierung von Daten – verstanden in einem weiteren Sinne, einschließlich verschiedener Phasen der Datenherstellung, Datenanreicherung und Datenveredelung – interagieren häufig unterschiedliche Akteure mit unterschiedlichen Zielen miteinander und tragen dabei in unterschiedlichen Rollen zur Generierung von Daten bei. Ein relevanter **Beitrag eines Akteurs** (natürliche oder juristische Person) **zur Generierung von Daten** kann darin bestehen, dass

a) sich die in den Daten gespeicherten Informationen in ihrer Bedeutung auf diesen Akteur, oder auf einen mit diesem Akteur verbundenen (z. B. ihm gehörenden) Gegenstand, beziehen;

b) die Daten durch eine Aktivität dieses Akteurs, oder durch Verwendung eines ihm gehörenden Gegenstands (z. B. eines Sensors), generiert wurden; oder

c) die Daten durch Software oder eine andere Komponente (z. B. Sensoren) generiert wurden, welche dieser Akteur geschaffen hat oder in welche er investiert hat.

Dabei kommt der unter a) genannten Situation, dass ein Akteur Subjekt der in den Daten gespeicherten Information ist, bei natürlichen Personen eine herausgehobene Bedeutung zu, ist sie doch zugleich Anknüpfungspunkt für das verfassungsrechtlich geschützte Recht auf informationelle Selbstbestimmung und Datenschutz.

Ein Beitrag zur Generierung von Daten führt angesichts der spezifischen Charakteristika von Daten sowie – bei personenbezogenen Daten – angesichts der untrennbaren Verknüpfung mit Persönlichkeitsrechten nach Auffassung der DEK jenseits des geltenden Immaterialgüterrechts nicht zu exklusiven Eigentumsrechten an Daten (→ siehe und 5.2.4). Vielmehr folgen aus einem solchen Beitrag Datenrechte eines Akteurs in Gestalt von **Mitsprache- und Teilhaberechten**, mit denen korrespondierende Pflichten anderer Akteure einhergehen. Zwischen einem Akteur, der an der Generierung von Daten beteiligt war, und einem Akteur, der diese Daten faktisch kontrolliert, entsteht aus ethischer Sicht daher eine **dynamische Sonderbeziehung**. Diese Beziehung kann mehr oder minder langfristig sowie stärker oder schwächer ausgeprägt sein. Bezüglich personenbezogener Daten ist sie weitgehend durch das geltende Datenschutzrecht determiniert.

Die **Anerkennung und Ausgestaltung** von Datenrechten und korrespondierenden Datenpflichten in dynamischen Umgebungen hängt aus ethischer Sicht von den folgenden allgemeinen Faktoren ab, die dort, wo Datenrechte und Datenpflichten bereits gesetzlich konkretisiert wurden, regelmäßig auch der rechtlichen Beurteilung zugrunde liegen:

1 Modell der Datenrechte und Datenpflichten in Anlehnung an Vorentwürfe Nr. 2 (Februar 2019) und Nr. 3 (Oktober 2019) der Principles for a Data Economy des European Law Institute (ELI) und des American Law Institute (ALI), die der DEK zur Verfügung gestellt wurden. Die Vorentwürfe sind bislang weder von ALI noch von ELI verabschiedet worden und stellen noch nicht die offizielle Position einer oder beider Organisationen dar.

a) Umfang und Art des **Beitrags zur Datengenerierung** desjenigen Akteurs, der ein Datenrecht geltend macht;

b) **Gewicht des Individualinteresses** desjenigen Akteurs, der das Datenrecht geltend macht, an der Gewährung des Datenrechts (insbesondere an Unterlassung/Zugang/Korrektur/wirtschaftlicher Teilhabe);

c) Gewicht von ggf. **konfligierenden Individualinteressen** desjenigen Akteurs, dem gegenüber das Datenrecht geltend gemacht wird, oder Dritter, unter Berücksichtigung von Ausgleichsmöglichkeiten (z. B. Schutzmaßnahmen, Vergütung);

d) Gewicht von **Interessen der Allgemeinheit**;

e) **Machtverteilung** zwischen dem Akteur, der das Datenrecht geltend macht, und dem Akteur, dem gegenüber das Datenrecht geltend gemacht wird.

Diese Faktoren wirken im Wege eines beweglichen Systems zusammen, d. h., dass etwa ein besonders stark ausgeprägtes Allgemeininteresse am Datenzugang den besonders schwach ausgeprägten Beitrag zur Datengenerierung ausgleichen kann. Dabei sind die in Teil B dargelegten allgemeinen Grundsätze und Prinzipien stets zu berücksichtigen, so dass es nicht zu einer Aushöhlung zentraler Individualinteressen durch tatsächliche oder bloß vermeintliche Allgemeininteressen kommen kann. Die Faktoren bestimmen auch die **Konkretisierung und Ausgestaltung** z. B. von Formaten, Fristen, Schutzmaßnahmen oder finanzieller Entschädigung. Dazu gehört auch die Frage, ob nur auf Ansuchen desjenigen, der ein Datenrecht geltend macht (z. B. Datenzugangsanspruch), oder auch proaktiv (z. B. Datenveröffentlichungspflicht) gehandelt werden muss.

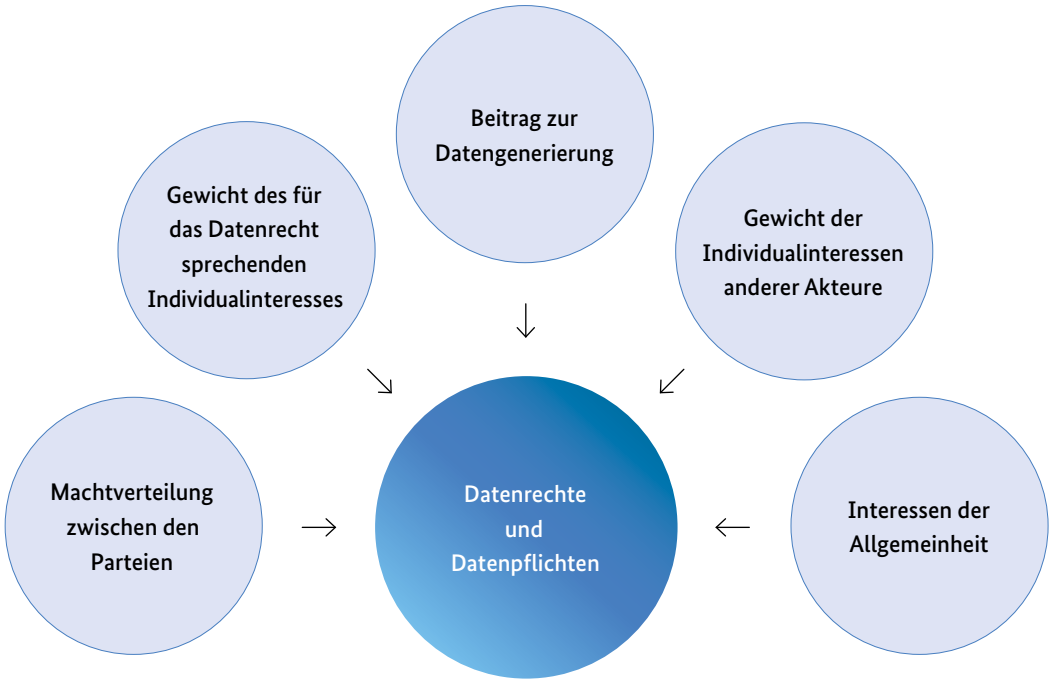


Abbildung 6: Allgemeine Faktoren zur Ausgestaltung von Datenrechten und korrespondierenden Datenpflichten

Die **Betroffenenrechte** der DSGVO sind eine besonders wichtige und – weil einheitlich an der Qualifikation von Daten als personenbezogen anknüpfend – in gewisser Weise typisierte Ausprägung dieser Grundsätze speziell zum Schutz derjenigen natürlichen Person, auf die sich die Information bezieht. Die hier formulierten Grundsätze können allerdings auch für nicht-personenbezogene Daten herangezogen werden, und sie gelten nicht nur für Individuen, sondern auch für juristische Personen und für Kollektive.

2.2 Konkretisierung der allgemeinen Grundsätze anhand typischer Szenarien

Von der **Zielrichtung** her können Datenrechte insbesondere auf eine Unterlassung der Datennutzung (bis hin zur Löschungspflicht), auf Zugang zu Daten (z. B. Offenlegung, Übertragung, volle Portabilität), auf eine Korrektur von Daten oder auf wirtschaftliche Teilhabe ausgerichtet sein.

2.2.1 Unterlassungs-Szenarien

Vielfach wird eine Situation gegeben sein, in welcher ein Akteur von einem anderen Akteur verlangt, eine bestimmte Datennutzung zu unterlassen. In Bezug auf personenbezogene Daten geht die DSGVO sogar vom Grundsatz der Unterlassungspflicht aus, sofern nicht eine Rechtsgrundlage gegeben ist und die übrigen Anforderungen eingehalten werden.[2] Auch jenseits des Anwendungsbereichs der DSGVO und ganz allgemein kann sich das Gewicht des Individualinteresses desjenigen Akteurs, der das Datenrecht geltend macht, aus ethischer Sicht zu einem **Unterlassungsanspruch** – bis hin zu einem Anspruch auf Löschung der Daten – verdichten, wenn die Datenverarbeitung

a) diesem oder einem anderen Akteur Schaden zufügen könnte; und

b) unvereinbar ist mit den Umständen, unter denen der Beitrag zur Datengenerierung geleistet wurde, insbesondere, weil

   (i) dies zu einem anderen Zweck erfolgte und nicht zu erwarten ist, dass der Akteur den Beitrag freiwillig geleistet hätte, wenn er die jetzige Datenverarbeitung vorhergesehen hätte; oder

   (ii) sein Einverständnis aus übergeordneten Gründen unwirksam wäre.

Bevor ein Unterlassungsanspruch bejaht werden kann, ist das so konkretisierte Individualinteresse jedoch noch mit den oben (→) genannten weiteren Faktoren in Abwägung zu bringen. Unterlassung kann daher etwa nicht begehrt werden, wenn die Datenverarbeitung ganz ausnahmsweise aus zwingenden Gründen (z. B. Verfolgung von Straftaten) dennoch gerechtfertigt ist.

2 Art. 6 Abs. 1, Art. 9 Abs. 1 DSGVO.

In Bezug auf **nicht-personenbezogene Daten** kann ein Unterlassungsanspruch beispielsweise in Wertschöpfungsketten und Kundenbeziehungen Bedeutung haben, in denen nicht-personenbezogene Daten von großer wirtschaftlicher Bedeutung sind, die Interessen der Beteiligten an einer Unterlassung aber durchaus gewichtig sein können (→ unten 5.3).

Beispiel 1

Von den Sensoren moderner Landmaschinen gesammelte nicht-personenbezogene Daten (Bodenqualität, Wetter usw.) werden von den Herstellern für die Erbringung zahlreicher Dienstleistungen (Precision Farming, Predictive Maintenance u.a.) genutzt. Würde der Hersteller die Daten auch an mögliche Investoren oder an Verpächter weiterleiten, erhielten diese damit Informationen, welche dem landwirtschaftlichen Betrieb bei künftigen Verhandlungen über seine Flächen schaden können. Es ist nicht davon auszugehen, dass ein landwirtschaftlicher Betrieb freiwillig zu diesem Zweck an der Generierung der Daten mitgewirkt hätte. Bei der ethischen Bewertung eines Unterlassungsanspruchs ist neben dem konkreten Machtverhältnis auch zu berücksichtigen, dass der Betrieb einen sehr gewichtigen Beitrag zur Generierung der Daten geleistet hat. An schutzwürdigen Interessen Dritter kämen nur das Erwerbsinteresse des Herstellers und ein Allgemeininteresse an der korrekten Information für Investoren, Verpächter etc. in Betracht.

Ein **Verzicht** auf einen eigentlich begründeten Unterlassungsanspruch ist aus ethischer Sicht nur begrenzt möglich. Er verbietet sich von selbst, wenn das Einverständnis zur Datennutzung im Sinne der Voraussetzung b. (ii) aus übergeordneten Gesichtspunkten heraus unwirksam wäre, etwa weil es gegen das Gesetz oder die guten Sitten verstoßen würde, da nach unserer Rechts- und Werteordnung keine beliebige Selbst- oder Fremdschädigung akzeptiert werden kann. Soweit dies nicht der Fall ist, kann gegebenenfalls bei Erfüllung strenger Anforderungen an den Verzicht – etwa durch eine gesonderte Vereinbarung ohne Druck oder Koppelung mit anderen Leistungen – Freiwilligkeit gesichert werden, womit Voraussetzung b) (i) entfallen würde.

In könnte der landwirtschaftliche Betrieb die Datenweitergabe an Dritte gestatten – beispielsweise aufgrund einer individuellen Vereinbarung mit entsprechender Vergütung und ohne, dass die Nutzung des Traktors davon abhängig gemacht würde.

Für **personenbezogene Daten** folgen Unterlassungspflichten zwar regelmäßig bereits aus dem geltenden Datenschutzrecht, doch können die genannten Kriterien etwa herangezogen werden, um zu entscheiden, ob **die materiellen Grenzen der Einwilligung** überschritten werden (→ unten 3.2.1), oder um die Abwägung berechtigter Interessen zu konkretisieren.

Beispiel 2

Vom Nutzer eines sozialen Netzwerks wird mittels Daten über das Nutzungsverhalten ein umfassendes Persönlichkeitsprofil angelegt, das u.a. die Punkte „psychisch labil" und „Esoterik" beinhaltet. Er wird in der Folge fast täglich – oft in zeitlichem Zusammenhang mit Postings, die psychische Anspannung signalisieren – mit Angeboten über teure persönliche Horoskope, Leistungen von „Energetikern" usw. konfrontiert, welche er vielfach annimmt. Bei Einrichtung seines Nutzerkontos hatte er ein Kästchen mit folgendem Text angeklickt: „Ich möchte, dass meine Daten im Hinblick auf persönliche Präferenzen und Eigenschaften ausgewertet werden, um Dienste, auch von Drittanbietern, besser personalisieren zu können (Profiling)". Diese „Einwilligung" macht die Verarbeitung aber nicht zulässig. Das kann auf verschiedene Weise begründet werden, u.a. damit, dass die Verarbeitung zu diesem Zweck dem Nutzer erheblichen Schaden zufügt und dies mit den Umständen, unter denen er die Daten generiert hat, unvereinbar ist (etwa weil er bei Kenntnis der Zusammenhänge mit diesem Zweck die Daten nicht generiert hätte, und weil die Rechtsordnung die Ausnutzung solcher psychischer Zustände missbilligt, vgl. § 138 BGB).

Vielfach wird es Unterlassungspflichten geben, die durch keine Einwilligung oder Abwägung relativiert werden können, wobei oft von „roten Linien" oder **„absoluten Grenzen"** die Rede ist. Diese Grenzen müssen nicht datenspezifisch sein; und die meisten sind es auch nicht. Beispielsweise wäre eine dem Demokratieprinzip zuwiderlaufende Beeinflussung von Wahlen mit oder ohne Nutzung von Daten zu unterlassen. Eine datenspezifische absolute Grenze ist nach Auffassung der DEK etwa die Totalüberwachung von Menschen.

Beispiel 3

Eine Angestellte verpflichtet sich bei Abschluss des Arbeitsvertrags, die Standortfunktion von SmartWatch und Mobiltelefon sowie eine Reihe von Daten erhebenden Applikationen (u.a. zum Monitoring des Schlafverhaltens und von Emotionen) auch im Privatleben stets eingeschaltet zu lassen und die Geräte dem Arbeitgeber jederzeit auf Aufforderung zum Auslesen der Daten zur Verfügung zu stellen. Selbst wenn die Angestellte in jede dieser Maßnahmen eingewilligt haben sollte, und selbst wenn sie sich aus freien Stücken für diesen Arbeitgeber entschieden hat und ebenso gut das Angebot eines anderen Arbeitgebers hätte annehmen können, ergibt sich mindestens in der Gesamtschau eine vollständige oder annähernde Totalüberwachung, die mit der Menschenwürde, der Selbstbestimmung und der Privatheit nicht vereinbar ist.

Umgekehrt können die für Unterlassungs-Szenarien geltenden Kriterien auch indirekt relevant werden, wenn es um ethische oder gar rechtliche **Pflichten zur Datennutzung** geht. Eine solche kann insbesondere dann bestehen, wenn einen Akteur die Pflicht trifft, wichtige Rechtsgüter zu schützen und er über Daten verfügt, deren Nutzung geeignet ist, diesen Schutz zu gewährleisten oder zu verbessern. Die Pflicht zum Schutz wichtiger Rechtsgüter kann dann zu einer Pflicht zur Nutzung dieser Daten führen, jedenfalls sofern kein entgegenstehender begründeter Unterlassungsanspruch eines anderen Akteurs besteht.

Beispiel 4

Ein Krankenhaus hat Probleme mit einem multiresistenten Keim. Um bessere Erkenntnisse zur Anfälligkeit bestimmter Patienten zu erhalten, so dass diese möglicherweise gezielt in ein anderes Haus verlegt werden können, wäre es erforderlich, die Gesundheitsdaten derjenigen Patienten zu analysieren, die in der letzten Zeit mit dem Keim angesteckt wurden. In einer solchen Situation hat das Krankenhaus generell die Pflicht, neue Patienten bestmöglich vor einer Infektion zu schützen und dazu alle verfügbaren und zumutbaren Vorsichtsmaßnahmen zu ergreifen. Dies umfasst auch die Nutzung der Gesundheitsdaten der Patienten, die sich bereits angesteckt hatten, sofern dies neue Patienten schützen kann und gegenüber den bereits angesteckten Patienten keine Unterlassungspflicht besteht.

2.2.2 Zugangs-Szenarien

Bei der Frage des Zugangs zu Daten wird es zunächst sehr viele Situationen geben, in denen sich die Zugang suchende und die die Daten faktisch kontrollierende Partei „handelseinig" werden. Sofern keine überwiegenden Interessen Dritter oder der Allgemeinheit dagegen sprechen, insbesondere keinem Akteur nach den oben genannten Kriterien ein Unterlassungsanspruch zusteht, sind solche **freiwilligen Arrangements** zu begrüßen. Angesichts des hohen Wertschöpfungspotenzials, das mit der Verfügbarkeit und Auswertung von Daten einhergehen kann, wird jedoch auch intensiv diskutiert, unter welchen Voraussetzungen und Bedingungen ein Zugang zu Daten aus ethischer Sicht gewährt werden soll oder gar muss.[3]

Dabei ist zunächst an Situationen zu denken, in denen die Erfüllung einer besonderen, vielfach sogar gesetzlich festgelegten, **Pflicht oder Aufgabe** (z. B. Strafverfolgung, Sorge für die öffentliche Gesundheit) den Zugang zu Daten erfordert. Ein etwaiges Recht, Zugang zu Daten zu erhalten, folgt dann den für diese Pflicht oder Aufgabe geltenden Regeln, wobei vor allem dem **Grundsatz der Verhältnismäßigkeit** überragende Bedeutung zukommt und stets etwaige Unterlassungsansprüche (→ oben) betroffener Akteure zu prüfen sind.

Zudem kann es zu einer Geltendmachung von selbständigen Ansprüchen auf Zugang zu Daten kommen, etwa innerhalb **bestehender Wertschöpfungssysteme**, in denen meist viele Akteure in unterschiedlichen Rollen (z. B. als Zulieferunternehmen, Hersteller, Händler, Endnutzer) zur Generierung von Daten beitragen und dabei sowohl die eigenen Rollen als auch die Rollen anderer Akteure prinzipiell kennen und akzeptieren (→ näher unten 5.3). Das Individualinteresse eines Akteurs, welches für die Zugangsgewährung geltend gemacht wird, kann dann insbesondere darin bestehen, dass die Daten erforderlich sind für die

3 Siehe anstelle vieler: Europäische Kommission: Aufbau einer europäischen Datenwirtschaft, COM(2017) 9 final, 10.1.2017, S. 11 ff (abrufbar unter: *https://ec.europa.eu/transparency/regdoc/rep/1/2017/DE/COM-2017-9-F1-DE-MAIN-PART-1.PDF*); Europäische Kommission: Aufbau eines gemeinsamen europäischen Datenraums, COM(2018) 232 final, 25.4.2018, S. 10 ff (abrufbar unter: *https://ec.europa.eu/transparency/regdoc/rep/1/2018/DE/COM-2018-232-F1-DE-MAIN-PART-1.PDF*).

a) bestimmungsgemäße Nutzung eines im Rahmen des Wertschöpfungssystems genutzten Gutes (z. B. Reparatur einer vernetzten Maschine durch den Endnutzer);

b) Qualitätskontrolle und -verbesserung einer im Rahmen des Wertschöpfungssystems erbrachten Leistung (z. B. seitens eines Zulieferunternehmens);

c) Wahrheitsfindung bzw. Beweisführung (z. B. in einem Rechtsstreit mit Dritten);

d) Vermeidung von wettbewerbswidrigen Effekten (z. B. Lock-in-Effekten); oder eine

e) neue Wertschöpfung mit Hilfe der Daten (z. B. durch Entwicklung eines Smart Service).

Beispiel 5

Ein Zulieferunternehmen stellt die Motoren für die Landmaschinen in her. Damit das Zulieferunternehmen die Qualität seiner Motoren überprüfen und stetig verbessern kann, wäre es für das Unternehmen sehr wichtig, Zugang zu bestimmten Traktordaten zu erhalten. Diese werden allerdings in der Cloud des Herstellers gespeichert, der dem Zulieferunternehmen keinen Zugang gewähren will. In dieser Situation wäre zu berücksichtigen, dass der Zulieferer einen signifikanten Beitrag zur Generierung der Motorendaten geleistet hat und die Daten dringend für die Qualitätsverbesserung einer Leistung im selben Wertschöpfungssystem benötigt, an dem auch der Hersteller beteiligt ist. Neben den konkreten Machtverhältnissen wäre auch zu berücksichtigen, dass alle Beteiligten, einschließlich der Allgemeinheit, ein Interesse an guter Motorenqualität haben. Auf der Seite des Herstellers könnten ökonomische Interessen, insbesondere Geheimhaltungsinteressen, zu berücksichtigen sein.

Zugangsansprüche auf ähnlicher Basis werden auch diskutiert, wenn die Zugang suchende Partei und die Partei, die die Daten faktisch kontrolliert, zwar noch nicht Teil desselben Wertschöpfungssystems sind, ein **neues Wertschöpfungssystem** aber gerade geschaffen werden könnte. In solchen Situationen fällt die Bewertung anhand der allgemeinen Kriterien meist anders aus, schon weil die Zugang suchende Partei oft überhaupt keinen Beitrag zur Generierung der Daten geleistet hat und es eher **allgemeine Wohlfahrtserwägungen** oder spezielle Erwägungen, etwa zur Sicherung des **Wettbewerbs**, sind, welche für einen Datenzugang ins Feld geführt werden können (→ näher unten 5.5).

Beispiel 6

Angenommen in hätte der Hersteller – ein beherrschendes Unternehmen auf dem Traktorenmarkt – Boden- und Wetterdaten über Jahrzehnte gesammelt. Ein Start-up erkennt das Potenzial, mit den Daten eine Datenbank für Investoren aufzubauen und begehrt nun Zugang zu den Daten. Hier wäre zu berücksichtigen, dass das Start-up selbst keinen Beitrag zur Generierung der Daten geleistet hat. Ob und welches Allgemeininteresse am Datenzugang besteht, hängt davon ab, ob der Hersteller seine Marktmacht missbraucht, sowie davon, welche Bedeutung die Brechung der Marktmacht weniger Unternehmen für die gedeihliche Entwicklung der europäischen Wirtschaft hat (falls das Start-up überhaupt in Europa arbeitet). In jedem Fall gilt es zu bedenken, dass Geschäftsgeheimnisse und andere Interessen Dritter – wie etwa des Herstellers und der landwirtschaftlichen Betriebe in Beispiel 1 – durch die Datenoffenlegung potenziell massiv beeinträchtigt werden.

Zu den allgemein anerkannten Prinzipien von **Open Government Data**, also der Zurverfügungstellung von Daten der öffentlichen Hand an Private, gehören Prinzipien wie „standardmäßig offen“ und „verwendbar von allen zu jedem Zweck“.[4] Viele wollen diese Prinzipien im Sinne weitergehender Open-Data-Konzepte auch auf Daten ausweiten, die bei Privaten entstanden sind und von ihnen faktisch kontrolliert werden. Eine schwierige ethische Frage ist im Zusammenhang mit Open Data, inwieweit eine typisierte, d. h. den konkreten Einzelfall nicht mehr berücksichtigende Beurteilung von Interessen der Allgemeinheit erfolgen darf.

Beispiel 7

Eine Stadt erhebt zur Erleichterung der Verkehrsplanung (u.a. Anpassung der Taktung öffentlicher Verkehrsmittel) in großem Umfang Mobilitätsdaten mithilfe von Smartphone-Signalen. Die Daten sind theoretisch „anonymisiert“, doch lässt sich bei Zusammenführen mit anderen Datensätzen und etwas Zusatzwissen ein bestimmter Smartphone-Besitzer mit 95%iger Wahrscheinlichkeit identifizieren. Für diese Daten interessieren sich u.a.: ein Forscher, der Erkenntnisse für die optimale Gestaltung von Erholungsflächen im Stadtgebiet gewinnen möchte; ein Start-up mit der Geschäftsidee einer Online-Detektei, bei der man gegen Entgelt Mobilitätsprofile seines Ehepartners, Konkurrenten usw. abfragen kann; ein Forschungsinstitut, das im Auftrag der Regierung eines ausländischen Staates Erkenntnisse über die politischen Aktivitäten ihrer Staatsbürger erlangen soll. Bei individueller Beurteilung wären die drei Zugangsverlangen sehr unterschiedlich zu bewerten. Es stellt sich daher die Frage, unter welchen Bedingungen die Stadt im Hinblick auf viele mögliche gemeinwohlfördernde Nutzungen die Daten offen bereitstellen darf oder gar muss.

Die DEK betont in diesem Zusammenhang die Bedeutung der individuellen (denkbaren) Unterlassungsansprüche derjenigen Akteure, die zur Generierung der Daten beigetragen haben – insbesondere derjenigen, auf die sich die Daten beziehen. Das bedeutet, dass nicht nur unter Abwägung des Schädigungspotenzials und des zu erwartenden Gemeinwohlnutzens alle möglichen und zumutbaren Schutzmaßnahmen (einschließlich ständig neu zu verbessernder Anonymisierungstechniken) zu ergreifen sind, sondern dass sich eine pauschalierende Zugangsgewährung je nach Schädigungspotenzial auch ganz verbieten kann (→ näher unten).

2.2.3 Korrektur-Szenarien

Daten können qualitativ schlecht sein. Insbesondere können der Kontext unpassend, die Kodierung **falsch** oder die Daten in einem Maße **unvollständig** sein, dass die mit ihrer Hilfe gewonnenen Ableitungen falsch werden. In derartigen Konstellationen kann sich ein ethisch begründeter Anspruch eines Akteurs, der an der Generierung von Daten beteiligt war, auf Korrektur der zu Grunde gelegten Daten oder der mit ihrer Hilfe gewonnen Ableitungen ergeben. Da grundsätzlich weder ein schützenswertes Individualinteresse noch ein Allgemeininteresse an der Verarbeitung falscher oder unvollständiger Daten besteht, sind die Hürden für einen derartigen Anspruch gering. In der Regel ist es ausreichend, dass

a) die Verarbeitung der falschen oder unvollständigen Daten diesem Akteur (insbesondere einem Akteur, auf den sich die Informationen beziehen) Schaden zufügen kann; und

b) die Korrektur unter Berücksichtigung von Schwere und Wahrscheinlichkeit des Schadens einerseits und dem für die Korrektur erforderlichen Aufwand andererseits nicht unverhältnismäßig ist.

4 Siehe Erwägungsgrund 16 der Richtlinie (EU) 2019/1024 über offene Daten und die Weiterverwendung von Informationen des öffentlichen Sektors (PSI-Richtlinie); Principles 1 und 3 der auf dem G8-Gipfel am 18. Juni 2013 unterzeichneten G8 Open Data Charter; Principle 1 der im September 2015 auf dem Gipfel der Open Government Partnership unterzeichneten International Open Data Charter.

Beispiel 8
Die bei dem Hersteller in Beispiel 5 gespeicherten Daten betreffend die Motoren des Zulieferunternehmens stellen sich als grob fehlerhaft heraus. Dies ist für das Zulieferunternehmen nicht nur deswegen misslich, weil das Zulieferunternehmen mit diesen Daten seiner Aufgabe der Qualitätssicherung nur unvollkommen nachkommen kann, sondern auch deswegen, weil die Motorendaten mit den Motorendaten anderer Motorenhersteller gepoolt und ausgewertet werden und schlechte Leistungswerte der Motoren des betroffenen Zulieferunternehmens dessen Chancen, Aufträge anderer Hersteller zu erlangen, schmälern dürften. Hier kann die Verarbeitung falscher Daten dem Zulieferunternehmen Schaden zufügen, und Anhaltspunkte für eine Unverhältnismäßigkeit des Aufwands sind nicht gegeben.

Wenn der Aufwand für die Korrektur zu groß, der mögliche Schaden aber schwerwiegend ist, ist regelmäßig ein Unterlassungsanspruch gegeben (→ oben).

2.2.4 Szenarien wirtschaftlicher Teilhabe

Situationen, in denen ein Akteur Daten nutzt, zu deren Generierung andere Akteure beigetragen haben, und in denen der Daten nutzende Akteur mit Hilfe der Daten Wertschöpfung betreibt, sind alltäglich und grundsätzlich auch erwünscht. Sofern nach den genannten Kriterien (→ oben) kein Unterlassungsanspruch besteht, ist dies von den Akteuren, die zur Generierung der Daten beigetragen haben, hinzunehmen. Der **gemeinwohlbezogene Charakter der hier vertretenen Datenrechte und Datenpflichten** steht einem generellen Anspruch auf Vergütung solcher Akteure normalerweise entgegen. Vielmehr müssen sich solche Akteure mit kollektiven Teilhabemöglichkeiten – insbesondere über die Besteuerung von Wertschöpfung – zufrieden geben.

Soweit ein Vergütungsanspruch nicht aus einem wirksamen Vertrag folgt, kommt individuelle Vergütung allenfalls als Ausgleichsmaßnahme im Einzelfall in Betracht, etwa soweit die entschädigungslose Ausübung eines Datenrechts konkret unverhältnismäßig erschiene (→ vgl. oben 2.1, Faktor c). Nur ganz **ausnahmsweise** kann nach Auffassung der DEK einem Akteur, der zur Generierung von Daten beigetragen hat, auch ohne entsprechenden Vertrag aus ethischer Sicht eine **eigenständige Vergütung** für die Datennutzung durch andere zustehen. Dies sollte aber nur der Fall sein, wenn

a) der Beitrag des Akteurs zur Datengenerierung einen besonderen **Aufwand** erfordert hat oder **besonders einzigartig** ist und aus wirtschaftlicher Perspektive nur schwer durch Beiträge anderer Akteure ersetzbar wäre; und

b) mit Hilfe der Daten eine ganz außergewöhnlich **hohe Wertschöpfung** betrieben wird; und

c) es dem Akteur aufgrund der Umstände, unter denen der Beitrag zur Datengenerierung geleistet wurde, **nicht möglich oder nicht zumutbar** war, über eine **Vergütung** zu verhandeln.

Die Höhe einer solchen, nur ganz ausnahmsweise geschuldeten Vergütung muss angemessen sein und darf insbesondere nicht den prinzipiellen Anreiz, mit Hilfe von Daten Wertschöpfung zu betreiben, gefährden. Sie muss auch berücksichtigen, dass der die Wertschöpfung betreibende Akteur in der Regel wirtschaftliche Risiken eingegangen ist.

2.3 Kollektive Aspekte von Datenrechten und Datenpflichten

Zu klären ist, ob und gegebenenfalls inwieweit die Überlegungen zu Unterlassung, Zugang, Korrektur und wirtschaftlicher Teilhabe auch auf **Kollektive** im Sinne definierter Gruppen von Personen (z. B. indigene Völker im Fall der Verwendung ihrer genetischen Daten) übertragbar sind, d. h. ob auch Kollektiven im Zusammenhang mit der Nutzung „ihrer" Daten bestimmte Datenrechte zustehen können. So könnte beispielsweise erwogen werden, ob der Bevölkerung eines Staates oder der EU für die Nutzung von Daten, die von dieser Bevölkerung generiert wurden, aus ethischer Sicht ein Recht auf wirtschaftliche Teilhabe (etwa in der Form von Steuern oder Transferleistungen) zustehen kann. Nach Auffassung der DEK kann dies prinzipiell der Fall sein.

Beispiel 9
Ein Internetkonzern verdient Milliardensummen durch Nutzerdaten, die weltweit bei der Nutzung der Dienste des Konzerns anfallen. Obgleich jährlich auch ein Milliardengewinn mit Daten von Nutzern aus der EU erwirtschaftet wird, zahlt der Konzern in der EU so gut wie keine Steuern. Hier stellt sich die Frage, ob den Konzern aus ethischen Gründen eine Pflicht treffen sollte, die Allgemeinheit in der EU wirtschaftlich in der Form von Steuern an der Wertschöpfung teilhaben zu lassen. Dies berührt grundlegende Fragen der Verteilungs- und der Teilhabegerechtigkeit und damit einer gerechten Wirtschaftsordnung. Es können aber etwa auch Aspekte wie Marktmacht oder die besondere Einzigartigkeit von Beiträgen (z. B. Audiodaten in einer bestimmten Sprache zur Entwicklung neuer sprachgesteuerter Dienste) in die Beurteilung mit einfließen.

Die Berücksichtigung von Gruppen und Kollektiven ist aufgrund des **relationalen Charakters vieler Daten** allgemein wichtig. Dieser relationale Charakter kommt beispielsweise zum Vorschein, wenn zahlreiche digitale Dienste von Nutzern verlangen, Daten ihrer Kontakte oder „Freunde" preiszugeben. Dies wird bei den hier vertretenen Datenrechten und korrespondierenden Datenpflichten insofern mitberücksichtigt, als den „Freunden" sowohl eigene Rechte auf Unterlassung, Zugang usw. zustehen können als auch jedenfalls ihre Interessen bei der Abwägung über ein geltend gemachtes Datenrecht stets mit einzubeziehen sind (→ oben 2.1). Darüber hinaus können bestimmte Daten, zu deren Generierung ein Akteur beigetragen hat, aber auch indirekt **Aufschluss über andere Akteure** geben, welche als solche individuell nicht – auch nicht im weitesten Sinne – zur Generierung dieser Daten beigetragen haben. Dieser Punkt ist bei genetischen Daten besonders evident, betrifft aber auch andere Datenarten. Damit eng verwandt ist die Situation, dass individuelle Daten, selbst in aggregierter Form, Einflüsse mit möglicherweise negativen **Drittwirkungen** jenseits des datenliefernden Individuums entfalten können.

Beispiel 10
Eine Krankenversicherung setzt Anreize zum Gesundheitstracking durch das Angebot reduzierter Prämien: Die Vorteile derer, die Daten preisgeben, können sich in höheren Beiträgen für jene, die Daten nicht preisgeben wollen, niederschlagen, d. h. der Vorteil des einen ist dann der Nachteil des anderen.

Auch Fragen der **Repräsentativität** von Trainingsdaten für algorithmische Systeme können als Relationalitätsproblem gefasst werden: Die mangelnde Bezüglichkeit zwischen jenen, die Trainingsdaten geliefert haben, und jenen, auf die die trainierten Systeme angewendet werden, kann zu systematischen Verzerrungen und einer möglichen Diskriminierung führen (→ näher Teil F, 2.6).

Um dieser Herausforderung gerecht zu werden, müssen individualistische Ansätze von Datenrechten in Ethik, Recht und Technikgestaltung um **relationale Konzeptionen von Datenrechten** erweitert werden (vgl. auch die Diskussion um Group Privacy). Das bedeutet, dass ein Beitrag zur Generierung von Daten, der durch einen Angehörigen einer relevanten Gruppe geleistet wurde, unter Umständen auch anderen Angehörigen dieser Gruppe zuzurechnen ist, so dass diesen trotz Fehlens eines individuellen Beitrags zumindest aus ethischer Sicht eigene Rechte auf Unterlassung, Zugang usw. zustehen können.

3. Anforderungen an die Nutzung personenbezogener Daten

3.1 Personenbezogene Daten und Daten juristischer Personen

Personenbezogene Daten sind alle Informationen, die sich auf eine **identifizierte oder identifizierbare natürliche Person** beziehen. Als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind, identifiziert werden kann (Art. 2 Nr. 1 DSGVO).

Wenngleich im Folgenden nur personenbezogene Daten im rechtlichen Sinn betrachtet werden, möchte die DEK daran erinnern, dass der **Schutzbedarf von Unternehmen und juristischen Personen** nicht ganz in den Hintergrund treten darf. Durch die Vernetzung aller Maschinen, den Austausch von Daten zwischen den Fabrikkomponenten und die Speicherung aller Produktionsdaten in digitalen Zwillingen in Industrie-4.0-Anlagen ist auch die Gefährdungslage für juristische Personen nochmals erhöht: So kann zum Beispiel aus der Verknüpfung von Einzeldaten (z. B. aus dem Wirkbetrieb von Geräten) ein praktisch lückenloses Bild interner Betriebsabläufe entstehen, das durch fehlende Schutzmechanismen in die Hände von betriebsfremden Akteuren (Konkurrenten, Verhandlungspartnern, Behörden, Übernahmeinteressenten usw.) gelangen kann. Dies stellt nach Auffassung der DEK eine ethisch bedenkliche Gefährdung der digitalen Selbstbestimmung von Unternehmen und juristischen Personen sowie auch – da zu einem großen Teil Datenflüsse zu Drittstaaten stattfinden – für die **digitale Souveränität Deutschlands und Europas** dar, der entgegenzuwirken ist.

Ein wichtiger juristischer Ansatzpunkt für den datenbezogenen Schutz von Unternehmen ist der **Schutz von Geschäftsgeheimnissen**, insbesondere durch das Gesetz zum Schutz von Geschäftsgeheimnissen (GeschGehG). Bei dessen Auslegung und Anwendung muss sichergestellt werden, dass der Schutz sensibler Unternehmensdaten angesichts seiner zentralen Bedeutung für eine faire und wettbewerbliche Wirtschaftsordnung und das darauf fußende wirtschaftliche und gesellschaftliche Wohlergehen vollumfänglich gewahrt bleibt. Allerdings ist die dem GeschGehG zugrundeliegende Richtlinie 2016/943 in verschiedener Hinsicht nicht ausreichend auf die Realität von IoT und Industrie 4.0 zugeschnitten. Die DEK fordert die Bundesregierung daher auf, den **datenbezogenen Schutz deutscher und europäischer Unternehmen zu verbessern**.

Die von der DEK im Folgenden für personenbezogene Daten unterbreiteten Handlungsempfehlungen – etwa in Bezug auf die risikoadäquate Auslegung des geltenden Rechtsrahmens (→ unten) oder auf datenschutzfreundliches Design von Produkten und Dienstleistungen (→ unten) – gelten dabei in sachgerechter Abwandlung bzw. Abschwächung auch für den Schutz von Daten, die sich auf Unternehmen und juristische Personen beziehen.

3.2 Digitale Selbstbestimmung als Aufgabe für die gesamte Rechtsordnung

3.2.1 Kooperatives Verhältnis zwischen den geltenden Rechtsregimen

Die Nutzung personenbezogener Daten in den verschiedensten Kontexten ist einerseits eine unverzichtbare Grundlage unserer Wirtschaft und Gesellschaft. Sie steht aber andererseits stets in einem Spannungsverhältnis zu individuellen Grundrechten. Beim Grundrecht auf informationelle Selbstbestimmung als Teil des allgemeinen Persönlichkeitsrechts handelt es sich im Kern um den Schutz der Menschenwürde. Das **Datenschutzrecht**, insbesondere die DSGVO, konkretisiert diese Maßstäbe und bindet öffentliche und private Stellen.

Die DSGVO stellt eine der großen Errungenschaften europäischer Rechtsetzung dar, welche derzeit weiteren Ländern als Inspirationsquelle dient. Allerdings dürfen die Erwartungen an diesen Rechtsakt nicht überspannt werden. Die DSGVO ist auf Datenschutz fokussiert, nicht aber auf die umfassende Wohlfahrt des Einzelnen und der Allgemeinheit in der Datengesellschaft. Sie ist auch für sich betrachtet nicht geeignet, alle Schäden, welche der Einzelne durch Verarbeitung seiner personenbezogenen Daten erleiden könnte, abzuwenden und in diesem Sinne umfassenden Integritätsschutz zu gewährleisten. Insbesondere soweit es um den Schutz von Rechtsgütern und Interessen geht, die **vom Datenschutzrecht nicht speziell adressiert** werden (z. B. Vermögensinteressen, Leben und körperliche Unversehrtheit, psychische Integrität, Ehre), bleibt die gesamte Rechtsordnung berufen. Dies gilt auch dann, wenn personenbezogene Daten im Spiel sind.

Die **datenschutzrechtliche Einwilligung** stellt einen zentralen Mechanismus zur Gewährleistung informationeller Selbstbestimmung im digitalen und analogen Bereich dar. Allerdings ist der Rechtsordnung ein inhaltlich schrankenloses Selbstbestimmungsrecht – einschließlich der Freiheit zur beliebigen Selbst- oder Fremdschädigung – nicht bekannt und auch ethisch nicht vertretbar. Eine freiwillige und informierte Einwilligung des Einzelnen als Ausdruck seiner grundrechtlich geschützten Handlungsfreiheit sollte zwar nur in eng begrenzten Ausnahmefällen von der Rechtsordnung eingeschränkt oder gar verboten werden. Ebenso wie beim Abschluss von Verträgen oder bei der Einwilligung in Eingriffe in die körperliche Unversehrtheit sind aber auch bei der datenschutzrechtlichen Einwilligung materielle Schranken anzuerkennen.

Nach Auffassung der DEK hat sich gezeigt, dass der Einzelne durch Anzahl und Komplexität der ihm abverlangten Entscheidungen bezüglich einer datenschutzrechtlichen Einwilligung ebenso wie durch die Unabschätzbarkeit aller Auswirkungen einer Datenverarbeitung **systematisch überfordert** wird. Die DEK sieht in einem unsachgemäßen Umgang mit dem Rechtsinstitut der Einwilligung seitens der Anbieter digitaler Dienste eine von mehreren Ursachen eines **Vertrauensverlusts** in der digitalen Gesellschaft. So kann der Einzelne derzeit vielfach nicht mehr darauf vertrauen, dass Staat und Rechtsordnung Rahmenbedingungen schaffen, in denen er sich sicher und relativ sorglos bewegen kann, ohne die Zufügung massiver Schäden durch Dritte befürchten zu müssen. Ebenso wie im Vertragsrecht zwischen Unternehmern und Verbrauchern die Inhaltskontrolle Allgemeiner Geschäftsbedingungen eine Art der „rationalen Gleichgültigkeit“ ermöglicht und Verbraucher selbst bei Bagatelltransaktionen umfassend schützt, gilt es, den gleichen Zustand durch **Inhaltskontrolle von Einwilligungserklärungen** zu erreichen[5]. Bei dieser Inhaltskontrolle sind prinzipiell die Wertungen der gesamten Rechtsordnung zu berücksichtigen.

3.2.2 Risikoadäquate Auslegung des geltenden Rechtsrahmens

Die DEK weist nachdrücklich darauf hin, dass der geltende Rechtsrahmen infolge der neuartigen Gefährdungslagen durch die umfassende Sammlung, Nutzung und Auswertung von personenbezogenen Daten in einer Weise ausgelegt und angewendet werden muss, dass diesen Gefährdungslagen bereits durch das geltende Recht soweit als möglich Rechnung getragen werden kann.

Unabhängig von der Erfüllung datenschutzrechtlicher Anforderungen existieren eine Reihe **absoluter Grenzen**, die eine Datenverarbeitung nicht überschreiten darf. Datennutzungen jenseits dieser Grenzen gilt es nach Möglichkeit bereits durch grundrechtskonforme Auslegung und Anwendung des geltenden Rechts[6] zu unterbinden. Dies betrifft nach Auffassung der DEK beispielsweise:

5 Vgl. auch Erwägungsgrund 42 zur DSGVO.
6 Infrage kommen insbesondere die Kontrolle von Allgemeinen Geschäftsbedingungen (§§ 307 ff BGB), die Grundsätze über Sittenwidrigkeit (§ 138 BGB) und sittenwidrige vorsätzliche Schädigung (§ 826 BGB) sowie vertragliche und vertragsähnliche Schutz- und Treuepflichten (§ 241 Abs. 2 BGB).

- Mit den Grundrechten unvereinbare **Eingriffe in den Kern der Privatsphäre und die Integrität der Persönlichkeit** durch Profiling und/oder Scoring (etwa bestimmte Formen der Ermittlung von Persönlichkeitszügen, Emotionen oder zu erwartenden Verhaltensweisen);

- Das Bewirken einer mit der Menschenwürde unvereinbaren **Totalüberwachung**, auch im Wege einer „Überwachungs-Gesamtrechnung" oder der Erstellung eines „Super-Scores";

- **Sittenwidrige Ausnutzungen** besonderer Notlagen oder eines pathologischen Gesundheitszustands;

- Dem Demokratieprinzip zuwiderlaufende **Beeinflussungen politischer Wahlen.**

Ethisch verwerfliche Irreführung oder Manipulation im geschäftlichen Verkehr – wozu auch Geschäftspraktiken gehören sollten, die auf die Hergabe von personenbezogenen Daten abzielen – sind bereits nach geltendem Recht und unabhängig von einem datenschutzrechtlichen Rechtsbruch als **irreführendes oder aggressives Verhalten** nach dem Gesetz gegen den unlauteren Wettbewerb (UWG) einzustufen und lösen entsprechende Rechtsfolgen aus (z. B. Anfechtung wegen Täuschung oder Drohung; Unterlassung und Schadensersatz). Dazu können nach Auffassung der DEK etwa auch gehören:

- Sog. **Addictive Designs,** d. h. technologische Gestaltungen, die geeignet sind, die Verhaltensfreiheit des Nutzers in Bezug auf die Nutzung (und das Beenden der Nutzung) durch unzulässige Beeinflussung wesentlich zu beeinträchtigen, vor allem durch Mechanismen, die ein Suchtverhalten verursachen;

- Sog. **Dark Patterns,** d. h. technologische Gestaltungen primär von Benutzungsschnittstellen, die geeignet sind, einen Nutzer über bestimmte Punkte zu täuschen und/oder ihn manipulativ zu veranlassen, eine bestimmte – möglicherweise auch wirtschaftlich relevante – Entscheidung zu treffen.

Absolute Grenzen der Datenverarbeitung sind auch zum Schutz vor **unangemessener vermögensmäßiger Benachteiligung** gezogen. Diesem Schutz dienen verschiedene Mechanismen des geltenden Rechts[7]. Beispiele für missbräuchliche Vertragsbedingungen bzw. Verletzung von Schutz- und Treuepflichten wären nach Auffassung der DEK etwa:

- Verwehrung oder unangemessene Erschwerung des Zugangs zu Gerätedaten, die für die übliche **Nutzung** eines Geräts einschließlich Reparatur durch eine unabhängige Werkstatt erforderlich sind (z. B. Gewährung nur nach Maßgabe von Art. 12 DSGVO, z. B. nur innerhalb eines Monats bzw. sogar von drei Monaten);

- Verwehrung oder unangemessene Erschwerung eines betriebsnotwendigen Datenzugangs für den **Zweiterwerber** einer vernetzten Sache (z. B. bei Verkauf einer mit Smart-Home-Technologie ausgestatteten Immobilie);

- Erschwerung des Anbieterwechsels durch sog. **Lock-in** veredelter Daten (z. B. Verweigerung der Herausgabe von Datenanalysen, für die der Nutzer wirtschaftlich betrachtet bereits bezahlt hat, und die nicht dem Schutz von Betriebs- und Geschäftsgeheimnissen des Unternehmers unterliegen);

- Verarbeitung von Nutzerdaten durch den Produzenten oder ein anderes Glied der Lieferkette zu einem Zweck, der den **wirtschaftlichen Interessen** des Nutzers signifikant zuwiderläuft (z. B. zum Zweck der Preisdifferenzierung, wenn damit die Abschöpfung der maximalen individuellen Zahlungsbereitschaft intendiert wird).

7 Vgl. die in Fn. 6 genannten Instrumente.

Social Media Monitoring

Social Media Monitoring ist die systematische **Beobachtung** der Inhalte sozialer Medien zu einem bestimmten Thema. Es hat sich zu einem Instrument der Datenverwertung entwickelt und macht sich dabei den Umstand zunutze, dass soziale Netzwerke die Kommunikationsmöglichkeiten der Nutzer erweitern und das digitale Verhalten zugleich einer konstanten Beobachtung aussetzen.

Unternehmen bedienen sich der nutzergenerierten Daten Sozialer Netzwerke häufig, z. B. zu Zwecken der Marktforschung und des Marketings. Öffentliche Stellen machen von Social Media Monitoring bislang seltener Gebrauch, aber auch sie nutzen es – beispielsweise durchsucht die Finanzverwaltung mittels eines Webcrawlers öffentlich verfügbare Inhalte im Internet, um gezielt nach gewerblichen Verkäufern zu suchen, die keine Umsatzsteuer abführen.

Mittels Social Media Monitoring zusammengetragene Informationen können über algorithmische Systeme einer weiteren, eingriffsintensiveren **Nutzung und Verwertung** (insb. der Erstellung von Persönlichkeitsprofilen zu kommerziellen Zwecken) zugeführt werden. Dies erfolgt dann rechtmäßig, wenn die Interessensabwägung nach Art. 6 Abs. 1 lit. f DSGVO positiv ausfällt oder ein sonstiger Rechtfertigungsgrund vorliegt. Die Tatsache, dass der Betroffene die Daten selbst öffentlich gemacht hat, rechtfertigt ausweislich Erwägungsgrund 51 DSGVO für sich allein eine weitere Nutzung und Verwertung noch nicht.

Die Grenze der Rechtmäßigkeit der Beobachtung ist nach Ansicht der DEK jedenfalls da erreicht, wo das Monitoring auf öffentliche Informationen erstreckt wird, deren Reichweite die betroffene Person bei Öffentlichmachung nicht einschätzen konnte (z. B. unbedachte Äußerungen von Kindern insgesamt) oder deren Sensibilität zu hoch ist (z. B. Äußerungen über Suizidabsichten). Auch sollten Daten von Bewerbern selbst dann, wenn der Betroffene sie selbst öffentlich gemacht hat, bei der Einstellung nicht verwendet werden, wenn sie zu tief in die persönliche Integrität eingreifen oder wenn sie sich nicht ihrem klaren Schwerpunkt nach auf die berufliche Vergangenheit beziehen (z. B. Äußerungen zur sexuellen Orientierung). Das Gleiche gilt für eine sonstige systematische Auswertung von Daten aus dem Privatleben (z. B. Tracking Daten).

Gerade für die weitergehende, eingriffsintensivere Nutzung und Verwertung können sich im Rahmen der Interessenabwägung Grenzen der Zulässigkeit ergeben (z. B. zielgerichtete gewerbliche Ansprache auf Basis der sexuellen Ausrichtung oder Ausnutzung emotional labiler Situationen). Insbesondere Anbieter Sozialer Netzwerke sind technisch in der Lage, Kommunikationsvorgänge, die über zentrale Plattformen laufen, im Detail auszuwerten; selbst wenn die Inhalte durch eine Ende-zu-Ende-Verschlüsselung einem allgemeinen Zugriff entzogen sein mögen, sind ihnen weitgehende Analysen auf Basis von Metadaten möglich. Das Auswerten von Kommunikation zwischen Individuen oder in geschlossenen Gruppen sollte rechtlich auch privaten Anbietern in Anlehnung an das Fernmeldegeheimnis untersagt sein. Insoweit empfiehlt die DEK der Bundesregierung, darauf hinzuwirken, dass diese Verbote im Rahmen der anstehenden Verabschiedung der e-Privacy-Verordnung kurzfristig umgesetzt werden.

3.2.3 Bedarf nach Konkretisierung und Verschärfung des geltenden Rechtsrahmens

Ein den Vorgaben der Verfassung genügender Rechtsgüterschutz wird in vielen Fragen der digitalen Gesellschaft zurzeit allenfalls mittels Auslegung und rechtsfortbildender Konkretisierung von unbestimmten Rechtsbegriffen und Generalklauseln durch Aufsichtsbehörden und Gerichte im Einzelfall bewirkt. Diese Situation ist nach Auffassung der DEK unangemessen. Zwar haben unbestimmte Rechtsbegriffe und Generalklauseln den Vorteil der Flexibilität und Zukunftsoffenheit. Dennoch dauert ihre Konkretisierung für neue und insbesondere digitale Zusammenhänge durch die behördliche und gerichtliche Praxis häufig Jahre bis Jahrzehnte, so dass in der Zwischenzeit sowohl ein **strukturelles Vollzugsdefizit** des geltenden Rechts als auch ein **Mangel an Rechtssicherheit** zu verzeichnen ist. Auf Grund der besonderen Grundrechtssensibilität und der Ungewissheit, ob und in welchem Zeitraum sich eine den verfassungsrechtlichen Anforderungen genügende Rechtspraxis entwickeln wird, sieht die DEK es als zentrale Aufgabe des demokratisch legitimierten Gesetzgebers an, den Ordnungsrahmen zeitnah verbindlich festzuschreiben.

Angesichts der Risiken, die für den Einzelnen in kritischen Bereichen durch **persönlichkeitssensible Profilbildungen** (sog. **Profiling**, ggf. mit der Folge von **Scoring**) erwachsen, hält die DEK insbesondere in diesen Bereichen eine Verschärfung des geltenden Rechtsrahmens für dringend geboten, um den Gefahren der Manipulation und der Diskriminierung des Einzelnen wirkungsvoll begegnen zu können.

Profilbildung

„Profiling“ ist in **Art. 4 Nr. 4 DSGVO** als jede Art der automatisierten Verarbeitung personenbezogener Daten definiert, die darin besteht, dass diese personenbezogenen Daten verwendet werden, um bestimmte persönliche Aspekte, die sich auf eine natürliche Person beziehen, zu bewerten, insbesondere um Aspekte bezüglich Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel dieser natürlichen Person zu analysieren oder vorherzusagen.

Profilbildungen stellen letztlich **Ableitungen** (Schlussfolgerungen) auf der Grundlage bestimmter Ausgangsdaten dar, die sich v. a. bestimmter Methoden des statistischen Schließens bedienen (→ Teil C, 2.2.2). Diese Ableitungen können wirkliche oder vermeintliche „Eigenschaften“ eines Einzelnen betreffen (z. B. „psychische Stabilität“, „Vertrauenswürdigkeit“, „Sozialverträglichkeit“) und/oder prognostischer Natur sein, wenn sie das künftige Verhalten eines Einzelnen zum Gegenstand haben (z. B. ein bestimmtes Konsumverhalten).

Neben der Profilbildung wird auch häufig versucht, aus dem beobachteten Verhalten eines Benutzers in der Interaktion mit digitalen Systemen diesen mit der Hilfe sog. Matching-Algorithmen einem vordefinierten **Stereotyp-Schema** zuzuordnen (z. B. bei Reisebuchung: Sportfan, Kulturreisender, Familienmensch, Wandergast, Vertreter, Gourmet). Mit dem für einen individuellen Benutzer instanziierten Stereotyp sind typische Vorlieben, Ziele und Persönlichkeitsmerkmale gespeichert, die dann in die weitere algorithmische Verarbeitung eingehen.

Es werden nicht immer die Profile als solche gespeichert, sondern es werden **ad-hoc-Ableitungen** (insbesondere Verhaltensvorhersagen) dynamisch und in Echtzeit (z. B. „ist jetzt kaufbereit für Schuhe“) aus Rohdaten generiert.

Profilbildungen schlechthin zu verbieten, schösse angesichts des Umstandes, dass der durch Profile ermöglichte Grad an Personalisierung zahlreicher digitaler Angebote von vielen Nutzerinnen und Nutzern als komfortabel und hilfreich empfunden wird, über das Ziel hinaus. Die DEK empfiehlt der Bundesregierung jedoch, sich beispielsweise im Rahmen der anstehenden Evaluierung der DSGVO dafür einzusetzen, die **DSGVO um spezifische Regelungen zu Profiling-Verfahren zu ergänzen**, die über die bereits bestehende Regelung des Art. 22 DSGVO zur Zulässigkeit automatisierter Entscheidungen hinausgehen, oder sich sogar für einen eigenen europäischen Rechtsakt einzusetzen, der den Gefahren durch Profilbildungen für die Grundrechte Einzelner effektiv begegnet. Falls sich eine hinreichend wirkungsvolle europäische Lösung in absehbarer Zeit als nicht realistisch erweisen sollte, sollte im Rahmen des europarechtlich Zulässigen eine nationale gesetzliche Regelung für den Umgang mit grundrechtsgefährdenden Profilbildungen angedacht werden.

In Bezug auf Profilbildungen sollten aus Sicht der DEK insbesondere folgende Aspekte eine (horizontale und/oder sektorale) gesetzliche Regelung erfahren, sofern sie von der DSGVO bei zutreffender Auslegung nicht ohnehin bereits vorgegeben sind:

a) Normierung **absoluter Grenzen** in der Form von gesetzlichen Verboten bestimmter **kritischer Einsatzzwecke** (z. B. Verwendung von Profilen, die aus Daten aus dem Privatleben gewonnen wurden, bei der Bewerberauswahl) und von Profilbildungen bei **besonders sensiblen personenbezogenen Daten**, etwa in Zusammenhang mit Emotionserkennungssoftware und biometrischen Daten, und bei Datenverarbeitungen mit **unvertretbarem Risikopotenzial** für die betroffenen Personen oder für die Gesellschaft;

b) Normierung von **Zulässigkeitsvoraussetzungen** für kritische Profilbildungen, einschließlich Qualitätsanforderungen hinsichtlich Aussagekraft und Treffsicherheit der gebildeten Profile (→ näher hierzu Teil F, 4.2.1), und einem risikoadäquaten System von Einwilligungslösungen (sog. Opt-in) und Widerspruchslösungen (sog. Opt-out), wobei letztere nur bei sehr geringem Risiko infrage kommen;

c) Konkretisierung des **Verhältnismäßigkeitsgrundsatzes** u.a. bezüglich der Anforderungen an die Art und den Umfang der zur Profilbildung herangezogenen Daten, der zulässigen Tiefe der zu Zwecken einer Profilbildung erfolgenden Schlussfolgerungen, und vor allem der Zwecke, für die Profilbildungen zulässigerweise eingesetzt werden dürfen;

d) Spezifische Kennzeichnungs-, **Informations- und Auskunftspflichten** bezüglich der Profilbildungen als solcher – und zwar einschließlich bezüglich der Existenz und des Zweckes von algorithmischen Systemen, die für **ad-hoc-Ableitungen** geeignet sind, sowie der bereits erfolgten, kritischen Ableitungen –, nicht erst der im Anschluss erfolgenden automatisierten Entscheidung;

e) Praktikable **Einwirkungsmöglichkeiten** einer betroffenen Person auf die über sie gebildeten Profile einschließlich der Möglichkeit zur Löschung/Korrektur/Überprüfung; dazu gehört auch das Recht auf einen „digitalen Neuanfang“ durch Löschung der gebildeten Profile, z. B. mit Erreichen der Volljährigkeit, wie es eine EU-Expertengruppe jüngst vorgeschlagen hat.[8]

8 High-Level Expert Group on Artificial Intelligence: Policy and Investment Recommendations for Trustworthy AI, 26.06.2019, S. 14, 40 (abrufbar unter: *https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60343*).

Sprachassistenten

Sprachassistenten bieten große Chancen im Hinblick auf Komfort und – insbesondere für Menschen mit Einschränkungen – den erleichterten Zugang zu digitalen Techniken. Sie bergen aber auch Gefahren für die Selbstbestimmung der Betroffenen.

Sprachassistenten erheben, oftmals auch ohne explizite Aktivierung, ihre Umgebungsgeräusche. Die dabei erhobenen Sprachaufnahmen der Nutzer sowie Dritter sind **biometrische Daten** im Sinne der DSGVO. Neben der Echtzeitanalyse der Sprachaufnahmen zur Reaktion auf die eingesprochene Aufforderung findet regelmäßig eine automatische Protokollierung gewisser Daten in einer Log-Datei statt (sog. Logging). Die analysierte personenspezifische Stimmfärbung sowie das Sprachmuster lassen sich verwenden, um die jeweilige Person **eindeutig zu identifizieren** oder **Sprachemotionen** zu analysieren. Ein solches Profiling greift besonders tief und invasiv in den Kernbereich der Persönlichkeitsrechte ein und droht die strukturelle Ungleichheit zwischen Angebots- und Nachfrageseite im Markt weiter zu vergrößern. Die Möglichkeit, das gesprochene Wort neu zu kombinieren bzw. digital nachzuformen (sog. deep fakes), eröffnet weiteres hohes **Missbrauchspotenzial**.

Faktisch verschwimmt für den individuellen Nutzer neben der Kenntnis des „Wie“ das Wissen um das „Ob“ der Datenverarbeitung Die Verwendung einer authentisch **menschlich klingenden Stimme** kann, insbesondere bei technisch unerfahrenen Menschen, zudem zu einer weitergehenden Preisgabe persönlichkeitssensibler Daten führen. Überdies zeichnen Sprachassistenten oftmals nicht nur lokal auf, sondern vernetzen sich über einen virtuellen Assistenten als Schaltzentrale und Herzstück moderner Wohnräume, vermehrt mit anderen Smart-Home-Produkten.

Die DEK sieht die umfassende Profilbildungsgefahr, die im Zusammenhang mit Sprachassistenten von der Zusammenführung verschiedenster Soft- und Hardwarekomponenten ausgeht, kritisch. Zudem können die Einfachheit, der Komfort sowie augenscheinliche Vorzüge der Verknüpfung mit weiteren Geräten den Nutzer letztlich in eine „Plug & Play-Falle“ tappen lassen. Geeignete Maßnahmen zur Reduzierung der Risiken, die von Sprachassistenten ausgehen, wären neben den Verboten besonders kritischer Profilbildungen und Einsatzzwecke nach Auffassung der DEK etwa:

a) Bindende technische Vorgaben zur Implementierung von Datenschutz „by design“ und „by default“ (→ siehe auch unten), insbesondere grundsätzlich **rein lokale Verarbeitung von Sprachdateien** (und Löschbarkeit) und Beschränkung einer Datenweiterleitung an den Betreiber oder Dritte auf bereits in Maschinensprache übersetzte Befehle (z. B. eine Bestellung);

b) Bindende technische Vorgaben zur **Abschaltbarkeit** von Mikrofon und Internetverbindung sowie **Sichtbarmachung**, ob das Mikrofon an- oder ausgeschaltet ist (→ siehe ebenfalls unten);

c) Dem Medium angemessene Ausgestaltung von **Transparenzpflichten** (→ siehe Teil F, 4.1), indem die wichtigsten Offenlegungen in der jeweiligen Situation oder in regelmäßigen Abständen auch **akustisch** erfolgen.

Jenseits derartiger spezialgesetzlicher Schutzmaßnahmen sollte die Bundesregierung prüfen, inwieweit losgelöst von den Zielen des Datenschutzrechts – und damit außerhalb des Anwendungsbereichs der DSGVO – vorrangig auf europäischer, sonst auf nationaler Ebene auf weitere Regelungen hingewirkt werden sollte, um den notwendigen Ordnungsrahmen für einen angemessenen Umgang mit Daten zu schaffen bzw. abzurunden. Im Zuge dessen empfiehlt die DEK insbesondere (→ zu Beispielen jeweils oben):

a) Ausdrückliche gesetzliche Normierung von datenspezifischen **Klauselverboten für die AGB-Kontrolle** (§§ 308, 309 BGB) und datenspezifischen **Schutz- und Treuepflichten** (§ 241 Abs. 2 BGB);

b) Ausdrückliche gesetzliche Normierung datenspezifischer **Deliktstatbestände** in Konkretisierung des Tatbestands sittenwidriger vorsätzlicher Schädigung (etwa in Gestalt eines neuen § 826a BGB);

c) Ausdrückliche gesetzliche Normierung datenspezifischer irreführender und aggressiver **Geschäftspraktiken**, wie z. B. Addictive Designs und Dark Patterns, durch Erweiterung der „Black List" des UWG; wegen der vollharmonisierenden Wirkung der europäischen Richtlinie über unlautere Geschäftspraktiken müsste diese Änderung allerdings zunächst auf europäischer Ebene ansetzen.

Erfolgt die Profilbildung durch **staatliche Stellen**, sind mögliche Effekte im Sinne eines kumulativen Grundrechtseingriffs bzw. einer Überwachungsgesamtrechnung ebenso zu berücksichtigen wie mögliche Nebenfolgen oder „Kollateralschäden". Besonderes Missbrauchspotenzial sieht die DEK in der Vernetzung einzelner Teilsysteme, wodurch Daten und Analyseerkenntnisse aus ganz unterschiedlichen Sach- und Lebensbereichen zusammengeführt werden. Dies führt zu einer erheblichen Verdichtung der Überwachung. Die Verknüpfung personenbezogener Informationen über verschiedene Überwachungssysteme hinweg und die Zusammenführung von Profilen wird dabei durch Techniken der intelligenten Mustererkennung (insbes. der Gesichtserkennung) erleichtert. Vor diesem Hintergrund empfiehlt die DEK zum einen, entsprechende Mustererkennung nur dort zu nutzen, wo dies für die Erfüllung staatlicher Aufgaben **unbedingt erforderlich** ist und zudem – über das nachrichtendienstliche Trennungsgebot hinaus – **klare gesetzliche Grenzen für den Austausch von Informationen** und Mustern zwischen den Behörden zu definieren. Dies kann auch die Neuregelung von Verwendungs- und Verwertungsverboten umfassen, insbesondere für den Austausch zwischen präventiv und repressiv tätigen staatlichen Stellen.

3.2.4 Bedarf nach einer Vereinheitlichung der Datenschutzaufsicht für den Markt

Die Datenschutzaufsicht über die Wirtschaft ist in Deutschland zwischen Bundes- und Landesbehörden verteilt. In Einzelfragen lassen sich Abweichungen in Aussagen zu datenschutzrechtlichen Anforderungen und eine divergierende Vollzugspraxis beobachten, die die betroffenen Akteure vor Herausforderungen stellt. Während im System der europäischen Mitgliedstaaten der Europäische Datenschutzausschuss (EDPB) als Institution für eine einheitliche Anwendung der DSGVO eingeführt wurde und im Einzelfall auch über Weisungsbefugnisse verfügt, erreicht das föderale Miteinander der Datenschutzbehörden der Bundesländer in Deutschland **bisher keine ähnliche Verbindlichkeit und Einheitlichkeit**.

Sofern sich die Abstimmung unter den deutschen Datenschutzaufsichtsbehörden nicht verstärken und formalisieren lässt und so die einheitliche und kohärente Anwendung des Datenschutzrechts gewährleistet werden kann, ist zu überlegen, die Datenschutzaufsicht im Markt durch **eine neue Behördenstruktur** zu vereinheitlichen. Eine solche Vereinheitlichung erlaubt den Aufbau spezialisierter Expertise, der für die Durchsetzung des Datenschutzrechts in einem technisch hochdynamischen Umfeld erforderlich ist. Dabei müsste gewährleistet sein, dass die einheitliche Behörde entweder selbst oder durch intensive Kooperation mit anderen Behörden auch die Durchsetzung **sonstiger datenrelevanter Rechtsmaterien**, die in engem funktionellem Zusammenhang mit dem Datenschutzrecht stehen (z. B. das Zivil- oder Lauterkeitsrecht), gewährleistet. Die Konzentration von Kompetenz für die Datenschutzaufsicht über den Markt in einer Stelle könnte ferner die Stimme Deutschlands im Europäischen Datenschutzausschuss – in dem alle Mitgliedstaaten bereits jetzt durch eine Datenschutzaufsichtsbehörde mit nationaler Zuständigkeit vertreten sind – weiter stärken. Schließlich sollte eine Zentralisierung der Behördenkompetenz mit der Konzentration der gerichtlichen Kontrolle der datenschutzaufsichtlichen Maßnahmen im Markt bei einem Gericht einhergehen, damit dieses gleichfalls eine entsprechende Expertise und eine kohärente Rechtsprechung entwickeln kann.

Organisationsrechtlich sind **verschiedene Modelle** denkbar. Im Zuge seiner Zuständigkeit zur Regelung des Rechts der Wirtschaft könnte der Bund die Kompetenz der Datenschutzaufsicht über die Wirtschaft (nicht-öffentlicher Bereich) auf den Bundesbeauftragten für den Datenschutz und die Informationsfreiheit übertragen und diesen entsprechend ausstatten. Dieser könnte durch verschiedene Außenstellen eine Präsenz der Datenschutzaufsicht in der Fläche garantieren (ähnlich dem Bundesamt für Migration und Flüchtlinge oder der Bundesbank). Denkbar ist auch die Bildung einer gemeinsamen Einrichtung der Länder qua Staatsvertrag nach den Modellen etwa im Rundfunkbereich oder der gemeinsamen Zentralstellen der Länder für Sicherheitstechnik und Gesundheitsschutz. Hier müsste die Unabhängigkeit der Datenschutzaufsicht durch die gemeinsame Einrichtung im Staatsvertrag gesichert werden. In jedem Fall ist – um eine angemessene Schlagkraft zu gewährleisten – auf eine bessere **personelle und sachliche Ausstattung** der Behörden zu achten.

Die Zuständigkeit der **Landesdatenschutzbehörden für den öffentlichen Bereich** sollte schon aus verfassungsrechtlichen Gründen in jedem Fall unangetastet bleiben.

3.3 Personenbezogene Daten als Vermögensgut

3.3.1 Ökonomisierung personenbezogener Daten

Personenbezogenen Daten kommt eine enorme wirtschaftliche Bedeutung zu. Der grundrechtliche Schutz der Persönlichkeit umfasst anerkanntermaßen auch die Entscheidung des Einzelnen, manche **Aspekte seiner Persönlichkeit gegen Entgelt zur Verfügung zu stellen** (z. B. Recht am eigenen Bild) und damit zu vermarkten.[9] Ebenso wie aber dem Einzelnen eine Vermarktung seiner Daten nicht vollkommen verwehrt ist, ist es auch nicht vollkommen ausgeschlossen, dass personenbezogene Daten auf Initiative Dritter hin wirtschaftlich verwertet werden. Der in diesem Zusammenhang teilweise bemühte Vergleich mit dem Handel menschlicher Organe hinkt in mehrfacher Hinsicht, u.a. weil Daten – anders als menschliche Organe – ein nicht-rivales Gut sind und die Tatsache, dass ein Anderer personenbezogene Daten verarbeitet, für sich betrachtet der betroffenen Person noch nicht unbedingt schadet; der Schaden wird erst durch einen bestimmten Kontext oder Zweck der Datenverarbeitung bewirkt.

Mit der Herleitung des informationellen Selbstbestimmungsrechts aus der Menschenwürde wird allerdings deutlich, dass der wirtschaftlichen Verwertung personenbezogener Daten dort **Grenzen** gezogen sind, wo auch ganz allgemein die Grenzen der Verarbeitung personenbezogener Daten verlaufen (→ oben 3.2.1 und), einschließlich der materiellen Grenzen der Einwilligung. Die wirtschaftliche Verwertung personenbezogener Daten ist in diesem Zusammenhang weder generell strengeren Regeln unterworfen noch generell privilegiert. Bei der Anwendung der allgemein geltenden Regeln müssen wirtschaftliche Aspekte allerdings in vielen Zusammenhängen berücksichtigt werden (z. B. hat wirtschaftlicher Druck Bedeutung für die Freiwilligkeit einer Einwilligung).

3.3.2 Daten als Eigentum und die Frage eines finanziellen Ausgleichs

Die DEK sieht derzeit **keine hinreichenden Gründe,** zusätzliche eigentumsähnliche Verwertungsrechte einzuführen, welche eine wirtschaftliche Partizipation an mithilfe von Daten generierten Gewinnen ermöglichen würden (oft unter dem Stichwort „**Dateneigentum**“ oder „Datenerzeugerrecht“ diskutiert).[10] Dem Einzelnen stehen bereits jetzt aufgrund des Datenschutzrechts oder des allgemeinen Zivilrechts genügend Rechtspositionen mit Drittwirkung zu, deren Einschränkung er theoretisch nur gegen Zahlung eines entsprechenden Entgelts dulden müsste. Wenn ihm die Aushandlung eines solchen Entgelts nicht gelingt, liegt das an Umständen (z. B. fehlende Verhandlungsmacht und/oder schlecht funktionierender Wettbewerb), die nichts mit dem Fehlen eines weiteren eigentumsähnlichen Verwertungsrechts zu tun haben.

Die Asymmetrie der Verhandlungsposition ließe sich allerdings theoretisch durch die Einführung von **Verwertungsgesellschaften**, die eigentumsähnliche Verwertungsrechte an Daten kollektiv wahrnehmen, ändern. Eine eigentumsähnliche wirtschaftliche Komponente personenbezogener Daten stünde allerdings in einem potenziellen **Spannungsverhältnis zum Datenschutz**, insbesondere zur Freiwilligkeit und jederzeitigen Widerruflichkeit der Einwilligung und zum Löschungsanspruch. Zudem würden **zweifelhafte finanzielle Anreize** zur Produktion möglichst vieler personenbezogener Daten geschaffen und würden gerade besonders vulnerable Personen (z. B. Minderjährige, einkommensschwache Bevölkerungsgruppen) zur Preisgabe möglichst vieler Daten animiert. Eine eventuelle Einpreisung der Vergütungen durch die Industrie könnte zudem zu einer verhältnismäßigen **Mehrbelastung datenschutzbewusster Personen** führen.

9 Siehe z. B. § 22 Gesetz betreffend das Urheberrecht an Werken der bildenden Künste und der Photographie (KunstUrhG).
10 Siehe anstelle vieler: Europäische Kommission: Aufbau einer europäischen Datenwirtschaft, 10.01.2017, COM(2017) 9 final (abrufbar unter: *https://ec.europa.eu/transparency/regdoc/rep/1/2017/DE/COM-2017-9-F1-DE-MAIN-PART-1.PDF*); Arbeitsgruppe „Digitaler Neustart“ der Konferenz der Justizministerinnen und Justizminister der Länder: Bericht vom 15. Mai 2017, S. 29 ff (abrufbar unter: *https://www.justiz.nrw.de/JM/schwerpunkte/digitaler_neustart/zt_bericht_arbeitsgruppe/bericht_ag_dig_neustart.pdf*).

Die genannten Argumente verfangen zwar nicht in gleichem Maße in Bezug auf anonymisierte Daten. Angesichts der Vielzahl von Akteuren, die einen Beitrag zur Generierung und Veredelung von Daten leisten, würde ein faires Vergütungssystem allerdings ein Maß an **Komplexität** erreichen und ein Ausmaß an Allzeitüberwachung zwecks Messung von Datenflüssen erfordern, das außer Verhältnis zu jedem möglichen Gerechtigkeitsgewinn stünde. Hinzu kämen mögliche negative Konsequenzen für die **Datenqualität**, da Anreize geschaffen würden, z. B. durch Anlegen falscher Geräteprofile „künstlich" Daten zu produzieren, die ein verzerrtes Bild der Realität liefern. Die DEK empfiehlt daher **auch bezüglich anonymisierter Daten keine Einführung von Verwertungsrechten**, die als Ausschließlichkeitsrechte ausgestaltet sind.

3.3.3 Daten als „Gegenleistung"

Eine Vielzahl digitaler Inhalte und Dienstleistungen (z. B. Suchmaschinen, soziale Netzwerke, Messenger-Dienste, Online-Spiele) werden Endnutzern ohne monetäre Gegenleistung angeboten. Die Finanzierung erfolgt auf andere Weise, insbesondere durch Leistungen Dritter für personalisierte Werbung und sonstige personalisierte Informationsangebote an die Nutzer sowie für deren Nutzerprofile und Nutzerscores. Dies hat zur plakativen Bezeichnung personenbezogener Daten als „Gegenleistung" für den digitalen Inhalt oder die Dienstleistung geführt, so etwa im ursprünglichen – im Gesetzgebungsverfahren jedoch wieder geänderten – Entwurf von Art. 3 Nr. 1 der Richtlinie über digitale Inhalte.[11] Inwieweit das beschriebene wirtschaftliche Modell überhaupt mit dem **Koppelungsverbot** aus Art. 7 Abs. 4 DSGVO vereinbar ist,[12] wird letztlich durch den EuGH zu klären sein.

Wenngleich die plakative Bezeichnung zur allgemeinen Bewusstseinsbildung beigetragen hat, plädiert die DEK dafür, **von der Bezeichnung von Daten als „Gegenleistung" abzusehen**. Zum einen sind personenbezogene Daten Teil der Persönlichkeit und genießen verfassungsrechtlichen Schutz. Zum anderen könnte die Einordnung als Gegenleistung nicht intendierte Implikationen nach sich ziehen. So könnte sie etwa als Argument dafür dienen, datenbezogene Allgemeine Geschäftsbedingungen nicht mehr in vollem Umfang der Inhaltskontrolle zu unterwerfen, oder dafür, dass ein Widerruf der Einwilligung, Löschungsverlangen usw. vertragliche Sanktionen gegen den Verbraucher auszulösen vermag.

In diesem Zusammenhang sollte der deutsche Gesetzgeber Freiräume bei der Umsetzung der Richtlinie (EU) 2019/770 über digitale Inhalte und digitale Dienstleistungen nicht in einer Weise nutzen, welche den Einzelnen von der Geltendmachung seiner datenschutzrechtlichen Rechtspositionen abhalten könnte. Insbesondere sollte der Anbieter im Fall des Widerrufs der Einwilligung zur Datenverwendung seine Leistung zwar mit sofortiger Wirkung einstellen dürfen, doch sollten **Zahlungsansprüche wegen einer bereits erbrachten Leistung ausgeschlossen** sein, ebenso wie ein nachträgliches **automatisches Zurückfallen in ein Bezahlmodell.**

Als Ausweg aus dem Koppelungsverbot werden verstärkt **Bezahlmodelle** diskutiert. Allerdings stellt jede noch so geringfügige finanzielle Belastung – insbesondere für vulnerable Bevölkerungsgruppen – einen Nachteil dar, der die Betroffenen abschrecken und zur übermäßigen Preisgabe ihrer personenbezogenen Daten bewegen kann. Auch ist eine überdurchschnittlich starke finanzielle Belastung besonders datenschutzbewusster Personen zu befürchten. Daher sollte vorrangig angestrebt werden, die **Finanzierung durch gewerbliche Nutzer**, die bislang unentgeltlich Gebrauch von bestimmten digitalen Inhalten oder Dienstleistungen machen, zu erreichen (z. B. die Seite eines Unternehmens bei einem sozialen Netzwerk).

11 Europäische Kommission: Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates über bestimmte vertragsrechtliche Aspekte der Bereitstellung digitaler Inhalte, 9.12.2015, COM(2015) 634 final (abrufbar unter: *https://ec.europa.eu/transparency/regdoc/rep/1/2015/DE/1-2015-634-DE-F1-1.PDF*).

12 Europäischer Datenschutzbeauftragter: Stellungnahme 4/2017 zu dem Vorschlag für eine Richtlinie über bestimmte vertragsrechtliche Aspekte der Bereitstellung digitaler Inhalte, 14. März 2017, S. 19 (abrufbar unter: *https://edps.europa.eu/sites/edp/files/publication/17-03-14_opinion_digital_content_de.pdf*).

Bezahlmodelle können allerdings auch das Bewusstsein von Verbrauchern für den monetären Wert der eigenen Daten stärken und Transparenz schaffen. Aus diesen Gründen kann nach Auffassung der DEK das **alternative Angebot eines Bezahlmodells** einen ethisch akzeptablen Ausgleich zur Herstellung der notwendigen Freiwilligkeit darstellen. Dabei ist jedoch zu beachten, dass der Preis nicht missbräuchlich und marktunüblich hoch sein darf, sondern eine auch aus Verbrauchersicht realistische Alternative zur Preisgabe personenbezogener Daten darstellen muss. Ferner ist aus ethischer Sicht sicherzustellen, dass es nicht zu einer Quersubventionierung durch datenschutzbewusste Nutzer kommt und dass die Bedürfnisse sozial schwacher Bevölkerungsgruppen etwa durch entsprechende staatliche Transferleistungen berücksichtigt werden.

3.3.4 Daten als Grundlage personalisierter Risikoeinschätzung

Bei der **personalisierten Risikoeinschätzung** (z. B. einmalig bei der Kreditvergabe oder laufend bei Versicherungen mit Telematiktarifen) geht es um höhere Granularität von preisrelevanten Vorhersagen durch Nutzung algorithmischer Systeme. Es handelt sich hierbei letztlich um einen sektorspezifischen Anwendungsfall einer bestimmten Form von Profilbildung und eines darauf aufbauenden Scoring (→ zu Profilbildungen allgemein bereits oben 3.2.3 und unten Teil F, 4.2.2). Die Verarbeitung zusätzlicher personenbezogener Daten bedarf bei der personalisierten Risikoeinschätzung regelmäßig der Einwilligung der betroffenen Personen. Diese werden zunächst jene Personen erteilen, welche sich dadurch ökonomische Vorteile erhoffen. Dabei kann die Einwilligung einer Person signifikante Auswirkungen auf andere Personen haben und ethisch unerwünschte Kettenreaktionen (sog. Unraveling-Effekte) auslösen. Dies kann dazu führen, dass der in die Datenverarbeitung Einwilligende unter unverhältnismäßigem Druck steht und die Freiwilligkeit der Einwilligung gefährdet wird.

Beispiel 11

Besonders gesunde Versicherte willigen in die Datenverarbeitung durch eine Krankenversicherung ein. Um nicht in den Verdacht zu geraten, zu den Versicherten mit schlechterer Gesundheit zu gehören, geraten andere unter Druck, ebenfalls einzuwilligen.

Sofern die Parameter durch das Verhalten des Einzelnen beeinflussbar sind, können derartige Modelle zudem erheblichen **Einfluss auf die private Lebensgestaltung** entfalten. Gerade im Versicherungssektor kommt aus ethischer Sicht noch der Aspekt hinzu, dass das Streben nach immer höherer Granularität der Risikoeinschätzung dem **Grundprinzip der kollektiven Risikoübernahme** durch die Gemeinschaft aller Versicherten zuwiderläuft. Im Extremfall „vollständigen" Wissens auf der Seite des Versicherers und entsprechender Anpassung des Preises an das individuelle Risiko hat sich der Gedanke einer Versicherung ad absurdum geführt.

Nach Ansicht der DEK stellen sich daher aus ethischer Sicht insbesondere folgende Anforderungen an eine personalisierte Risikoeinschätzung:

a) Die Datenverarbeitung darf **nicht den Kern privater Lebensführung** betreffen, sondern nur Bereiche, in denen der Einzelne ohnehin in Kontakt mit der Außenwelt tritt und damit rechnen muss, dass man Schlüsse aus seinem Verhalten zieht. Ethisch akzeptabel wäre danach bei einer Kfz-Versicherung etwa die Registrierung der gefahrenen Kilometer oder von Verstößen gegen die StVO, nicht dagegen des rein privaten, wenn auch möglicherweise risikorelevanten Verhaltens im Fahrzeug (z. B. Frequenz des Gähnens, Gespräche mit Beifahrern) oder gar des Gesundheitszustands (z. B. Herzschwäche) oder der sonstigen Lebensführung (z.B. Einkaufsverhalten betreffend Kaffee oder Alkohol);

b) Zwischen den verarbeiteten Daten und dem zu bestimmenden Risiko muss ein **klarer ursächlicher Zusammenhang** bestehen, und die Verknüpfung darf keine **Diskriminierung** darstellen (→ siehe dazu unten Teil F, 2.6);

c) Es darf sich nicht um Daten handeln, die unmittelbar Schlussfolgerungen mit **Wirkung für Angehörige oder sonstige Dritte** zulassen;

d) Es muss umfassende **Transparenz** bezüglich der Auswirkungen, die bestimmte Parameter und deren Gewichtung auf die Gestaltung des Preises oder der sonstigen Konditionen haben, gegeben sein, und der Einzelne muss klare und verständliche Erläuterungen erhalten, wie er die Konditionen verbessern kann (→ siehe dazu Teil F, 2.7);

e) Um unerwünschte Kettenreaktionen in Grenzen zu halten, darf die Differenz zwischen den „optimalen" Konditionen und den bei Verweigerung der Einwilligung zu erreichenden Konditionen ein Höchstmaß nicht überschreiten (z. B. **maximale Preisdifferenz**).

3.3.5 Daten als Reputationskapital

Personenbezogene Daten, Profile und Scores erhalten im Zusammenhang mit **personalisierten wirtschaftlichen Konditionen** (personalisierte Preise, personalisiertes Ranking, personalisierte Produkte und Dienstleistungen) eine Funktion als Reputationskapital. Bei der personalisierten Verhaltensprämierung zu Zwecken der **Kundenbindung** (z. B. durch Rabatte in Abhängigkeit von der Einkaufsmenge des Vormonats) werden zwar Anreize zur Einwilligung in die Verarbeitung personenbezogener Daten geschaffen und besteht eine Tendenz, die private Lebensführung zu beeinflussen. Der DEK liegen jedoch keine Hinweise vor, dass in Zusammenhang mit Kundenbindungsprogrammen die soeben (→ oben) dargestellten ethischen Grenzen in der deutschen Wirtschaft derzeit überschritten werden. Die Entwicklung sollte jedoch weiter beobachtet werden.

Bei der **klassischen Preisdifferenzierung** und Maßnahmen ähnlicher Wirkung sieht die DEK den Schwerpunkt der Problematik im Bereich der Regulierung algorithmischer Systeme (→ dazu im Detail Teil F). Zu einem Problem der Datennutzung wird Preisdifferenzierung allerdings dann, wenn Verbrauchern suggeriert wird, die Preise generell durch Preisgabe möglichst vieler personenbezogener Daten oder durch bestimmte, an die relevanten Kriterien angepasste Verhaltensweisen (etwa Online-Einkauf über einen Computer einer bestimmten Marke) senken zu können bzw. wenn Verbraucher, die die Einwilligung in die zur personalisierten Preissetzung erforderliche Datenverarbeitung verweigern, im **Durchschnitt stets höhere Preise** zahlen. Letzteres wäre nach Ansicht der DEK nicht zuletzt ein ethisch bedenklicher Angriff auf die Freiwilligkeit der Einwilligung.

Darüber hinaus erlangen **echte Reputationsdaten**, die auch für außenstehende Dritte sichtbar sind (z. B. durch „Sterne" indizierte Zuverlässigkeit als Vertragspartner im Rahmen einer Online-Plattform), immer größere wirtschaftliche und immaterielle Bedeutung. Derartige echte Reputationsdaten werden teilweise durch die neue **Verordnung (EU) 2019/1150 zur Fairness und Transparenz** für gewerbliche Nutzer von Online-Vermittlungsdiensten erfasst.[13] Dabei wurde ein behutsamer und weitgehend auf Transparenzanforderungen und Selbstregulierung aufbauender Regelungsansatz gewählt. Die DEK begrüßt im Grundsatz diesen behutsamen Ansatz. Sie weist jedoch darauf hin, dass insbesondere die Abhängigkeit einzelner Branchen von echten Reputationsdaten zu starken Lock-in-Effekten führen kann, die den Wettbewerb gefährden und problematisch sind, falls die Daten bei einem Wechsel des Online-Vermittlungsdienstes nicht mitgenommen werden können.

13 Vgl. deren Art. 9 zum Datenzugang sowie viele allgemeine Bestimmungen, etwa zu Allgemeinen Geschäftsbedingungen und Ranking.

Beispiel 12
Ein Kleinstunternehmer, der über eine Online-Plattform Fahrdienstleistungen anbietet und sich ein gutes Bewertungsprofil erworben hat, möchte die Plattform wechseln und sein Bewertungsprofil übernommen haben.

Die DEK sieht die Probleme, die mit einer allgemeinen gesetzlichen Verpflichtung zur Übernahme von Bewertungsprofilen verbunden wären. Sie empfiehlt der Bundesregierung jedoch, zu prüfen, unter welchen Bedingungen einem gewerblichen Nutzer doch ein Anspruch auf **Portabilität** seiner Bewertungsprofile zugesprochen werden kann, um auf europäischer Ebene auf eine weitergehende Regelung hinzuarbeiten.[14]

Die gesteigerte Bedeutung **sozialer Reputationsdaten** (Anzahl der „Likes", „Followers", „Freunde") sind demgegenüber Teil einer größeren Entwicklung unserer Gesellschaft, die – mit der begrenzten Ausnahme von sog. Influencern – nicht mehr primär unter dem Aspekt der „Ökonomisierung" personenbezogener Daten gesehen werden kann, sondern im Hinblick auf die systemischen gesellschaftlichen Auswirkungen zu diskutieren ist.

3.3.6 Daten als Handelsware

Zahlreiche Unternehmen erzielen mittlerweile zum Teil beträchtliche Gewinne dadurch, dass sie gesammelte personenbezogene Daten, Profile und Scores oder aus aggregierten Rohdaten vorgenommene statistische Auswertungen über einzelne Personen an Dritte weiterverkaufen oder bereits vorhandene Profile mit Schätzdaten weiter anreichern und diese dann in den Handel bringen. Derartige Geschäftsmodelle werden im Folgenden als **„Datenhandel"** bezeichnet.

Derzeit enthält die DSGVO keine spezifischen Regelungen zum Datenhandel. Sie qualifiziert derartige Geschäftsmodelle vielmehr schlicht als gewöhnliche Datenverarbeitungsprozesse, die den allgemeinen Regelungen der DSGVO unterliegen. Bei genauer Prüfung der geltenden Bestimmungen wird man oft zu dem Schluss gelangen müssen, dass Formen des Datenhandels gegen die Vorgaben der DSGVO verstoßen und daher rechtswidrig betrieben werden. Insgesamt besteht im Bereich des Datenhandels jedoch ein **erhebliches Vollzugsdefizit**. Die DEK würde es deshalb begrüßen, wenn die Datenschutzaufsicht in Bezug auf diese Branche mit besonderer Dringlichkeit tätig werden würde und der Europäische Datenschutzausschuss (EDSA), hilfsweise die Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (DSK), unter Konkretisierung des risikobasierten Ansatzes der DSGVO klar abgrenzbare Fallgruppen für verschiedene Formen des rechtmäßigen Datenhandels entwickeln würde. Dabei wäre klarzustellen, in welchen Fällen des Datenhandels es einer Einwilligung der betroffenen Person für die Weitergabe von Daten bedarf, in welchen Fällen nur ein Widerspruchsrecht besteht, und in welchen Fällen zwingende schutzwürdige Gründe sogar das Widerspruchsrecht ausschließen.

Über den Vollzug des bereits geltenden Datenschutzrechts hinaus sollte die Weitergabe von Daten an Dritte im Lichte der allgemeinen Prinzipien der Datenverarbeitung (Art. 5 DSGVO) nur in engen Grenzen zulässig sein. Daher empfiehlt die DEK der Bundesregierung, auf europäischer Ebene u.a. bei der anstehenden Evaluierung der DSGVO darauf hinzuwirken, dass die **DSGVO um datenhandelsspezifische Regelungen ergänzt** wird. Für die Ausgestaltung einer solchen künftigen Regelung sollten die folgenden **ethischen Gesichtspunkte,** die zum Teil schon in der DSGVO niedergelegt sind, berücksichtigt werden:

14 Vgl. etwa Artikel 6 und 7 des Entwurfs der „Model Rules on Online Intermediary Platforms" des European Law Institute, die der DEK zur Verfügung gestellt wurden.

a) Der Ausgangspunkt jeder Abwägung sollte in der informationellen Selbstbestimmung des Einzelnen liegen, sodass Datenhandel im Grundsatz der vorherigen **Einwilligung** der betroffenen Person unter Berücksichtigung der **materiellen Schranken** der Einwilligung (→ oben 3.2.1 und) bedarf;

b) Kann die Datenverarbeitung im Einzelfall auf eine andere Rechtsgrundlage als die Einwilligung gestützt werden, muss der Einzelne bereits vorab die Möglichkeit haben, ein **Widerspruchsrecht** auf einfache Weise auszuüben (z. B. Entfernen eines Häkchens unmittelbar vor Erhebung), und darf nicht erst auf gesonderte Kommunikationskanäle verwiesen werden;

c) Datenhandelsmodelle ohne jedwede **Wahlmöglichkeiten** des Betroffenen sollten nur sehr selten in Betracht kommen, und zwar lediglich dann, wenn und soweit die Weitergabe der Daten aufgrund eindeutig überwiegender öffentlicher Interessen des Gemeinwohls erforderlich ist. Diese Kategorie sollte vollständig durch den Gesetzgeber konkretisiert werden;

d) Die DSGVO enthält detaillierte Vorschriften zur Datenweitergabe an Auftragsverarbeiter und zur Datenweiterleitung in Drittstaaten. Zwar können im Lichte von Sinn und Zweck der DSGVO bei der Weitergabe an Dritte innerhalb des Gebiets der EU kaum niedrigere Anforderungen gelten und sollten diese Anforderungen etwa als „geeignete Garantien" in die allgemeinen Regelungen hineingelesen werden. Dennoch wäre dringend zu empfehlen, die Pflichten bei der Weitergabe von Daten an Dritte (z. B. Kontrollpflichten) ebenso wie diesbezügliche Haftungstatbestände ausdrücklich gesetzlich zu konkretisieren;

e) Verantwortliche sollten die konkrete Quelle, aus der ein Datum erhoben oder aus der es etwa durch automatisierte Schlussfolgerung generiert wurde, sowie die konkreten Einzelempfänger dokumentieren und offenlegen müssen, und zwar in einer standardisierten und maschinenlesbaren Form, welche die automatisierte Verwaltung etwa durch PMT/PIMS (→ dazu unten 4.3.) ermöglicht. Dadurch würde dem Umstand Rechnung getragen, dass Datenhändler in der **Wahrnehmung für Betroffene** bislang häufig im Verborgenen bleiben und dass eine bloße Benennung der Kategorien von Quellen oder Empfängern für den Betroffenen weitgehend nutzlos ist;

f) Aufgrund der Vielzahl von Datenhändlern können Betroffenenrechte nur dann effektiv geltend gemacht werden, wenn zentrale Mechanismen die **Geltendmachung** erleichtern oder übernehmen (z. B. die Datenschutzaufsicht, → dazu oben , oder PMT/PIMS, dazu unten 4.3);

g) Aufgrund des erhöhten Risikos und Kontrollverlusts in Folge von Streuungseffekten sollten Datenhändler einer **datenschutzrechtlichen Zertifizierungspflicht** unterworfen werden, die regelmäßige Auditierungen durch die Zertifizierungsstellen vorsieht. Die DEK empfiehlt, dass die unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder hierzu spezifische Zertifizierungskriterien aufstellen, die den von der DEK aufgezeigten Risiken und ihren Empfehlungen Rechnung tragen.

3.4 Daten und digitaler Nachlass

Moderne Kommunikationstechnologien und Kapazitäten der Datenverarbeitung ermöglichen eine nahezu lückenlose Aufzeichnung der privaten Aktivitäten eines Menschen über Jahrzehnte hinweg, ebenso wie deren automatisierte Auswertung. Gelangen die gesammelten Daten eines Menschen nach dessen Tod in die Hände der Erben oder eines sonstigen Dritten, bedeutet dies eine **neue Dimension von Gefährdung für die Privatheit**, und zwar sowohl für den Verstorbenen als auch vor allem für seine Kommunikationspartner zu deren Lebzeiten. Der oft angestellte Vergleich mit Tagebüchern und persönlichen Briefen hinkt, weil viele Kommunikationen (via Messenger, Chats, E-Mails usw.) funktional nicht an die Stelle von Briefen treten, sondern an die Stelle des flüchtig gesprochenen Wortes.

3.4.1 Vorrang von Verfügungen zu Lebzeiten

Die DEK sieht die vorrangig anzustrebende Lösung in bewussten und informierten Dispositionen des Betroffenen noch zu Lebzeiten. Vielfach unterbleiben solche Dispositionen allein aus Unsicherheit oder Unkenntnis der rechtlichen und tatsächlichen Möglichkeiten. Vor diesem Hintergrund hält es die DEK für gerechtfertigt, die **Diensteanbieter zu verpflichten**, Nutzer auf Dispositionsmöglichkeiten für den Fall der dauernden Einwilligungsunfähigkeit (z. B. infolge von Demenz) oder des Todes hinzuweisen sowie die technischen Möglichkeiten bereitzustellen, möglichst barrierefrei – d. h. ohne oder mit nur minimalem Medienwechsel – Dispositionen zu treffen. Dazu könnte das **Telemediengesetz (TMG) um eine entsprechende Vorschrift ergänzt** werden.[15]

Nach Auffassung der DEK sollte die – nur besonders zugespitzte – Situation beim Tod einer Person auch zum Anlass genommen werden, ganz allgemein über die Gestaltung digitaler Kommunikationsformen nachzudenken. Die DEK empfiehlt der Bundesregierung daher, eine Verpflichtung für Messenger-Dienste zu prüfen, die **standardmäßige Löschung** von Nachrichten nach einer bestimmten Frist als Option anzubieten. Entscheidet sich der Nutzer für diese Option, würde dann eine Nachricht nach Ablauf der Frist – so sie vom Empfänger oder Sender nicht manuell archiviert wurde – automatisch gelöscht.

3.4.2 Die Rolle von Intermediären

Die wachsende Sensibilität für das Thema hat auch neue Geschäftsmodelle hervorgebracht: Eine Vielzahl von Unternehmen bietet inzwischen Dienstleistungen rund um den digitalen Nachlass an (von der zentralen Verwahrung von Kontodaten und Passwörtern bis hin zur umfassenden Verwaltung des digitalen Nachlasses). Diese können sinnvolle Hilfestellungen sein. Sie sind aber zugleich auch mit Gefahren verbunden. Jene reichen von mangelnder Vorsorge für den Fall der Insolvenz oder sonstiger Auflösung des Unternehmens über Lücken in der Informationssicherheit bis hin zu echtem Betrug. Eine **Qualitätskontrolle** und vorsichtige **Regulierung** sowie die **Aufklärung** der Bevölkerung über mögliche Vorteile und Risiken erscheinen nach Auffassung der DEK zum Schutze der Bürger geboten.

15 Mario Martini: Juristenzeitung (JZ), 2012, S. 1145, 1154.

Die DEK empfiehlt darüber hinaus dem Staat, als Teil der Daseinsvorsorge für seine Bürger eine **zumindest staatlich beaufsichtigte Stelle** einzurichten, welche zu leistbaren Konditionen Basis-Dienstleistungen der digitalen Nachlasssicherung und Nachlassplanung auf dem aktuellen Stand der Informationssicherheitstechnik erbringt. Genau wie bei einem Testament eine Wahlmöglichkeit besteht, das Testament privat zu verwahren oder aber beim Notar oder Amtsgericht verwahren zu lassen, sollte eine vergleichbare Wahlmöglichkeit zwischen privater bzw. privatwirtschaftlicher Lösung und einer staatlichen Dienstleistung auch in Bezug auf den digitalen Nachlass bestehen.

3.4.3 Postmortaler Datenschutz

Die DEK empfiehlt keine prinzipielle Abkehr von den vom Bundesgerichtshof (BGH)[16] formulierten Grundsätzen eines **Übergangs auf die Erben**, da die unerwünschten und/oder überschießenden Wirkungen einer anderweitigen Default-Lösung (etwa eines gesetzlich angeordneten Treuhand-Modells oder einer Trennung vermögens- und persönlichkeitsbezogener Inhalte in Bezug auf ein und dasselbe Nutzerkonto) die möglichen Vorteile vielfach überwögen. Ist ein ganzes Nutzerkonto seiner Art nach ohne Vermögenswert, aber besonders persönlichkeitssensitiv (etwa ein Online-Konto in einer Gruppe „Anonymer Alkoholiker"), dürfte es jedoch vorzuziehen sein, es aufgrund des höchstpersönlichen Charakters ganz vom Erbrecht auszunehmen. Soweit – auch zum Schutz der Kommunikationspartner des Verstorbenen – das **Telekommunikationsgeheimnis** Platz greift, ist der Gesetzgeber ohnehin nach wie vor aufgerufen, die Normkollision mit dem grundrechtlich verbürgten Erbrecht aufzulösen, etwa durch einen entsprechenden Hinweis im Erbrechtsteil des BGB.

Der vom BGH formulierte Grundsatz des Übergangs auf die Erben ist an das Bestehen eines Vertragsverhältnisses gekoppelt. Soweit kein Vertragsverhältnis besteht oder wegen Höchstpersönlichkeit nicht auf die Erben übergeht, können diese nicht einschreiten. Da der **Schutz durch die DSGVO mit dem Tod erlischt**, stehen sodann, nach derzeitiger Gesetzeslage, auch keine datenschutzrechtlichen Eingriffsmöglichkeiten zur Verfügung, die Angehörige geltend machen könnten. Dass damit die personenbezogenen Daten Verstorbener in die nahezu unbegrenzte Verfügungsgewalt der jeweiligen Verantwortlichen übergehen, erscheint ethisch bedenklich. Die DEK empfiehlt der Bundesregierung daher, nach dem Vorbild mehrerer europäischer Staaten von der in Erwägungsgrund 27 zur DSGVO erwähnten Möglichkeit Gebrauch zu machen, Regelungen zum **postmortalen Datenschutz** zu erlassen. Dabei sollten Angehörige fundamentale Betroffenenrechte – etwa auf Löschung von Daten oder Korrektur unrichtiger Daten – auch nach dem Tod des Betroffenen geltend machen können. Zugleich wäre in geeigneter Weise sicherzustellen, dass Verfügungen, die der Verstorbene zu Lebzeiten getroffen hat – und wenn auch nur konkludent z. B. durch bewusste Öffentlich-Stellung seiner „Life Story" – zu respektieren sind.

16 Urteil des Bundesgerichtshofs vom 12. Juli 2018, Aktenzeichen III ZR 183/17.

3.5 Besondere Gruppen von Betroffenen

3.5.1 Beschäftigte

Durch die teilweise weitreichende Erfassung der Bewegungs- und Leistungsdaten der Arbeitnehmer in modernen Arbeitsumgebungen und durch die für bestimmte Kollaborationsformen notwendige Erstellung biometrischer Profile entstehen erhebliche **Gefahren für die informationelle Selbstbestimmung und das allgemeine Persönlichkeitsrecht** der Arbeitnehmer. Zu den zu bedenkenden Fragen gehören neben den Rechtsgrundlagen der Datenverarbeitung und der Mitbestimmung der Interessenvertretungen etwa: Anforderungen an eine Information der Beschäftigten (vgl. etwa Herausforderungen durch Multi-Sensor-Fusion) und je nach Kontext Schaffung von Widerspruchsmöglichkeiten; Einzelheiten zur Speicherung, Speicherdauer und zulässigen Offenlegung von Beschäftigtendaten gegenüber Dritten; Recht auf Korrektur falscher oder überholter Daten (etwa bei persönlichen Profilen) und angemessene Löschregelungen; Rahmenbedingungen für eine begrenzte Kontrolle und Überwachung von Beschäftigten; Begrenzung der Lokalisierung von Mitarbeitern und Ausschluss von umfassenden Bewegungsprofilen; Begrenzung von Verpflichtungen zum Teilen von Social Media Accounts und zum Datenzugriff des Arbeitgebers im Kontext von „Bring your Own Device"-Modellen; Rahmenbedingungen für den Einsatz von biometrischen Systemen; oder Begrenzung von psychologischen Untersuchungsmethoden.

Die DEK empfiehlt der Bundesregierung, die Sozialpartner einzuladen, ausgehend von den bereits in Tarifverträgen bestehenden Beispielen guter Übung eine gemeinsame Linie für gesetzliche Konkretisierungen des **Beschäftigtendatenschutzes** zu entwickeln. Dabei sollten auch die Belange von Personen in unüblichen Beschäftigungsformen berücksichtigt werden. Kollektivverträge und Betriebsvereinbarungen sollen auch weiterhin im Bereich des Beschäftigtendatenschutzes eine wichtige Rolle spielen. Schon wegen der gesteigerten Grundrechtsrelevanz sollten die zentralen Grundsätze des Beschäftigtendatenschutzes aber nicht ausschließlich an Kollektivverträge und Betriebsvereinbarungen überwiesen werden, zumal diese nicht alle Beschäftigten erfassen. Die gegenwärtig bestehende Rechtsunsicherheit über das Ausmaß, in dem Vorschriften der DSGVO anwendbar bleiben, erschwert überdies sichere Investitionen.

Die DEK hält die klassische datenschutzrechtliche **Einwilligung**, verglichen mit anderen Rechtsgrundlagen der Verarbeitung von Beschäftigtendaten, nicht in allen Kontexten für geeignet, da die notwendigen Rahmenbedingungen für die Freiwilligkeit der Einwilligung im Beschäftigungskontext schwierig zu erfüllen sind und die jederzeitige Widerruflichkeit und Löschungsverpflichtung nicht in allen Konstellationen mit den Bedürfnissen des Arbeitgebers in einen angemessenen Ausgleich gebracht werden kann. Der Fokus eines Beschäftigtendatenschutzes sollte daher auf spezifisch auf den Beschäftigungskontext zugeschnittene, **gesetzliche Rechtfertigungsgründe** gelegt werden, die ein hohes Maß an Schutz und einen angemessenen Grundrechtsausgleich gewährleisten. Diese können einwilligungsähnliche Elemente aufweisen, welche die typischerweise gegebenen Machtverhältnisse im Beschäftigungskontext berücksichtigen.

Bei der Ausgestaltung der **Mitbestimmungsrechte der Interessenvertretungen**[17] über die Verarbeitung personenbezogener Daten im Betrieb muss der bestehenden **Wissensasymmetrie** zwischen Arbeitgeber- und Arbeitnehmerseite über die Wirkungsweise und Details der Verarbeitungsvorgänge angemessen Rechnung getragen werden. Es müssen daher Modelle gefunden werden, die den Interessenvertretungen über die geltenden Mechanismen hinaus den Rückgriff auf externen Sachverstand ermöglichen, wobei auf eine angemessene Einbindung des betrieblichen Datenschutzbeauftragten, aber auch auf den Schutz von Geschäftsgeheimnissen zu achten ist. Angesichts der ständigen Fortentwicklung datenverarbeitender Systeme im Betrieb (Software-Updates, selbstlernende Elemente usw.) sollte eine Fortentwicklung von punktueller Zustimmung hin zu **dauerhafter Begleitung von Prozessen** durch die Interessenvertretungen erfolgen.

Die Weiterentwicklung des Beschäftigtendatenschutzes sollte sich auch mit der Phase der **Bewerbung** um einen Arbeitsplatz und der **Begründung des Arbeitsverhältnisses** befassen. So ist beispielsweise darauf zu achten, dass das geltende Recht zu unzulässigen Fragen des Arbeitgebers im Bewerbungsverfahren und bei der Einstellung (z. B. nach dem Bestehen einer Schwangerschaft) weder durch den Einsatz sog. Human-Resources-Algorithmen noch durch die Aufforderung unterlaufen werden darf, dem Arbeitgeber Zugang zu Social-Media-Konten zu gewähren.

Bei einer Weiterentwicklung des Beschäftigtendatenschutzes ist darauf zu achten, dass auch diejenigen Personen erfasst werden, die in **unüblichen Beschäftigungsformen** arbeiten. Durch die Zunahme unüblicher Beschäftigungsformen in der Plattformökonomie verfügen die betreffenden Personen nicht über die klassischen Arbeitnehmer- und Mitspracherechte. Es kann zu einem enormen Machtungleichgewicht zwischen dem Auftraggeber bzw. dem Plattformbetreiber einerseits und dem Auftragnehmer bzw. den über die Plattform Arbeitenden andererseits kommen, das sich auch auf den Datenschutz und die informationelle Selbstbestimmung auswirken kann. Dem ist durch geeignete rechtliche Vorschriften – idealerweise auf EU-Ebene – und die Weiterentwicklung institutioneller Rahmenbedingungen, etwa durch eine Interessenvertretung, entgegenzuwirken.

3.5.2 Patienten

Mit Blick auf die Vorteile eines digitalisierten Gesundheitswesens spricht sich die DEK grundsätzlich für einen **raschen Ausbau digitaler Infrastrukturen sowie Prüfungs- und Bewertungsverfahren für digitale Versorgungsleistungen** innerhalb des Gesundheitssektors aus. Der qualitative und quantitative Ausbau digitalisierter Versorgungsmaßnahmen sollte die informationelle Selbstbestimmung des Patienten und seine Gesundheitskompetenz stärken.[18]

Bereits jetzt werden im Zusammenhang mit Versorgungsleistungen eine Vielzahl personenbezogener Daten verarbeitet. Bei ihnen handelt es sich im Regelfall um Gesundheitsdaten und genetische Daten, also um besondere Kategorien personenbezogener Daten i.S.d. Art. 9 DSGVO. Die **besondere Schutzbedürftigkeit dieser Daten bei gleichzeitiger Stärkung der Selbstbestimmung** von Patienten und Krankenversicherten, auch im Bereich der Forschung (→ unten), ist bei der Ausgestaltung einer zukünftig maßgeblich digitalen Gesundheitslandschaft umfassend zu berücksichtigen.

17 Derzeit etwa für den Betriebsrat § 87 Abs. 1 Nr. 6 Betriebsverfassungsgesetz (BetrVG) und für den Personalrat § 75 Abs. 3 Nr. 17 Bundespersonalvertretungsgesetz (BPersVG).

18 Deutscher Ethikrat, Big Data und Gesundheit, Stellungnahme, 30.11.2017 (abrufbar unter: *https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-big-data-und-gesundheit.pdf*).

In diesem Zusammenhang betont die DEK die Dringlichkeit des Auf- und Ausbaus der **elektronischen Patientenakte** (ePA), um die Qualität, Transparenz und Wirtschaftlichkeit der medizinischen Versorgung zu verbessern.[19] Unter Berücksichtigung der zentralen Bedeutung der ePA für die Digitalisierung des Gesundheitswesens weist die DEK darauf hin, dass bei der Implementierung der ePA in erhöhtem Maße sowohl auf Aspekte der Informationssicherheit als auch auf die Wahrung der Patientenhoheit zu achten ist; so sollte etwa das bestehende Kryptosicherheitskonzept der dezentralen Verwaltung von Schlüsseln bei den Versicherten (sog. PIN) erhalten bleiben. Zudem sollte die ePA auch im Falle der Einwilligungsunfähigkeit des Patienten auf der Grundlage der auch ansonsten anwendbaren Regelungen zur gesetzlichen Vertretung unabhängig von der Art der Krankenversicherung nutzbar sein.

Die Bedeutung digitaler Gesundheitsdienstleistungen und Produkte, die nicht kollektiv finanziert werden (sog. **zweiter Gesundheitsmarkt**), nimmt beständig zu – auch vor dem Hintergrund, dass die gesetzlichen Krankenversicherungen bislang lediglich vereinzelt digitale Versorgungsangebote bereitstellen. Neben Fitness-, Gesundheits- und Wellness-Angeboten ist der Möglichkeit des digitalen **Selbst-Monitoring** durch Apps sowie entsprechende Wearables eine signifikante Relevanz im Kontext des digitalisierten Gesundheitswesens beizumessen. Die Qualität dieser Apps und damit auch die Verwertbarkeit der dadurch erhobenen Daten ist jedoch vielfach nicht hoch und auch nicht umfassend geprüft. Dies birgt für die betroffenen Patienten und Nutzer ein zuweilen beträchtliches Gesundheitsrisiko. Zudem sollte es den Patienten nicht zugemutet werden, die Qualität der jeweiligen Produkte und Dienstleistungen, allem voran mit Blick auf den Datenschutz und die Informationssicherheit, eigenständig zu bewerten, noch sollte die digitale Gesundheitsversorgung eine Frage der individuellen finanziellen Leistungsfähigkeit sein. Mit Blick auf diesen Befund begrüßt die DEK die vorgesehene Etablierung eines Verfahrens zur Prüfung und Bewertung entsprechender Apps durch das Bundesinstitut für Arzneimittel und Medizinprodukte.

3.5.3 Minderjährige

Die DEK begrüßt die Bemühungen, sowohl auf gesetzgeberischer Ebene als auch auf der Ebene der Selbstregulierung, besondere **Schutzmechanismen** für die digitale Selbstbestimmung Minderjähriger zu entwickeln. Diese sollen erstens einem stärkeren Datenschutz, dem Schutz vor Profilbildung, Manipulation durch Dark Patterns und Addictive Designs usw. dienen und zweitens einem besseren Schutz vor nicht altersgerechten (gewaltverherrlichenden usw.) Inhalten.

Allerdings erinnert die DEK auch daran, dass alle diese Schutzmechanismen ins Leere laufen, solange nicht ein zuverlässiges **Identitätenmanagement** gewährleistet ist und sichergestellt wird, dass Minderjährige auch als solche erkannt und behandelt werden. Eine vom Nutzer behauptete Altersangabe ist als Mittel der Überprüfung jedenfalls ungeeignet. Es wäre ethisch aber auch problematisch, zu fordern, dass Anbieter durch Erhebung – möglicherweise sogar besonders sensibler – personenbezogener Daten (etwa: Gesichtserkennung mit Datenübertragung in die Cloud des Anbieters) selbst eine Alterseinschätzung vorzunehmen haben oder aber die Last ganz den Erziehungsberechtigten aufzubürden, die damit leicht überfordert würden. Die DEK empfiehlt der Bundesregierung daher, die Entwicklung **familienadäquater Technologien** zu fördern, die eine selbstbestimmte Entwicklung der Minderjährigen ermöglichen und zugleich ihren Schutz zuverlässig gewährleisten.

19 Siehe hierzu bereits die Empfehlung der DEK für eine partizipative Entwicklung der elektronischen Patientenakte (ePA) vom 28.11.2018 (abrufbar unter: *www.datenethikkommission.de*).

In diesem Zusammenhang empfiehlt die DEK der Bundesregierung, insbesondere bei mobilen Endgeräten auf europäischer Ebene darauf zu dringen, dass die in der DSGVO festgelegten Prinzipien von **Datenschutz „by design" und „by default"** eingehalten werden, um den Schutz der informationellen Selbstbestimmung und der Privatheit Minderjähriger zu gewährleisten. Um die Hersteller der Betriebssysteme für mobile Endgeräte und die Anbieter digitaler Dienste dazu zu bringen, alle für die betreffenden Altersstufen geltenden rechtlichen Vorschriften einzuhalten und Dienste, die nicht altersgerecht sind, zu blockieren, müssten die deutschen und europäischen Datenschutzbehörden, die Kartellbehörden, die Medienaufsicht und die technischen Regulierungsbehörden in ihren jeweiligen Aufgaben- und Zuständigkeitsbereichen dazu beitragen, die Anforderungen durchzusetzen. Auch etwa die Akteure im Bereich der Schulen und Kindertagesstätten, in denen solche Systeme zum Einsatz kommen, sollten diese Anforderungen im Rahmen von Beschaffungen deutlich machen. Zur Notwendigkeit, Datenschutz „by design" und „by default" auch gegenüber Herstellern einzufordern (→ siehe näher unten).

Zu erwägen ist in diesem Zusammenhang darüber hinaus insbesondere die Einführung einer EU-weiten Verpflichtung für die Hersteller mobiler Endgeräte, ein Endgerät bereits beim Kauf irreversibel (oder nur mithilfe eines Schlüssels reversibel) und erkennbar als „Kinder-Endgerät" zu programmieren. Diese Programmierung hätte automatisch die Einhaltung aller für Kinder geltenden rechtlichen Vorschriften sicherzustellen und nicht altersgerechte Dienste zu blocken. Die Minderjährigen können den **bei Aktivierung eingestellten entsprechenden Status** ihres Geräts/Betriebssystems dabei nicht ohne Einverständnis der Eltern ändern. Eine solche Lösung hätte auch klare Vorteile gegenüber sog. Parental-Control-Apps, welche erstens vielfach ein eigenes Datenschutz- und Informationssicherheitsproblem darstellen und zweitens ethisch problematische Möglichkeiten der Totalüberwachung im privaten Bereich mit sich bringen.

3.5.4 Sonstige Pflege- und Schutzbedürftige

Die Verarbeitung von Daten vulnerabler Gruppen erfolgt vielfach zu deren eigenem Schutz, so etwa im Bereich der Pflege. Digitale Technologien ermöglichen beispielsweise älteren Menschen ein viel sichereres Verbleiben in der gewohnten Umgebung. Dies kann auch helfen, den negativen Auswirkungen des Fachkräftemangels in der Pflege entgegenzuwirken und eine bessere Versorgung sicherzustellen. Insbesondere **digitale Assistenzsysteme** können dabei, richtig eingesetzt, eine Brückentechnologie darstellen und sich adaptiv den unterschiedlichen Bedürfnissen verschiedener Menschen anpassen.

Sowohl das Recht auf Leben und auf körperliche Unversehrtheit als auch das Recht auf informationelle Selbstbestimmung stellen Grundrechte dar, die im Wege praktischer Konkordanz miteinander in Einklang zu bringen sind. Dabei sind insbesondere die **Gefahren für Leben bzw. Gesundheit** auf der einen Seite und die Intensität des Eingriffs in die informationelle **Selbstbestimmung** auf der anderen Seite zu berücksichtigen.

Bei der Überwachung durch professionelle Akteure im Pflegebereich bedarf es nach Auffassung der DEK der Erarbeitung von **Standards und Leitlinien** durch die Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (DSK). Diese sollten insbesondere vorgeben, auf welche Rechtsgrundlagen sich diese Akteure in welcher Situation stützen können und in welchen Fällen – insbesondere im Falle der Ermangelung einer **Einwilligung** der betroffenen Person oder ihres Betreuers – eine Maßnahme gegebenenfalls auf Art. 6 Abs. 1 lit. f oder lit. d DSGVO gestützt werden kann oder ganz unterbleiben muss. Auch Vorgaben für die Informationserteilung sollten darin enthalten sein, wobei nach Auffassung der DEK bereits im Vorfeld der Aufnahme in einer Einrichtung (z. B. Pflegeheim, Kindergarten, Schule) differenzierte Informationen über die Möglichkeiten der digitalen Überwachung erteilt und ggf. – soweit keine gesetzliche Rechtsgrundlage für die Datenverarbeitung besteht – auch differenzierte Einwilligungen eingeholt werden müssten. Derartige Standards und Leitlinien wären zugleich geeignet, für die Träger von Einrichtungen und das Pflegepersonal mehr Rechtssicherheit zu schaffen und Haftungsrisiken zu verringern. Zur Klarstellung, dass auch eine antizipierte Einwilligung der betroffenen Person in einer **Patientenverfügung** möglich ist, sollte § 1901a BGB entsprechend angepasst werden.

Als besonders schutzbedürftig sind grundsätzlich auch Personen anzusehen, die sich im häuslichen Bereich und somit im sicher gewähnten Zentrum ihrer räumlichen Privatsphäre bewegen. Auch hier entstehen im Zusammenhang mit neuen Technologien wachsende potenzielle **Überwachungsmöglichkeiten von Privatpersonen durch andere Privatpersonen** (z. B. die Überwachung von Partnern, Kindern oder Menschen mit Behinderung), bis hin zu einer ethisch äußerst bedenklichen Möglichkeit einer privaten Totalüberwachung. Da es vielfach an einer hinreichenden Sensibilität für das Thema fehlt, empfiehlt die DEK der Bundesregierung, aber auch den in vielen Punkten zuständigen Landesregierungen, diesbezüglich **bewusstseinsbildende Maßnahmen** zu ergreifen. Darüber hinaus empfiehlt die DEK der Bundesregierung, die Entwicklungen weiter zu beobachten, sieht jedoch derzeit noch keinen Bedarf für gesetzliche Maßnahmen (z. B. neue Straftatbestände).

3.6 Datenschutz durch Technikgestaltung

Diejenigen, die ethisch begründete Datenrechte wahrnehmen oder korrespondierende Datenpflichten befolgen müssen – seien es etwa Bürger, Unternehmen oder staatliche Stellen – müssen dazu auch in der Lage sein. Es bedarf dafür auch **technischer Voraussetzungen**, insbesondere befähigender Technologien. Durch solche Befähigungen darf allerdings nicht die Verantwortung für den Schutz grundlegender Rechte und Freiheiten auf individuelle Nutzer überwälzt werden. Hier ist vielmehr der Staat mit einer entsprechenden **Regulierung** gefordert, die den Schutz dieser grundlegenden Rechte und Freiheiten prinzipiell gewährleistet, ohne dass der Einzelne tätig werden müsste.

3.6.1 Datenschutzfreundliches Design von Produkten und Dienstleistungen

Mit Art. 25 DSGVO, der mit „Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen“ überschrieben ist, werden **Datenschutz „by design“ und „by default“** zur Pflicht für Verantwortliche. Dies bedeutet, dass Datenschutz im Sinne der Grundsätze aus Art. 5 DSGVO risikoadäquat bei der Technikgestaltung berücksichtigt werden muss. Die dafür notwendigen technischen und organisatorischen Maßnahmen müssen sowohl bereits vor der Verarbeitung, nämlich wenn der Verantwortliche die Mittel für die Verarbeitung festlegt, als auch während der eigentlichen Verarbeitung getroffen werden.

Datenschutz „by design“ und „by default“

Datenschutz „by design“ stellt die Wahl der technischen und organisatorischen Maßnahmen unter Bedingungen wie den Stand der Technik, die Implementierungskosten, die Verarbeitung und das Risiko für die Rechte und Freiheiten natürlicher Personen. **Datenschutz „by default“** ist nicht an solche Bedingungen geknüpft, muss also stets umgesetzt werden. In der Praxis werden allerdings oft überschießende personenbezogene Daten wie z. B. Identifikatoren verarbeitet, die Verarbeitung ist nicht ausreichend beschränkt, die Speicherfristen sind zu lang und es können mehr Personen auf die Daten zugreifen als nötig.

Aus diesem Grund wurden für das **„Privacy Engineering“** die weiteren Schutzziele Nichtverkettbarkeit, Transparenz und Intervenierbarkeit entwickelt, die mittlerweile als sogenannte Gewährleistungsziele Teil des Standard-Datenschutzmodells (SDM) der deutschen Datenschutzaufsichtsbehörden geworden sind.[1] Das SDM definiert ähnlich den IT-Grundschutz-Katalogen des BSI Bausteine, die von Verantwortlichen und Technikgestaltern herangezogen werden können, um die für ihren jeweiligen Schutzbedarf angemessenen technischen und organisatorischen Maßnahmen zu treffen. Bislang sind erst einige Bausteine verfügbar, weitere werden folgen. Die Anlehnung an die IT-Grundschutz-Kataloge und die Normenreihe ISO 2700x führt dazu, dass viele Entwickler mit dem grundsätzlichen Konzept vertraut sind und die rechtlichen Anforderungen besser bei der Konzeptionierung und Implementierung von technischen Systemen umsetzen können.

Auch die Frage, inwieweit eine **Zentralisierung oder eine Dezentralisierung** bei der Gestaltung von technischen Systemen zu bevorzugen ist, muss im Einzelfall geklärt werden. Zentralisierte Systeme erlauben in der Regel ein höheres Maß an Kontrolle und Einflussnahme durch die Betreiber. Dies kann gewollt sein, z. B. um Datenschutz- oder Informationssicherheitsfunktionalität durchzusetzen. Es kann aber auch kritisch werden, da die zentralisierte Datenhaltung und Steuerung der Verarbeitung ein höheres Missbrauchspotenzial aufweist – einerseits als Angriffsziel von Dritten, die an die Daten herankommen oder die Verarbeitung sabotieren wollen, andererseits durch den Betreiber selbst, beispielsweise durch eine Nutzung des großen Datenbestands zu anderen Zwecken als vorgesehen. Dezentralisierung kann in geeigneter Gestaltung dagegen gewährleisten, dass Daten nicht oder nicht einfach verknüpft werden können oder dass sich die Verfügbarkeit des Gesamtsystems schwerer stören lässt.

1 Arbeitskreis Technik der Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder: Das Standard-Datenschutzmodell – Eine Methode zur Datenschutzberatung und -prüfung auf der Basis einheitlicher Gewährleistungsziele V.1.1 – Erprobungsfassung, 2018 (abrufbar unter: *https://www.datenschutzzentrum.de/sdm/*).

Besondere Praxisrelevanz haben datenschutzrechtliche Designvorgaben in Bezug auf **Endgeräte**. Diese können am Körper (sog. Wearables, z. B. SmartWatch oder intelligente Textilien) oder zumindest in Körpernähe tragbar (z. B. Smartphone), anderweitig beweglich (z. B. vernetztes Auto) oder auch unbeweglich sein (z. B. Smart-Home-Einrichtungen). Dem Design der Softwaresysteme für solche Endgeräte kommt umso größere ethische Relevanz zu, je mehr sie in Körpernähe bzw. in sehr privaten und intimen Bereichen (z. B. Badezimmer, Schlafzimmer) zum Einsatz kommen, je stärker besonders vulnerable Personen (z.B. Kinder und Jugendliche, Pflegebedürftige, Personen mit Behinderungen) betroffen sind und je tiefer sie in die Persönlichkeit einer Person eindringen. Eine besondere Herausforderung an selbstbestimmungsfreundliches Design stellt das hohe Maß an (Selbst)Verantwortung dar, welches den Nutzern bei Zusammenstellung, Konfiguration und Betrieb der Geräte zugestanden bzw. abverlangt wird.

Die DEK empfiehlt der Bundesregierung, die Erforschung und Entwicklung **technischer Standards** für Endgeräte verstärkt zu fördern. Ferner empfiehlt die DEK nachdrücklich, auf europäischer Ebene auf die Einführung **technischer Vorgaben** zur Wahrung von Selbstbestimmung und digitaler Produktsicherheit im privaten Bereich zu dringen, insbesondere für den Bereich von **Endgeräten für Verbraucher**. Vorgaben an Endgeräte sollten nach Auffassung der DEK jedenfalls Folgendes umfassen:

- Produkte müssen auf dem Stand der Technik und dem Schutzbedarf angemessen vor **Cyberangriffen und zweckfremder Verwendung** von Daten geschützt werden, wobei insbesondere für sensible Daten (z.B. Gesundheitsdaten) geeignete Garantien vorliegen müssen. Die Wahrung eines hohen Grades an Cyberresilienz ist dabei eine Gemeinschaftsaufgabe von Staat, Wirtschaft und jedem Einzelnen;

- Es muss zu jedem Zeitpunkt klar ersichtlich sein, welche **Funktionen momentan aktiviert** sind, insbesondere ob GPS, Kamera, Mikrofon oder andere Sensoren eingeschaltet sind, ob eine Verbindung zum Internet besteht und ob Daten nach außerhalb des geschlossenen lokalen Bereichs übertragen werden;

- Die **Übertragung von Daten** nach außerhalb des lokalen Bereichs muss auf einfache Weise abzuschalten sein, und mittlerweile lokal gespeicherte Daten dürfen auch beim nächsten Einschalten nicht ohne den Willen des Nutzers übertragen werden (dies muss auch für einzelne Applikationen gelten, etwa auf Smartphones oder Smart-TV);

- Soweit **Basisfunktionen des Geräts auch ohne solche Datenübertragung** technisch möglich sind, müssen sie bei Abschalten der Übertragung erhalten bleiben (z.B. intelligenter Kühlschrank muss noch kühlen);

- Geräte sollten mit einem **„User Onboarding“**-Ansatz ausgeliefert werden, wobei das Onboarding bei erster Inbetriebnahme automatisch erfolgen und sich auch für Zweitnutzer nach Belieben wiederholen lassen sollte. Dabei sollte nicht nur die Funktionsweise erläutert werden, sondern ebenso die Erfassung und weitere Verarbeitung von Nutzerdaten;

- Endgeräte, die direkt mit dem Internet verbunden (z. B. Router) und mittels eines Passworts abgesichert sind, sollten nicht ohne eine vorherige Änderung des Initialpassworts in Betrieb genommen werden können. Systemseitig sollten nur **Passwörter**, die dem Stand der Technik entsprechen, zugelassen werden.

Nachvollziehbarkeit und Transparenz

Datenschutz „by design“ umfasst ebenso die Nachvollziehbarkeit und Transparenz der Systeme, einschließlich der Anwendungen, Skripte, Quellen und Elemente zu jedem Entwicklungs- und Prozesszeitpunkt. Die DEK begrüßt die laufenden Bemühungen, Best-Practice-Modelle für gute Allgemeine Geschäftsbedingungen (AGB) und Verbraucherinformationen zu entwerfen („One Pager“). Dabei sollen Verbraucher im Rahmen eines Mehrebenen-Ansatzes in einem ersten Schritt einfache, konzentrierte Informationen über die wesentlichen Datenverarbeitungen erhalten und – wenn gewünscht – in einem weiteren Schritt zu den ausführlichen AGB und Datenschutzinformationen geleitet werden. Dies wird aber nicht ausreichen, das Problem der unzureichenden und/oder den Verbraucher überfordernden und damit ihr Ziel verfehlenden Information zu lösen.

Um dem Verbraucher eine informierte Kaufentscheidung zu ermöglichen, sollten auf europäischer Ebene unter maßgeblicher Einbeziehung der Wirtschaft und Zivilgesellschaft einheitliche, maschinenlesbare und intuitiv verständliche Bildsymbole (**Piktogramme**) eingeführt werden, die wesentliche digitale Merkmale von Produkten, einschließlich digitalen Produkten (z. B. Apps), und Dienstleistungen vermitteln (z. B. für die Merkmale „Basisfunktionen nur mit Internetverbindung“, „Verfügt über Internetverbindung für Komfortfunktionen“, „Übermittelt Nutzerdaten“ und „Nutzer-Tracking“) und zusätzlich – insbesondere für graduell in unterschiedlichem Ausmaß gegebene Produktmerkmale – durch **Farbcodierungen** unterstützt sein können. Der Bundesregierung wird empfohlen, bei der Europäischen Kommission auf die Entwicklung solch standardisierter Bildsymbole gemäß Art. 12 Abs. 8 DSGVO hinzuwirken.

Die Förderung der Entwicklung zertifizierter **elektronischer Einkaufsassistenten**, die im Ladengeschäft oder Webshop ein Produkt identifizieren und Produktinformationen adressatengerecht aufarbeiten, kann zusätzliche Transparenz für Verbraucher schaffen.

Von der Gestaltung der Produkte, Dienste und Anwendungen hängt es ganz wesentlich ab, inwieweit die Verantwortlichen und Verarbeiter ihre Datenschutzpflichten erfüllen können. Jedoch sind Hersteller, die nicht selbst personenbezogene Daten verarbeiten, keine Adressaten der DSGVO. Die Verantwortlichen, die nicht auf Eigenentwicklungen zurückgreifen können oder wollen, müssen also eingebauten Datenschutz einfordern.[20] Die DEK empfiehlt der Bundesregierung daher, Maßnahmen zu ergreifen bzw. Maßnahmen anderer Akteure zu fördern, die zu einer verstärkten **Verantwortlichkeit der Hersteller** führen. Dies kann etwa geschehen durch:

- Unmittelbare **Vorgaben für Produktdesign und Produktsicherheit** durch den Gesetzgeber;
- Schaffung **wirksamer Rechtsbehelfe** entlang der Vertriebskette, mit deren Hilfe die Verantwortlichkeit für unzureichenden Datenschutz „by design“ und „by default“ auf die Hersteller[21] abgewälzt werden kann (vgl. gewisse Fortschritte der Abwälzung vom Verbraucher auf den Händler und entlang der Vertriebskette durch die neue EU-Richtlinie 2019/771 über den Warenkauf);

20 Vgl. Erwägungsgrund 78 zur DSGVO.
21 Christiane Wendehorst: Verbraucherrelevante Problemstellungen zu Besitz- und Eigentumsverhältnissen beim Internet der Dinge, Teil 2: Wissenschaftliches Rechtsgutachten, Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen, Dezember 2016, S. 120 (abrufbar unter: *http://www.svr-verbraucherfragen.de/wp-content/uploads/Wendehorst-Gutachten.pdf*).

- Gestaltung von **Ausschreibungen** sowie der Richtlinien für **Beschaffungsmaßnahmen** der öffentlichen Hand in einer Weise, die den Nachweis absoluter DSGVO-Konformität einschließlich der Einhaltung von Datenschutz „by design“ und „by default“ einfordert;

- Schaffung von **Anreizen** für ein besonders hohes Maß an Datenschutz „by design“ und „by default“, etwa durch entsprechende Bedingungen in staatlichen Förderprogrammen.

3.6.2 Datenschutzfreundliche Produktentwicklung

Datenschutz durch Technikgestaltung ist auch bei der Produktentwicklung und Produktweiterentwicklung zu beachten. Dies gilt insbesondere für die **Entwicklung algorithmischer Systeme**, bei denen typischerweise große Mengen an Datensätzen – etwa als Trainingsdaten – erforderlich werden (→ zu Einzelheiten Teil C, 2.2).

Datenschutzfreundliches Trainieren algorithmischer Systeme

Um die Datenschutz-Grundsätze in Art. 5 DSGVO beim Trainieren algorithmischer Systeme zu erfüllen, bestehen verschiedene Möglichkeiten. So hat etwa die Norwegische Datenschutzaufsichtsbehörde Datatilsynet im Januar 2018 Mittel und Methoden für ein datenschutzfreundliches Trainieren algorithmischer Systeme[1] vorgeschlagen:

1. Einsatz **datenminimierender Verfahren** bezüglich der Trainingsdaten, z. B. durch das Verwenden synthetischer Daten (beispielsweise über sog. Generative Adversarial Networks), durch föderales Lernen oder durch den Einsatz von datensparsamen Varianten, wie sie für neuronale Netze vorgeschlagen werden;

2. Einsatz von **Verschlüsselungsverfahren** wie Differential Privacy, Homomorphic Encryption oder anderer Verfahren, die Informationsabfragen erlauben, ohne einen Vollzugriff auf die Datenbank zu gewähren;

3. Einsatz **transparenzfördernder Verfahren**, um eine höhere Verständlichkeit und Nachvollziehbarkeit zu erreichen.

Die DEK sieht in all diesen Bereichen allerdings noch **Forschungsbedarf**. Dies betrifft auch Möglichkeiten des datenschutzfreundlichen Testens der algorithmischen Systeme.

1 Datatilsynet: Artifical intelligence and privacy, Report, Januar 2018, S. 27 f. (abrufbar unter: *https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf*).

Zusammenfassung der wichtigsten Handlungsempfehlungen

Anforderungen an die Nutzung personenbezogener Daten

1

Die DEK empfiehlt **Maßnahmen gegen ethisch nicht-vertretbare Datennutzungen**. Dazu gehören etwa Totalüberwachung, die Integrität der Persönlichkeit verletzende Profilbildung, gezielte Ausnutzung von Vulnerabilitäten, sog. Addictive Designs und Dark Patterns, dem Demokratieprinzip zuwiderlaufende Beeinflussung politischer Wahlen, Lock-in und systematische Schädigung von Verbrauchern sowie viele Formen des Handels mit personenbezogenen Daten.

2

Sowohl das Datenschutzrecht als auch die übrige Rechtsordnung (u. a. Zivilrecht, Lauterkeitsrecht) enthalten bereits eine Fülle von Instrumenten, die gegen derartige Datennutzungen eingesetzt werden können. Gemessen an Breitenwirkung und Schädigungspotenzial werden diese Instrumente indessen bislang nicht in ausreichender Weise genutzt – insbesondere gegenüber marktmächtigen Unternehmen. Dieses **Vollzugsdefizit** hat verschiedene Ursachen, die es systematisch anzugehen gilt.

3

Neben der Schärfung des Bewusstseins bei handelnden Akteuren (z. B. Aufsichtsbehörden) für die bereits bestehenden Möglichkeiten ist dringend eine **Konkretisierung und punktuelle Verschärfung des geltenden Rechtsrahmens** angezeigt. Dazu gehören etwa eine spezielle Normierung von datenspezifischen Klauselverboten, Schutz- und Treuepflichten, Deliktstatbeständen und unlauteren Geschäftspraktiken sowie die Schaffung eines weitaus konkreteren Rechtsrahmens für Profilbildungen und Scoring wie auch für den Datenhandel.

4

Um die Wirkungskraft der Aufsichtsbehörden zu erhöhen, bedürfen diese einer weitaus besseren personellen und sachlichen Ausstattung. Sofern es nicht gelingt, die Abstimmung unter den deutschen Datenschutzaufsichtsbehörden zu verstärken und zu formalisieren und so die einheitliche und kohärente Anwendung des Datenschutzrechts zu gewährleisten, ist eine **Zentralisierung der Datenschutzaufsicht für den Markt** in einer – mit einem weiten Mandat ausgestatteten und eng mit anderen Fachaufsichtsbehörden kooperierenden – Behörde auf Bundesebene zu erwägen. Die Zuständigkeit der Landesdatenschutzbehörden für den öffentlichen Bereich soll hingegen unangetastet bleiben.

5

Die Anerkennung von **„Dateneigentum“** im Sinne eines dem Sacheigentum oder dem geistigen Eigentum nachgebildeten Ausschließlichkeitsrechts an Daten würde nach Auffassung der DEK bestehende Probleme nicht lösen und stattdessen eine Reihe neuer Probleme schaffen. Sie wird daher **nicht empfohlen**. Die DEK empfiehlt auch nicht die Anerkennung genereller wirtschaftlicher Verwertungsrechte an personenbezogenen Daten, wie sie etwa durch Verwertungsgesellschaften geltend gemacht werden könnten.

6

Wenngleich die plakative Bezeichnung zur allgemeinen Bewusstseinsbildung beigetragen hat, plädiert die DEK dafür, **von der Bezeichnung von Daten als „Gegenleistung“ abzusehen**. Unabhängig von der künftigen Auslegung des sog. Koppelungsverbots durch die Aufsichtsbehörden und den EuGH fordert die DEK, dass Verbrauchern jeweils **zumutbare Alternativen** gegenüber der Freigabe von Daten zur auch kommerziellen Nutzung angeboten werden müssen (z. B. entsprechend ausgestaltete **Bezahlmodelle**).

7

Die Verwendung von Daten zur **personalisierten Risikoeinschätzung** (z. B. im Rahmen von Telematiktarifen bei bestimmten Versicherungen) sollte an **enge Voraussetzungen** geknüpft werden. So darf die Datenverarbeitung beispielsweise nicht den Kern privater Lebensführung betreffen, es muss ein klarer ursächlicher Zusammenhang zwischen Daten und Risiko vorliegen, und die Preisdifferenz zwischen personalisiertem und nicht personalisiertem Tarif sollte im Einzelnen noch festzulegende Prozentwerte nicht überschreiten. Weitere Anforderungen betreffen Transparenz, Nichtdiskriminierung und den Schutz dritter Personen.

8

Die DEK empfiehlt der Bundesregierung, Fragen rund um den **„digitalen Nachlass“** mit dem Urteil des BGH von 2018 nicht als erledigt anzusehen. Die praktisch lückenlose Aufzeichnung von digital geführter Kommunikation, die in vielen Fällen an die Stelle des flüchtig gesprochenen Wortes tritt, und ihre Aushändigung an Erben bedeutet eine neue Dimension von Gefährdung für die Privatheit. Ihr sollte mit einer Reihe von Maßnahmen begegnet werden, welche neue Pflichten von Dienstanbietern, Qualitätssicherung bei Angeboten digitaler Nachlassplanung sowie nationale Regelungen zum postmortalen Datenschutz umfassen.

9

Die DEK empfiehlt der Bundesregierung, die Sozialpartner einzuladen, ausgehend von den bereits in Tarifverträgen bestehenden Beispielen guter Übung eine gemeinsame Linie für gesetzliche Konkretisierungen des **Beschäftigtendatenschutzes** zu entwickeln. Dabei sollten auch die Belange von Personen in unüblichen Beschäftigungsformen berücksichtigt werden.

10

Mit Blick auf die Vorteile eines **digitalisierten Gesundheitswesens** spricht sich die DEK für einen raschen Ausbau digitaler Infrastrukturen innerhalb des Gesundheitssektors aus. Der qualitative und quantitative Ausbau digitalisierter Versorgungsmaßnahmen sollte die informationelle Selbstbestimmung des Patienten stärken. Hierzu gehört der partizipative Auf- und Ausbau der elektronischen Patientenakte (ePA) sowie die Weiterentwicklung von Verfahren zur Prüfung und Bewertung digitaler Gesundheitsanwendungen im ersten und zweiten Gesundheitsmarkt.

11

Die DEK fordert, dem erheblichen Vollzugsdefizit des geltenden Rechts betreffend den **Schutz von Kindern und Jugendlichen** im digitalen Raum abzuhelfen. Insbesondere sollten Technologien – einschließlich eines effektiven Identitätenmanagements – sowie Standardoptionen entwickelt und verpflichtend vorgesehen werden, welche einen zuverlässigen Schutz der Kinder und Jugendlichen gewährleisten und zugleich familienadäquat sind, indem sie Erziehungsberechtigte weder überfordern noch eine übermäßige Überwachung im privaten Bereich ermöglichen oder gar hierzu animieren.

12

Was den Umgang mit Daten **pflege- und schutzbedürftiger Menschen** betrifft, sollte für professionelle Akteure im Pflegebereich durch Standards und Leitlinien mehr Rechtssicherheit geschaffen werden. Zugleich ist eine gesetzliche Klarstellung zu erwägen, dass – soweit eine Datenverarbeitung auf die Einwilligung des pflege- und schutzbedürftigen Menschen gestützt werden muss – in Patientenverfügungen auch bestimmte Dispositionen in Bezug auf die Datenverarbeitung (z. B. für den Fall der dauernden Einwilligungsunfähigkeit infolge von Demenz) getroffen werden können.

13

Die DEK empfiehlt, eine Reihe verbindlicher Vorgaben für **datenschutzfreundliches Design von Produkten und Dienstleistungen** einzuführen und damit die an Verantwortliche im Sinne der DSGVO gerichteten Vorgaben von Datenschutz „by design“ und „by default“ bereits auf der Ebene der Hersteller wie auch der Diensteanbieter wirksam werden zu lassen. Dies betrifft insbesondere Vorgaben für Verbraucherendgeräte. In diesem Zusammenhang sind auch einheitliche Bildsymbole (Piktogramme) einzuführen, die dem Verbraucher eine informierte Kaufentscheidung ermöglichen.

14

Ferner bedarf es einer Reihe weiterer Maßnahmen auf verschiedenen Ebenen, um für Hersteller effektive **Anreize zur Implementierung eines datenschutzfreundlichen Designs** zu schaffen. Neben wirksamen Rechtsbehelfen entlang der Vertriebskette, mit deren Hilfe Hersteller mit in die Verantwortung für unzureichenden Datenschutz „by design“ und „by default“ genommen werden können, ist insbesondere an Vorgaben in Ausschreibungsbedingungen und Beschaffungsrichtlinien für die öffentliche Hand sowie an Bedingungen bei Förderprogrammen zu denken. Das Gleiche gilt für datenschutzfreundliche **Methoden der Produktentwicklung**, einschließlich des Trainierens algorithmischer Systeme.

15

Trotz des berechtigten Fokus auf Datenschutz natürlicher Personen darf der **Schutzbedarf von Unternehmen und juristischen Personen** nicht in den Hintergrund treten. Durch die umfassende Verknüpfbarkeit von Einzeldaten kann ein lückenloses Bild interner Betriebsabläufe entstehen und in die Hände von Konkurrenten, Verhandlungspartnern, Übernahmeinteressenten usw. gelangen. Dies stellt aufgrund umfangreicher Datenflüsse in Drittstaaten u. a. eine Gefährdung der digitalen Souveränität Deutschlands und Europas dar. Viele Handlungsempfehlungen sind daher sinngemäß auch auf die Daten juristischer Personen zu übertragen. Die DEK fordert die Bundesregierung auf, Schritte zu unternehmen, um den **datenbezogenen Schutz von Unternehmen zu verbessern**.

4. Verbesserung des kontrollierten Zugangs zu personenbezogenen Daten

Daten – auch personenbezogene Daten – sind eine **zentrale Ressource** der Datenwirtschaft und Schlüssel für viele wohlfahrtsfördernde Anwendungen. Die rasante Entwicklung digitaler Technologien – auch solcher, von denen jeder Einzelne enorm profitiert – wurde unter anderem durch die Auswertung der Daten von Milliarden von Nutzern weltweit ermöglicht. Auch wenn bei personenbezogenen Daten zunächst immer der Datenschutz im Mittelpunkt der Betrachtung steht, stellt sich doch verstärkt die Frage, inwieweit die generelle Verbesserung eines kontrollierten Zugangs zu personenbezogenen Daten – im Sinne des Prinzips der Wohlfahrt durch Nutzung und Teilen von Daten (→ oben 1.3.) und innerhalb des vom Datenschutzrecht vorgegebenen Rahmens – ethisch vertretbar oder sogar wünschenswert wäre.

4.1 Ermöglichung von Forschung mit personenbezogen Daten

4.1.1 Vorüberlegungen

Forschung stellt die Basis nahezu all unserer technischen Errungenschaften dar. Unter Bedingungen zunehmender Digitalisierung kommt datenbasierter Forschung dabei eine **herausragende Bedeutung** zu. Diese wird von der DSGVO bereits anerkannt und vom nationalen Recht, namentlich dem BDSG und den Landesdatenschutzgesetzen, punktuell gestärkt. Die DEK unterstreicht auch die signifikante Bedeutung der Verarbeitung genetischer, biometrischer und weiterer **Gesundheitsdaten** zu Forschungszwecken, zur Förderung der Prävention sowie zur Entwicklung neuer diagnostischer und therapeutischer Maßnahmen. Gerade der Einsatz Künstlicher Intelligenz verspricht in bestimmten Bereichen große Fortschritte, er ist aber je nach Fragestellung auf umfangreiche Datenbestände angewiesen. Die Freigabe von Gesundheitsdaten für Forschungszwecke wird auch immer wieder unter dem **Begriff der „Datenspende**“ diskutiert. Dieser Begriff ist jedoch **irreführend**, weil Daten im Unterschied zur Spende eines Organs oder einer Geldspende beliebig oft sowie gleichzeitig und auch vom Datengeber selbst weiterverwendet werden können.

Soweit die Forschungstätigkeit maßgeblich auf eine gemeinwohlorientierte Datennutzung ausgerichtet ist (etwa zur Gesundheitsvorsorge, zur Entwicklung nachhaltiger Mobilitätskonzepte oder allgemein zur Verbesserung von Lebensbedingungen), empfiehlt die DEK, vorhandene **datenschutzrechtliche Privilegierungstatbestände** auszuschöpfen und Forschung im Rahmen von Abwägungen als ein besonders gewichtiges Interesse zu werten.[22] Ergänzend sollten die Bundesländer vorhandene Regelungsbefugnisse, beispielsweise im Kontext des Hochschulrechts oder aber auch im Rahmen des Datenschutzrechts, innovationsfreundlich sowie im Geiste des vorgenannten Forschungsprivilegs ausfüllen. Der Begriff der wissenschaftlichen Forschung ist dabei – auch unter Einbeziehung der Rechtsprechung des Bundesverfassungsgerichts – weit zu verstehen. Nicht entscheidend ist dabei, ob die jeweilige Forschungstätigkeit durch öffentliche oder durch private Stellen betrieben wird.

Die DEK gibt zu bedenken, dass innerhalb des Spannungsverhältnisses zwischen den Grundrechtspositionen der Forschenden sowie der informationellen Selbstbestimmung der Betroffenen stets ein **angemessener Ausgleich** zu suchen ist. Im Rahmen der gesetzlich erforderlichen Abwägungen ist der **Schutz sensibler Daten** und damit einhergehend die Rechte der Betroffenen, wie beispielsweise Patienten oder Versicherte, besonders zu gewichten. Dabei kann sich zum Beispiel die Verschwiegenheitspflicht, die an bestimmte Berufsgeheimnisträger, wie Ärzte (vgl. § 203 Strafgesetzbuch), adressiert ist, auf die Arbeit von Forschungsinstitutionen auswirken, soweit diese auf Daten angewiesen sind, die bei jenen Berufsgeheimnisträgern erhoben werden bzw. gespeichert sind. Dies erfordert die Berücksichtigung der zum Schutz informationeller Selbstbestimmung normierten Verfahrensvorkehrungen.

22 Vgl. Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder: Orientierungshilfe der Aufsichtsbehörden für Anbieter von Telemedien, 2019, S. 14 (abrufbar unter: *https://www.datenschutzkonferenz-online.de/media/oh/20190405_oh_tmg.pdf*).

4.1.2 Rechtsklarheit und Rechtssicherheit

Bereits das derzeit geltende Recht ermöglicht und fördert datenbasierte Forschung. Allerdings stellen sich im Detail **Auslegungsfragen**, die der weiteren Klärung durch Aufsichtsbehörden und Gerichte bedürfen. So ist beispielsweise noch nicht abschließend geklärt, ob die **Weiterverwendung** für Forschungszwecke von Daten, die zu einem anderen Zweck (z. B. Gesundheitsversorgung) einmal rechtmäßig erhoben wurden, wegen Art. 5 Abs. 1 lit. b DSGVO und im Lichte von Erwägungsgrund 50 bei „geeigneten Garantien" i.S.v. Art. 89 DSGVO automatisch rechtmäßig ist, oder ob dafür ebenso eine eigene Rechtsgrundlage in Art. 6 Abs. 1-3 oder Art. 9 DSGVO gegeben sein muss wie für die erste Erhebung (bei Gesundheitsdaten wären dies z. B. gemäß § 27 BDSG eine ausdrückliche Einwilligung oder ein „erhebliches Überwiegen" der Forschungsinteressen). Teilweise wird auch vertreten, dass sich nur derjenige auf das Weiterverarbeitungsprivileg berufen könne, der die Daten selbst erhoben hat. Ebenso besteht Unsicherheit in Bezug auf die Reichweite des Forschungsbegriffs im Zusammenhang mit der **Entwicklung und Weiterentwicklung von Produkten**.

Auch wenn der rechtliche Rahmen für datenbasierte Forschung in Deutschland – auch in Bezug auf Gesundheitsdaten und andere besondere Kategorien von Daten – durchaus vorhanden ist, fehlt es diesem Regelungsrahmen schon aufgrund der föderalen Struktur und den grundgesetzlich festgeschriebenen Gesetzgebungskompetenzen von Bund und Ländern in Details an Einheitlichkeit. Dies führt aus Sicht der Forschung zu **Rechtsunsicherheit**, die zusätzlich dadurch verstärkt wird, dass verlässliche Auslegungshilfen, insbesondere was die Anforderungen an eine wirksame Einwilligung und das „erheblich überwiegende Interesse" des Forschenden i.S.d. § 27 BDSG betrifft, noch ausstehen. Diese Rechtsunsicherheit könnte die datenbasierte Forschung in Deutschland beeinträchtigen. Die DEK regt daher an, dass – etwa durch die Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (DSK), unter Einbeziehung relevanter Stakeholder aus Politik, Gesundheitswirtschaft und Zivilgesellschaft – **Handlungsempfehlungen sowie Auslegungskriterien** für einen praktikablen und rechtssicheren Umgang mit den betreffenden Normen ausgearbeitet werden (→ zu Standards der Pseudonymisierung und Anonymisierung unten 4.2).

Zur **weiteren Harmonisierung** der verschiedenen Regelungen im Bereich der Forschung (unterschiedliche mitgliedstaatliche Regelungen in der EU, Aufgabenteilung zwischen BDSG und Landesdatenschutzgesetzen, Spezialregelungen für besondere Bereiche) empfiehlt die DEK der Bundesregierung,

a) auf eine Synchronisierung der forschungsspezifischen **Rechtsgrundlagen** im BDSG und in den Landesdatenschutzgesetzen sowie in bereichsspezifischen Gesetzen zu dringen;

b) auf **europäischer Ebene** Vorhaben voranzutreiben, die auf eine stärkere Harmonisierung der mitgliedstaatlichen Regelungen zum Forschungsdatenschutz abzielen; sowie

c) auf ein **Notifizierungserfordernis** für mitgliedstaatliche Regelungen in diesem Bereich und auf die Einrichtung einer europäischen **Clearing-Stelle** für grenzüberschreitende Forschungsprojekte hinzuarbeiten.

4.1.3 Einwilligungsprozesse bei sensiblen Daten

Ein zentrales Instrument des Schutzes von Teilnehmern an Forschungsvorhaben (sog. Probanden), insbesondere an klinischer Forschung und Forschung mit Gesundheitsdaten sowie anderen besonders sensiblen Datenkategorien, ist die freiwillige, informierte und ausdrückliche Einwilligung der betroffenen Personen. Sie dient dazu, die **informationelle Selbstbestimmung** des Probanden zur Geltung kommen zu lassen. Zusätzlich stellt sie durch die verständliche Information über das Forschungsvorhaben sicher, dass die Teilnahme an der Studie den **Werten und Präferenzen** des Probanden entspricht. Als rechtlich verankertes Schutzinstrument fördert sie Transparenz und damit auch das Vertrauen in die Forschung. Nicht zuletzt trägt sie zur Integrität von Forschung und Forschenden bei.

Die Einholung einer informierten Einwilligung stellt die verantwortlichen Forschenden allerdings insbesondere im Kontext sensibler Daten vor erhebliche Herausforderungen. Wenn etwa ein neues Forschungsprojekt an Gesundheitsdaten durchgeführt werden soll, die bereits in einer Datenbank liegen, muss die betreffende Person kontaktiert werden, um erneut eine Einwilligung einzuholen, falls sie nicht – wie es als sog. breite Einwilligung (**broad consent**) im Bereich der Ethik diskutiert wird – schon ursprünglich in die Weiterverwendung ihrer Daten eingewilligt hat. Sollen Gesundheitsdaten aus der alltäglichen medizinischen Versorgung für Forschungszwecke verwendet werden, ist der Zugang zu den Patienten, um eine informierte Einwilligung einzuholen, ebenfalls mit hohen praktischen Hürden versehen. Die DEK empfiehlt vor diesem Hintergrund die Ausgestaltung und Aufbereitung entsprechender **Musterverfahren zur Einholung von Einwilligungen**, um die forschungsbezogene Verarbeitung in diesem Bereich zusätzlich zu erleichtern.

Unter ausdrücklicher Berücksichtigung des Grundrechtsgehalts der Einwilligung spricht sich die DEK ergänzend für die Entwicklung **innovativer Einwilligungsmodelle** im Forschungskontext aus. So sind bereits dynamische und für den Einzelfall angepasste Einwilligungserklärungen (**dynamic consent**) in Erprobung. Dabei ist dafür zu sorgen, dass der Einwilligende auch nach Abgabe der Einwilligung die Möglichkeit behält, die Kontrolle über seine Daten auszuüben. Zu diesem Zweck empfiehlt die DEK die Entwicklung und Ausgestaltung von Privacy Management Tools (PMT) und Personal Information Management Systems (PIMS) (→ unten) für den Forschungsbereich, wie z. B. **digitale Einwilligungsassistenten** oder Datenagenten. Derartige Einwilligungsassistenten können maßgeblich dazu beitragen, dass der Betroffene auch nach Beginn des Verarbeitungsprozesses den Überblick über seine Erklärungen behält, bei geänderter Sachlage erneut zur Abgabe entsprechender Erklärungen aufgefordert wird sowie auf einfache Weise seine Einwilligung widerrufen kann.

Insbesondere im Zusammenhang mit der Forschung an Gesundheitsdaten wird verstärkt die Förderung einer weitergehenden, unabhängig von einem konkreten Behandlungs- oder sonstigen Anlassfall erfolgenden Freigabe von Daten für die Forschung im Sinne einer pauschalen Einwilligung (**blanket consent**) diskutiert. Selbst wenn hierfür aus Sicht der Forschung gewichtige Gründe ins Feld geführt werden können, stehen diesem Konzept eine Reihe von Bedenken und Hindernissen gegenüber, darunter insbesondere das Erfordernis der Zweckbestimmtheit und der Informiertheit der Einwilligung. Selbst bei weitreichenden rechtlichen Absicherungen des Einwilligenden gegen missbräuchliche Verwendung seiner Daten und zum Schutz seiner Privatheit können seine Präferenzen und Werte nicht differenziert berücksichtigt werden.

Vor diesem Hintergrund empfiehlt die DEK die Prüfung eines innovativen Einwilligungsmodells, das als **„Meta-Consent“** in der Diskussion ist.[23] Unabhängig von einem konkreten Anlass entscheidet der Datengeber nach Beratung, für welche Art von Forschungsvorhaben er in welchem Forschungskontext welche Art von Einwilligung (spezifische oder breite Einwilligung) geben möchte. So kann er etwa bezüglich der folgenden Aspekte seine Festlegungen treffen:

- Forschungskontext (z. B. private oder öffentliche Forschung, kommerzielle oder nicht-kommerzielle Forschung, nationale, europäische oder internationale Forschung);
- Datenquellen (z. B. elektronische Patientenakte, Gewebe, Gesundheitsdaten, Lifestyle-Daten von Wearables);
- Art der Forschung (z. B. Präventionsforschung, Forschung zu Krebserkrankungen oder neurodegenerativen Erkrankungen, jede Art der Gesundheitsforschung).

Wenn die Daten anschließend für ein konkretes Forschungsvorhaben genutzt werden sollen, wird der Datengeber hierüber vorab **informiert** und erhält die Möglichkeit, dieser konkreten Datennutzung zu **widersprechen**.

Die konkrete Umsetzung des Modells im Einzelfall sollte auf jeden Fall unter der **Kontrolle** durch einen Treuhänder, eine Ethik-Kommission oder eine andere zuständige Stelle erfolgen, so dass die tatsächliche Umsetzung der Präferenzen des Einwilligenden gewährleistet ist. Die in einem Meta-Consent von dem Betroffenen niedergelegten Festlegungen zu seiner Einwilligung können von ihm jederzeit geändert werden. Auch hierfür sind die technischen und regulatorischen Voraussetzungen zu gewährleisten.

Beispiel 13

Der Datengeber legt fest, dass die Daten aus seiner elektronischen Patientenakte für öffentliche und kommerzielle Forschung genutzt werden dürfen. Zudem legt er fest, dass Blut- und Gewebeproben zur öffentlichen und kommerziellen Forschung zu degenerativen Erkrankungen genutzt werden dürfen. Seine Einwilligung beschränkt er bezüglich der Daten aus der elektronischen Patientenakte auf den europäischen Raum. Ein Unternehmen aus Spanien möchte sowohl Daten aus der elektronischen Patientenakte als auch Daten der Gewebeproben zur Demenzfoschung nutzen. Hierüber wird der Datengeber informiert und erhält vier Wochen Zeit, der Datennutzung zu widersprechen.

23 Thomas Ploug / Søren Holm: Bioethics, 2016 (30:9), S. 721, 721 ff.

Bei der Prüfung und Ausgestaltung des Modells ist zu berücksichtigen, dass die **Forschungsfreiheit** und das Weiterverarbeitungsprivileg im Vergleich zur geltenden Rechtslage nicht eingeschränkt wird. Vielmehr soll durch das Modell eines Meta-Consent betont werden, dass die Datengeber ihre **Werte und Präferenzen** im Hinblick auf die Verwendung ihrer Gesundheitsdaten für Forschungszwecke zum Ausdruck bringen können. Das würde zudem das Vertrauen der Gesellschaft in den Umgang mit Gesundheitsdaten stärken.

Es ist darüber hinaus zu bedenken, dass nicht nur die Nutzung, sondern auch die Nichtnutzung von Daten in ethischer Hinsicht zu verantworten ist, da so möglicher Fortschritt in wichtigen Bereichen verhindert wird. Zudem können ganz bestimmte **Gruppen vom Fortschritt ausgeschlossen** und damit diskriminiert werden. So können etwa für hochaltrige Personen mit mehreren chronischen Erkrankungen, die mehrere Medikamente gleichzeitig einnehmen, aus methodischen Gründen nur sehr eingeschränkt klinische Studien aufgesetzt werden. Durch eine qualitativ hochwertige Auswertung ihrer Gesundheitsdaten können aber wichtige Erkenntnisse über Wechselwirkungen zwischen Medikamenten und ihre Wirkung unter Alltagsbedingungen gewonnen und für eine weitergehende Forschung sowie die weitere Behandlung dieser Patienten fruchtbar gemacht werden.

Vor diesem Hintergrund und im Hinblick auf den auch im europäischen Kontext sowohl medizinisch als auch wirtschaftlich bedeutsamen Gesundheitssektor empfiehlt die DEK eine aktive **Förderung eines „lernenden Gesundheitssystems“**. In einem solchen System werden die Daten aus der alltäglichen Gesundheitsversorgung systematisch und qualitätsgestützt im Sinne der evidenzbasierten Medizin forschend genutzt, um mit den Ergebnissen die Versorgung kontinuierlich zu verbessern. Ein lernendes Gesundheitssystem bringt hohe Anforderungen an ein Mehrebenen-Governance-System mit sich und stellt den Patienten bzw. Versicherten ins Zentrum einer sektorenübergreifenden Gesundheitsversorgung.

4.1.4 Rechtlicher Diskriminierungsschutz

Die DEK weist allerdings auch darauf hin, dass bei der Ausgestaltung und Konzeption neuer, gesundheitsbezogener Forschungsvorhaben das erhebliche **Diskriminierungspotenzial** sensibler Daten (z. B. auf dem Arbeitsmarkt oder beim Abschluss von Versicherungen) zu berücksichtigen ist. Sowohl der technische Fortschritt in der Sequenzierung und Auswertung des menschlichen Genoms als auch die Auswertung von alltäglich erhobenen biologischen und Verhaltensdaten ermöglichen die Ermittlung von Risikoprofilen für zukünftige Erkrankungen, wobei es sich in aller Regel um die Angabe von Wahrscheinlichkeiten handelt. Im Falle genetischer Daten kann dies auch Auswirkungen auf Angehörige haben.

Vor diesem Hintergrund sollte die Bundesregierung die **Aufnahme eines korrespondierenden Tatbestandes innerhalb des Allgemeinen Gleichbehandlungsgesetzes** (AGG) sowie darüber hinaus spezifische **Verwertungsverbote** von Informationen über die Gesundheit einer Person – wie sie für genetische Informationen schon im Gendiagnostikgesetz festgelegt sind – prüfen.

4.2 Anonymisierung, Pseudonymisierung und synthetische Daten

Jeder Zugang zu personenbezogenen Daten hat sich in den Grenzen des geltenden Datenschutzrechts zu bewegen und muss sich an den dort statuierten Anforderungen an eine Datenverarbeitung – vom Zweckbindungsgrundsatz bis hin zu angemessenen Schutzmaßnahmen – messen lassen. Es ist daher für ein Unternehmen oder andere Anwender gegebenenfalls von ausschlaggebender Bedeutung, sicher sein zu können, sich entweder außerhalb des Anwendungsbereichs des Datenschutzrechts zu bewegen oder jedenfalls datenschutzkonform zu arbeiten. Einen Bedarf nach mehr **Rechtsicherheit** sieht die DEK dabei beispielsweise zu Fragen der Anonymisierung und Pseudonymisierung von Daten, zum Erkennen und Berücksichtigen des Personenbezugs von (vermeintlich anonymen) Datenbeständen und zu sog. synthetischen Daten.

Anonymisierte und pseudonymisierte Daten

Bei der **Anonymisierung** handelt es sich um eine Verarbeitung, die aus einem Bestand personenbezogener Daten den Personenbezug unwiederbringlich entfernt. Man unterscheidet zwei Anonymisierungsansätze, die sich einzeln oder kombiniert verwenden lassen: die **Randomisierung** und die Generalisierung. Unter Randomisierung versteht man eine Veränderung der Daten derart, dass eine Zuordenbarkeit zwischen anonymisierten Daten und der betroffenen Person nicht mehr gegeben ist. Dies kann beispielsweise dadurch erreicht werden, dass einzelne Datensätze verfälscht werden. Bei geeigneter Gestaltung der Randomisierung bleiben die statistischen Eigenschaften des ursprünglichen Datenbestands erhalten, z. B. wenn Werte nur vertauscht und nicht verändert werden. **Generalisierung** bezeichnet eine Vergröberung von Daten, beispielsweise durch Aggregation von detaillierten Einzelangaben wie Altersgruppen statt Geburtsdaten, Regionenbezeichnungen statt Postleitzahlen, Zeiträume statt sekundengenauem Zeitstempel.

Um einen Personenbezug in einem Datenbestand aufzufinden, sind drei Strategien wesentlich:

a) **Herausgreifen** („singling out"): Darunter versteht man die Möglichkeit, aus einem Datenbestand Datensätze zu einzelnen Personen zu isolieren, beispielsweise mit Hilfe singulärer Merkmale, mit denen einzelne Personen identifiziert werden können;

b) **Verknüpfbarkeit** („linkability"): Hiermit ist die Möglichkeit gemeint, mindestens zwei Datensätze, die dieselbe Person oder Personengruppe betreffen, mit Hilfe übereinstimmer Werte wie z.B. Kennungen, räumliche Koordinaten oder Zeitangaben zu verknüpfen. Diese Verknüpfung ermöglicht die Anreicherung der Daten zu derjenigen Person, zu der bislang weniger Daten vorhanden waren, und kann auf diese Weise zu einer Identifizierung dieser Person führen;

c) **Inferenz** („inference"): Darunter versteht man die Möglichkeit, den Wert eines Merkmals mit einer signifikanten Wahrscheinlichkeit von den Werten einer Reihe anderer Merkmale abzuleiten. Eine solche Ableitung ermöglicht ebenfalls eine Anreicherung der Daten zu einer Person und erhöht die Wahrscheinlichkeit eines Personenbezugs.

Ein anonymisierter Datenbestand ermöglicht – bezogen auf den Zeitpunkt der Beurteilung und die technologischen Möglichkeiten, die nach allgemeinem Ermessen wahrscheinlich genutzt werden (vgl. Erwägungsgrund 26 zur DSGVO) – keine (Wieder-)Herstellung eines Personenbezugs (sog. De-Anonymisierung) und bietet einem Angreifer, der auf die (Wieder-)Herstellung des Personenbezugs von einzelnen oder allen betroffenen Personen zielt, keinen ausreichenden Ansatzpunkt.

Sorgt man durch Veränderungen des Datenbestands, insbesondere durch das künstliche Hinzufügen von Unschärfen (je nach Kontext auch „noise" oder „blurring" genannt), dafür, dass Datensätze zu einer Person nicht herausgegriffen werden können, dass auf verkettungsermöglichende Daten verzichtet wird und dass keine Inferenzen gezogen werden können, beschränkt diese Veränderung in der Regel die Nutzbarkeit (sog. utility) der Daten. Sofern man die später gewünschten Auswertungen eines Datenbestands kennt, kann die Anonymisierung dafür optimiert werden, beispielsweise um den nötigen Detaillierungsgrad der Daten in den betroffenen Merkmalen nach Möglichkeit zu erhalten. Dasselbe gilt für den Vergleich verschiedener Datenbestände im Sinne einer Interoperabilität. Ist dieses Ziel bekannt, kann die Anonymisierung durch geeignete gleiche Gruppierungen und durch die Berücksichtigung möglicher Zusatzrisiken durch Informationen aus den weiteren Datenbeständen entsprechend gestaltet werden.

Bei der **Pseudonymisierung** lässt sich der resultierende Datenbestand ohne zusätzliche Informationen nicht mehr einer spezifischen betroffenen Person zuordnen. Diese zusätzlichen Informationen sind beispielsweise Zuordnungstabellen oder kryptographische Hash-Verfahren. Im Gegensatz zur Anonymisierung bleibt ein Personenbezug im Rechtssinne bestehen. Der Verantwortliche muss dafür Sorge tragen, dass diese zusätzlichen Informationen bei der weiteren Verarbeitung des pseudonymisierten Datenbestands besonders gegen einen (unberechtigten) Zugriff gesichert werden, da sich dadurch der Personenbezug herstellen lässt. Die DSGVO sieht die Pseudonymisierung als eine technisch-organisatorische Maßnahme zur Reduzierung des Risikos für die Rechte und Freiheiten natürlicher Personen und erwähnt sie an zahlreichen Stellen.

Anonymisierung und Pseudonymisierung sind jeweils Verarbeitungen von vorhandenen Daten. Davon abzugrenzen ist die **Verwendung von Pseudonymen** durch den Nutzenden. Dies kann durch bewusst selbstgewählte Kennungen (z. B. User-Namen bei Online-Diensten oder E-Mail-Adressen) geschehen; es können aber auch technisch berechnete Kennungen zum Einsatz kommen, z. B. bei der Online-ID-Funktion des elektronischen Personalausweises oder bei der Nutzung von datenschutzfördernden attributbasierten Berechtigungszertifikaten. Bei der Verwendung von Pseudonymen ist sehr häufig kein großer Schutz gegen das Herstellen eines Personenbezugs gegeben, insbesondere bei einem kontext- und kommunikationspartnerübergreifenden Einsatz, was eine Verkettung und Anreicherung von Daten in einem nutzerbezogenen Profil ermöglicht. Im Gegensatz dazu bieten ständig wechselnde und auf den jeweiligen Kontext beschränkte Transaktionspseudonyme einen größeren Schutz gegen eine Identifizierung.

Die Verfahren, die auf eine **Verschleierung des Personenbezugs** im Internet zielen, leisten zumeist keine wirkliche Anonymisierung, aber können dennoch zum Schutz gegen Identifizierung und Beobachtung beitragen. Simple Web-Proxy-Verfahren erlauben eine Nutzung des Internets mit der Kennung (hier: der IP-Adresse) eines zwischengeschalteten Servers, so dass für die angesurften Webserver die Zugriffe mehrerer Nutzender gleich aussehen, sofern diese nicht durch Cookies o.ä. weitere Identifikatoren mitsenden. Der Proxy-Server selbst hat in diesem Fall jedoch Kenntnis von den Kennungen der Nutzenden. In Verfahren mit einem größeren Schutz vor einer Identifizierung können solche Zwischenrechner hintereinander geschaltet werden, beispielsweise in Mix-Netzen wie Tor oder in Mix-Kaskaden wie JonDo. Auch hier kann durch eine zusätzliche Verrauschung mit künstlich erzeugtem „Dummy Traffic" eine Beobachtung der menschlichen Nutzenden erschwert werden.

4.2.1 Verfahren, Standards und Vermutungsregeln

Eine **Anonymisierung**, also die vollständige und nicht rückführbare Befreiung der Daten von jeglichem Personenbezug, unter Beibehaltung der maximalen Aussagekraft (utility) der Daten, ist vielfach faktisch ausgeschlossen. Sie ist allerdings auch oft nicht notwendig, da einerseits viele Zwecke bei genauerer Prüfung auch mit einer leicht geringeren Aussagekraft verfolgt werden können, andererseits bei der Verarbeitung im öffentlichen Interesse, zum Beispiel zum Zweck der Forschung, die DSGVO schon Ausnahmen vorsieht, die eine Verarbeitung auch von personenbezogenen Daten ohne Einwilligung erlauben. Gleichwohl gilt es, die Bemühungen um wirksame **Anonymisierungstechnologien und -verfahren** zu intensivieren, die eine Datenverarbeitung ganz außerhalb des Anwendungsbereichs der DSGVO ermöglichen.

Rechtssicherheit lässt sich letztlich nur im Wege der Entwicklung **standardisierter Technologien und Verfahren** gewährleisten, die gleichwohl stets die sich mit hoher Geschwindigkeit vollziehende technologische Entwicklung berücksichtigen müssen. Die DEK empfiehlt daher der Bundesregierung, im Interesse sowohl der betroffenen Personen als auch der Anwender insbesondere auf EU-Ebene auf die Entwicklung handhabbarer **Standards für Anonymisierung** zu dringen. Gleiches gilt für **Pseudonymisierungsmaßnahmen**, die der Risikolage für die Privatsphäre angemessen sind und wie sie im Rahmen des Digital-Gipfels der Bundesregierung derzeit bereits erarbeitet werden.

Anonymisierungsstandards sollten insbesondere klare Regeln für eine **gesetzliche** widerlegliche **Vermutung** einschließen, die dem Anwender Rechtssicherheit vermitteln, nicht dem Anwendungsbereich der DSGVO unterworfen zu sein. Hierbei ist zu berücksichtigen, dass in diesen Vermutungsregeln gegebenenfalls Einschränkungen zu definieren sind, beispielsweise in der zeitlichen Gültigkeit, wie dies auch im Bereich kryptographischer Verfahren der Fall ist,[24] oder in den zugelassenen Verarbeitungsformen, z. B. dass keine Veröffentlichung oder Zugänglichmachung gegenüber einer unbestimmten Zahl von Personen erfolgen darf. Solange keine rechtlichen Grundlagen für widerlegliche Vermutungsregeln bestehen, sollte die Entwicklung von technischen **Best-Practice-Verfahren** und die Erarbeitung von branchenspezifischen **Selbstverpflichtungen** (Codes of Conduct) unterstützt werden, um Erfahrungen zu gewinnen.

Eine Standardisierung der Verfahren zur Anonymisierung und Pseudonymisierung kann darüber hinaus in bestimmten Bereichen Regeln zur Entfernung des Personenbezugs vorgeben, die eine Vergleichbarkeit von verschiedenen Datenbeständen ermöglichen und damit die **Interoperabilität** verbessern. Die DEK empfiehlt, zumindest in den Bereichen, in denen bessere Interoperabilität gewünscht ist, kontextspezifische Regeln für die zu wählenden Gruppierungen (wie z. B. Wertebereiche von Altersgruppen, Postleitzahlen, IP-Adressen) zu spezifizieren. Von den Statistikämtern wird dies in Bezug auf ihre Datenbestände bereits jetzt so gehandhabt.

24 Bundesamt für Sicherheit in der Informationstechnik: Technische Richtlinie BSI TR-02102 Kryptographische Verfahren: Empfehlungen und Schlüssellängen, letzte Version von Februar 2019 (abrufbar unter: *https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/TechnischeRichtlinien/TR02102/BSI-TR-02102.pdf?__blob=publicationFile*).

Verfahren der Anonymisierung und der Pseudonymisierung beziehen sich auf Datenbestände, deren Personenbezug bekannt ist oder zumindest vermutet wird. Davon abzugrenzen sind Datenbestände, in denen kein Personenbezug vermutet wird, aber die bereits einzeln oder in Kombination zumindest dazu beitragen können, dass sich ein Personenbezug in vermeintlich anonymen Daten herstellen lässt. Die DEK empfiehlt, dass auch hierfür **standardisierte Prüfmethoden des Personenbezugs** entwickelt und vorgegeben werden, die es dem Anwender ermöglichen, eine Personenbeziehbarkeit festzustellen oder in ausreichendem Maße auszuschließen.

4.2.2 Verbot der De-Anonymisierung

Vermutungsregeln sollten allerdings durch angemessene **strafbewehrte Verbote der De-Anonymisierung** flankiert werden. Dies gilt für den Fall, dass bei bisher anonymen Daten, etwa durch die Entwicklung der Technik, ein Personenbezug hergestellt werden kann. Diese müssten so gefasst werden, dass die Forschung zum Erkennen und Entfernen eines Personenbezugs in Datenbeständen nicht behindert wird, denn zur Entwicklung von geeigneten Anonymisierungsstandards und zum Überprüfen ihrer Wirksamkeit ist es nötig, etwa bestehende Möglichkeiten der De-Anonymisierung weiter zu untersuchen. Auch dürfte die Einführung von strafbewehrten Verboten der De-Anonymisierung nicht dazu verleiten, die für eine Anonymisierung geltenden Standards herabzusetzen oder den Begriff der personenbezogenen Daten im Sinne der DSGVO zu verwässern. Anderenfalls würden auch Wettbewerbsnachteile für diejenigen Akteure entstehen, die sich mit technischen Mitteln um eine Anonymisierung bemühen und diese wichtige Technologie weiterentwickeln. Das Gleiche gilt für die Aufhebung von Pseudonymisierung, sofern diese Aufhebung nicht durch einen zu definierenden Katalog von Gründen gerechtfertigt ist.

4.2.3 Synthetische Daten

Von echten Daten abzugrenzen sind **synthetische Daten**, d. h. Daten, die künstlich generiert und nicht unmittelbar in der realen Welt erhoben wurden. Sie haben mehrere Vorteile gegenüber echten Daten.[25] Erstens lassen sich synthetische Daten in beliebiger Menge produzieren. Dies ist besonders wichtig für Simulationen, da echte Daten hier noch gar nicht angefallen sein können. Zweitens kann bei der Erzeugung synthetischer Daten dafür gesorgt werden, dass das gesamte Wertespektrum möglichst vollständig abgebildet wird, z. B. um das Verhalten eines technischen Systems bei ungewöhnlichen Datenkonstellationen zu testen. Drittens ist die Qualität von synthetischen Daten messbar. Je nach Bedarf im Einzelfall kann gewährleistet werden, dass die Eigenschaften eines Referenzdatenbestands aus der realen Welt erhalten bleiben, oder es lassen sich gezielt Verzerrungen, die in Echtbeständen vorkommen können, zur Vermeidung von Diskriminierungen herausnehmen. Solange der synthetische Datenbestand keinen Personenbezug aufweist, ist er anonym, und die DSGVO ist nicht anwendbar.

Die DEK empfiehlt der Bundesregierung, die **Forschung im Bereich synthetischer Daten zu fördern**. Dabei besteht u.a. Forschungsbedarf zu der Frage, inwieweit und in welchen Kontexten synthetische Daten die Verarbeitung echter Daten ersetzen können und wie eng die synthetischen Daten an die Eigenschaften von echten Daten angelehnt sein sollen. Die DEK empfiehlt, die Erzeugung und den Einsatz synthetischer Daten weiter zu untersuchen, beispielsweise im Hinblick auf ihre Datenqualität und auf die Vermeidung von Verzerrungen (Bias) und Diskriminierung.

25 Jörg Drechsler / Nicola Jentzsch: Synthetische Daten: Innovationspotenzial und gesellschaftliche Herausforderungen, Stiftung Neue Verantwortung, 2018 (abrufbar unter: *https://www.stiftung-nv.de/sites/default/files/synthetische_daten.pdf*).

4.3 Kontrollierter Datenzugang durch Datenmanagement- und Datentreuhandsysteme

4.3.1 Privacy Management Tools (PMT) und Personal Information Management Systems (PIMS)

Defizite in der Befähigung zur Wahrnehmung von Datenrechten werden vor allem in Bezug auf die **Kontrolle Einzelner über ihre personenbezogenen Daten** in einer zunehmend komplexer werdenden Umgebung gesehen. In der Regel verfügen betroffene Personen etwa über keine Dokumentation erteilter Einwilligungen. Das Teilen von Daten durch den ursprünglich Verantwortlichen kann zudem zu einer Streuung führen, die für betroffene Personen die bestehenden Intransparenzen und damit verbundenen Datenschutzrisiken noch deutlich erhöht (→ siehe zum Problem des Datenhandels oben). Es fehlt derzeit an hinreichenden Standards und Softwarewerkzeugen, mithilfe derer betroffene Personen etwaige Datenzugangsbeteiligungen und Datenweitergaben fortlaufend nachverfolgen und steuern und damit ihre Datenrechte effektiv wahrnehmen können.

Zur Lösung des Problems werden verstärkt technische und institutionelle Maßnahmen vorgeschlagen. Dazu gehören diverse sog. **Privacy Management Tools (PMT)**, die von Applikationen zur Erleichterung der nutzerseitigen Einwilligungsverwaltung (Dashboards etc.) bis hin zu KI-Tools, die individuelle Nutzerpräferenzen automatisch umsetzen („Datenagenten"), rangieren. Stehen nicht die Herstellung und der Support technischer Applikationen im Vordergrund, sondern Dienstleistungen, wird eher von **Personal Information Management Systems (PIMS)** gesprochen. Sie können von Single-Sign-On-Diensten über lokale Datensafes und Online-Speichersysteme bis zu mehr oder weniger umfassender Fremdverwaltung von Daten der Nutzenden (sog. Datentreuhand-Modelle) reichen. In der zuletzt genannten Variante können PIMS die digitale Selbstbestimmung unterstützen, indem sie die Ausübung von datenschutzrechtlichen Betroffenenrechten, wie das Erteilen und Widerrufen von Einwilligungen und die Wahrnehmung der Rechte auf Auskunft, Berichtigung, Löschung, Datenübertragbarkeit und Widerspruch teilweise für den Betroffenen übernehmen. Die DEK empfiehlt der Bundesregierung die **Förderung** von Innovationen und Standardisierungen für derartige Softwarewerkzeuge und Dienstleistungen.

4.3.2 Bedarf nach Regulierung von PMT/PIMS

Allerdings können von PMT/PIMS auch **Gefahren** ausgehen, wenn sie bestimmten, teilweise auch über die DSGVO hinausgehenden Anforderungen nicht genügen. So besteht bei fehlerhafter Ausgestaltung von PMT/PIMS die Gefahr, dass statt der Ermöglichung echter Selbstbestimmung betroffene Personen auf einen Weg der unbewussten oder sorglosen **Fremdbestimmung** geführt werden. Insbesondere würde es dem ethischen Wert der Selbstbestimmung letztlich widersprechen, wenn PMT/PIMS so ausgestaltet werden, dass Entscheidungen weitgehend (z. B. durch Blankomandate) von betroffenen Personen an die Betreiber von PMT/PIMS abgegeben oder Entscheidungen betroffener Personen durch diese interessenwidrig beeinflusst werden. PMT/PIMS müssen den betroffenen Personen als Hilfsmittel dienen, dürfen deren selbstbestimmte Entscheidungshoheit jedoch nicht ersetzen oder diese gar durch sog. Dark Patterns o.ä. (→ oben 3.2.2.) manipulieren.

Aufgrund der erhöhten Grundrechtsrelevanz und der fehlenden Möglichkeiten einer Qualitätskontrolle durch den Betroffenen selbst empfiehlt die DEK der Bundesregierung die Erarbeitung von **Qualitätsstandards für PMT/PIMS und die Einführung eines Zertifizierungs- und Überwachungssystems**. Letzteres sollte insbesondere für Systeme gelten, die im Namen der betroffenen Personen bzw. an ihrer Stelle agieren oder durch ihre technische Gestaltung die Entscheidungen der betroffenen Personen wesentlich steuern und kanalisieren. Sofern Daten unmittelbar durch die Betreiber von PMT/PIMS gespeichert werden (im Gegensatz zur ebenso möglichen dezentralen Speicherung und bloßen Verwaltung), bedarf es auch Vorkehrungen für den Fall der Insolvenz oder Auflösung.

PMT/PIMS können nur dann verlässlich arbeiten, wenn eine Kooperation mit allen betroffenen Verantwortlichen sichergestellt ist. Dabei ist eine hinreichende Breitenwirkung nur durch eine – unter sachgerechten Bedingungen stehende – **rechtliche Verpflichtung** für Verantwortliche im Sinne der DSGVO zu erreichen, die Kontrolle des Zugangs zu personenbezogenen Daten durch PMT/PIMS zu ermöglichen und beispielsweise sicherzustellen, dass jede datenschutzrelevante Information das PMT/PIMS erreicht und das PMT/PIMS in Bezug auf alle personenbezogenen Daten die Interessen der betroffenen Person wahrnehmen kann. Realistisch erscheint dabei zunächst ein **sektorspezifisches Vorgehen**, welches etwa für soziale Netzwerke zu erwägen ist.

Nach Ansicht der DEK kann der Betrieb derartiger Systeme entweder ohne Erwerbsabsicht, etwa durch **gemeinwohlorientierte Stiftungen** und ähnliche unabhängige Stellen und ohne jede Beteiligung von kommerziell motivierten Akteuren erfolgen, oder **privatwirtschaftlich** organisiert sein, wenn dabei der Betreiber an der Verwaltung, und nicht an der Nutzung der Daten verdient. In jedem Fall ist es notwendig, die besonderen Treuepflichten gegenüber der betroffenen Person gesetzlich präzise zu fassen, Akteure mit konfligierenden Interessen auszuschließen und insgesamt entsprechende Kontrollmöglichkeiten – etwa auch zur Minimierung von Bias und Diskriminierung – einzubauen. Bei den privatwirtschaftlichen Modellen muss auch sichergestellt werden, dass eine Unterminierung der Funktion als Interessenwalter der betroffenen Person bei einer Erwerbsabsicht ausgeschlossen ist. Betreiber, die Zugriff auf personenbezogene Daten erhalten, müssen ihren Sitz in der Europäischen Union haben.

Die DEK empfiehlt der Bundesregierung, auf eine entsprechende **Ergänzung der DSGVO** hinzuwirken, die einen konkretisierenden und rechtssicheren Rahmen für PMT/PIMS vorgibt. Über rechtliche Fragen zu Mandaten usw. hinaus ist einer übermäßigen zentralen Speicherung personenbezogener Daten entgegenzuwirken, die z.B. im Fall von Cyberangriffen Risiken für die Betroffenen erhöhen würde. Für eine automatisierte Realisierung der Dienste sind maschineninterpretierbare Formate und Kommunikationsprotokolle zu standardisieren.

4.3.3 PMT/PIMS als mögliche Schnittstelle zur Datenwirtschaft

Bei entsprechender Regulierung könnten PMT/PIMS auch doppelfunktional tätig werden. So könnte einerseits der Einzelne mit Hilfe von PMT/PIMS sein Recht auf informationelle Selbstbestimmung wirksam ausüben und Zweckbegrenzungen zuverlässig überprüfen, andererseits könnten aber auch – insbesondere unter Nutzung des Portabilitätsrechts aus Art. 20 DSGVO – Daten aus „Datensilos" geholt und für die europäische Datenwirtschaft freigesetzt werden. Der Grundgedanke von PMT/PIMS betrifft zwar zunächst nur die Verbesserung von Kontrolle des Einzelnen über seine personenbezogenen Daten. Dies beinhaltet an sich nicht, Datenzugang durch Dritte zu fördern. Eine mittelbare Datenzugangsfunktion wäre mit dem **Gedanken der treuhänderischen Verwaltung** aber dann zu vereinbaren, wenn sich der Datenzugang durch Dritte entweder als vom Betroffenen gewollte Datenfreigabe zur Förderung bestimmter Zwecke darstellt (→ etwa im Forschungskontext, oben) oder aber der wirtschaftlichen Verwertung der Daten im Interesse des Betroffenen dient und mit seiner ausdrücklichen Zustimmung erfolgt (→ zur Problematik der Ökonomisierung personenbezogener Daten, oben).

Sofern man eine zusätzliche Funktion von PMT/PIMS als Plattform für den rechtssicheren Datenzugang für Unternehmen anerkennen will, muss nach Auffassung der DEK sichergestellt werden, dass diese qualifizierten PMT/PIMS die Schutzfunktion der Betroffenenrechte nicht letztlich in ihr Gegenteil verkehren. Es bedarf der strengen Einhaltung der Grundsätze von Privacy und Ethics by Design. Insbesondere darf der Zweck nicht auf die möglichst weitgehende Datenverwertung und -streuung ausgerichtet sein. Die DEK betont, dass PMT/PIMS nicht ihre Funktion als eindeutige Interessenwalter der Betroffenen verlieren dürfen und ein **Interessenkonflikt ausgeschlossen** werden muss.

4.4 Datenzugang durch Datenportabilität

4.4.1 Förderung von Datenportabilität

Das Portabilitätsrecht in Art. 20 DSGVO (**Recht auf Datenübertragbarkeit**) stellt ein Instrument für den Betroffenen dar, in Bezug auf personenbezogene Daten, die der Betroffene einmal einem Unternehmen bereitgestellt hat, selbstbestimmt zu entscheiden, welche weiteren Unternehmen Zugang zu diesen Daten erhalten sollten. Es umfasst das Recht, die bereitgestellten Daten in einem „strukturierten, gängigen und maschinenlesbaren Format" zu erhalten oder einem anderen Verantwortlichen direkt übermitteln zu lassen. Datenportabilität wirkt vor allem in zwei Richtungen:

a) Bei einem Anbieterwechsel verhindert das Portabilitätsrecht unerwünschte Lock-in-Effekte und schützt damit sowohl den einzelnen Betroffenen in seiner wirtschaftlichen Dispositionsfreiheit als auch den freien Wettbewerb.

b) Unabhängig von einem Anbieterwechsel erlaubt das Portabilitätsrecht dem Betroffenen, seine Daten von einem Verantwortlichen heraus zu verlangen und anderen Unternehmen zur Verfügung zu stellen. Diese anderen Unternehmen erlangen dadurch – sofern sie zugleich eine eigene datenschutzrechtliche Rechtsgrundlage (z. B. Einwilligung oder Vertrag) für die Verarbeitung haben[26] – einen Datenzugang, den sie auf anderem Wege möglicherweise nicht erlangt hätten.

Die Anforderungen an ein „strukturiertes, gängiges und maschinenlesbares Format" werden in der Praxis bislang noch sehr unterschiedlich und uneinheitlich ausgelegt, obwohl sie Grundvoraussetzung für eine wirkungsvolle Ausübung des Portabilitätsrechts sind. Daher empfiehlt die DEK der Bundesregierung und den Datenschutzbehörden, in Anlehnung an Erwägungsgrund 68 zur DSGVO auf europäischer Ebene die Entwicklung **branchenbezogener Verhaltensregeln und Standards** zu unterstützen, damit Datenübertragbarkeit im Interesse aller Beteiligten einheitlich und praktisch wirkungsvoll umgesetzt werden kann.

Ohne Hinzutreten neuer Intermediäre (→ oben 4.3) dürfte die Anregung zur Ausübung des Portabilitätsrechts oft durch das Unternehmen, das einen neuen Kunden gewonnen hat, erfolgen. Dabei werden diejenigen Unternehmen besonders erfolgreich sein, die eine bequeme und automatisierte Ausübung des Portabilitätsrechts ermöglichen (z. B. Anbieter eines Kartendienstes ermöglicht per Klick das Portieren von Daten eines Mobilitätsdienstleisters). Aufgrund von Netzwerk- und Skaleneffekten besteht Grund zur Annahme, dass – zumindest mittelfristig – gerade daten- und marktmächtige Unternehmen die größten Profiteure des Portabilitätsrechts sein könnten. Der Bundesregierung wird daher empfohlen, die Entwicklungen **aufmerksam zu beobachten** und, soweit erforderlich, auf europäischer Ebene auf Maßnahmen zu dringen, die eine erleichterte Portabilität für die betroffenen Personen speziell von daten- und marktmächtigen Unternehmen zu anderen Marktteilnehmern, einschließlich Start-ups, fördern.

26 Zum Erfordernis einer derartigen eigenen datenschutzrechtlichen Rechtsgrundlage siehe etwa Art. 29-Datenschutzgruppe: Leitlinien zum Recht auf Datenübertragbarkeit, WP 242, rev. 01, S. 7 f.

4.4.2 Erweiterung des Portabilitätsrechts?

Derzeit wird eine Erweiterung des Portabilitätsrechts in verschiedener Hinsicht diskutiert, insbesondere was die Erweiterung auf andere als bereitgestellte (Roh-)Daten (z. B. auf bestimmte veredelte bzw. abgeleitete Daten) sowie ein Recht auf dynamische Echtzeit-Portabilität (d. h. Echtzeit-Streaming von Datenflüssen) betrifft. Die DEK legt der Bundesregierung im Sinne der soeben formulierten Empfehlung nahe, **derzeit auf keine rechtliche Änderung zur Erweiterung des bestehenden Portabilitätsrechts zu dringen**, da seine praktische Anwendung, die Aufsichtspraxis der Datenschutzbehörden und auch die Auslegung der DSGVO durch Gerichte so kurze Zeit nach Wirksamkeit des neuen Rechts zunächst abgewartet werden sollte.

4.4.3 Von Portabilität zu Interoperabilität und Interkonnektivität

Aufgrund von Netzwerkeffekten (z. B. bei Messenger-Diensten) dürfte die Datenportabilität allein nicht ausreichen, entstandenen und drohenden Daten- und Service-Oligopolen entgegenzuwirken und die Markteintrittshürden für neue Wettbewerber soweit zu senken, dass dominante Anbieter ernsthaft herausgefordert werden. Die DEK empfiehlt der Bundesregierung daher, auf die Einführung **sektorspezifischer Pflichten zur Interoperabilität** hinzuwirken, wie dies z. B. auch früher bei Postdienstleistungen und im Mobilfunk realisiert wurde. Dabei muss eine datenschutzkonforme Gestaltung der Interoperabilität einschließlich datenschutzfreundlicher Voreinstellungen gewährleistet sein, beispielsweise durch die Möglichkeit der Verwendung unterschiedlicher und wechselnder Kennungen statt nur eines übergreifenden Identifikators, durch Reduktion der Datensammelmöglichkeiten an zentralen Komponenten oder durch sonstige geeignete Realisierung auf verschiedenen Schichten im interoperablen technischen Zusammenwirken.

Diese Interoperabilitätsverpflichtungen könnten **asymmetrisch zwischen marktmächtigen Unternehmen und neuen Marktteilnehmern ausgestaltet** werden (z. B. wäre dann ein marktmächtiger Anbieter von Messenger-Diensten verpflichtet, den Kunden kleinerer Anbieter das unmittelbare Versenden von Nachrichten an seine Kunden und umgekehrt das Empfangen von deren Nachrichten zu ermöglichen). Jedenfalls ist dafür Sorge zu tragen, dass nicht über die Interoperabilitätsanforderungen ein umso stärkerer Fluss personenbezogener Daten hin zu daten- und marktmächtigen Unternehmen entsteht. Sofern dies gewährleistet werden kann, wäre es sinnvoll, etwa eine **Interkonnektivitätsverpflichtung für Kurznachrichtendienste und soziale Netzwerke** vorzusehen, um so den Konzentrationseffekten der Netzwerke entgegen zu wirken und dem Ziel der Datenportabilität, nämlich Wettbewerb und Markteintritt in der datenintensiven Wirtschaft zu fördern, noch intensiver zu dienen. Dies wäre auch eine Voraussetzung dafür, im Sinne der digitalen Souveränität Deutschlands bzw. Europas bestimmte Basisdienstleistungen der Informationsgesellschaft in Europa neu aufzubauen bzw. zu stärken.

4.5 Crowd Sensing zu gemeinwohlorientierten Zwecken

Auch das sog. Crowd Sensing will neue Datenressourcen für Datengesellschaft und Datenwirtschaft erschließen und verwendet dazu die technischen Geräte der Nutzenden als Sensoren. Diese erheben etwa in einem bestimmten Ortsbereich Daten und leiten sie an eine übergeordnete Instanz weiter, die die gesammelten Daten auswertet. Die DEK sieht die Potenziale, die diese Technologie mit sich bringen kann, insbesondere sofern ihr **Einsatz gemeinwohlorientierten Zwecken** dient. So kann Crowd Sensing beispielsweise in der Smart City für Echtzeit-Analysen der Verkehrslage, des Zustands der Infrastrukturen, der Luftqualität etc. genutzt werden. Zugleich sieht die DEK aber erhebliche Herausforderungen für eine ethisch angemessene Ausgestaltung. Denn die durch Crowd Sensing ermöglichten Analysen weisen typischerweise eine extrem hohe Granularität auf und können daher aus Sicht derjenigen, die die Daten beisteuern, sowie gegebenenfalls auch für Personen in ihrer Umgebung überaus sensibel sein. Um eine unerwünschte Rückführbarkeit auf die Nutzenden sowie möglicherweise betroffene weitere Personen auszuschließen oder anderweitigen Missbrauch zu vermeiden, bedarf es daher auch hier verstärkter Bemühungen um **Standards der Anonymisierung und Pseudonymisierung** (→ oben 4.2). Hinzu kommt, dass die im Zuge von Crowd Sensing getätigten Datenübertragungen die Ressourcen der Geräte der Nutzenden beeinträchtigen können und Sicherheitsfragen aufwerfen (→ unten Teil F, 8.3).

Diese Gesichtspunkte sind auch zu beachten, wenn sich die Nutzenden freiwillig und bewusst an Crowd-Sensing-Programmen beteiligen (sog. Participatory Sensing). Insoweit ist an die **materiellen Schranken der Einwilligung** zu erinnern (→ oben 3.2). Insgesamt muss auch beim Einsatz zu gemeinwohlorientierten Zwecken stets sichergestellt sein, dass die rechtlichen Vorgaben, insbesondere des Datenschutz- und des Verbraucherschutzrechts, vollumfänglich gewahrt bleiben. In diesem Fall ist ferner zu berücksichtigen, dass staatliche Entscheidungen und Maßnahmen regelmäßig nicht allein auf mittels Participatory Sensing gesammelte Daten gestützt werden dürfen, denn diese Daten sind durch die Freiwilligkeit der Teilnahme zwangsläufig **unvollständig und wahrscheinlich verzerrt**.

Sofern erörtert wird, ob personenbezogene Daten im Wege des Crowd Sensing ohne Kenntnis der Nutzenden erhoben, weitergeleitet und gesammelt werden dürfen (sog. Opportunistic Sensing), verstößt dies aus Sicht der DEK potenziell gegen elementare Grundsätze des Datenschutzes. Ob sich ein **gesetzlicher Zwang** zur Bereitstellung der eigenen technischen Geräte für die automatische Erhebung und Weiterleitung von Daten rechtfertigen lässt, wenn und soweit deren Analyse wichtigen Interessen des Gemeinwohls dienen kann, kann aus Sicht der DEK nur im konkreten Einzelfall entschieden werden.

Zusammenfassung der wichtigsten Handlungsempfehlungen

Verbesserung des kontrollierten Zugangs zu personenbezogenen Daten

16

Die DEK sieht in einer Datennutzung für gemeinwohlorientierte Forschungszwecke (z. B. zur Verbesserung der Gesundheitsfürsorge) enormes Potenzial, das es zum Wohle des Einzelnen und der Allgemeinheit zu nutzen gilt. Das geltende Datenschutzrecht erkennt dieses Potenzial durch eine Reihe weitreichender Privilegierungen prinzipiell an. Allerdings bestehen auch Unsicherheiten, insbesondere mit Blick auf die Reichweite des sog. Weiterverarbeitungsprivilegs sowie des Forschungsbegriffs im Zusammenhang mit der Entwicklung von Produkten. Dem muss aus Sicht der DEK durch entsprechende **gesetzliche Klarstellungen** begegnet werden.

17

Die Zersplitterung der Rechtslage, sowohl innerhalb Deutschlands als auch der EU Mitgliedstaaten untereinander, kann ein Hindernis für datengetriebene Forschung darstellen. Empfohlen wird daher eine **Harmonisierung der forschungsspezifischen Regelungen** sowohl auf Bundes- und Landesebene als auch der verschiedenen nationalen Regelungen innerhalb der EU. Auch die Einführung eines Notifizierungsverfahrens für mitgliedstaatliche Regelungen zum Forschungsdatenschutz sowie die Einrichtung einer europäischen Clearing-Stelle für grenzüberschreitende Forschungsprojekte könnte eine Erleichterung bringen.

18

Bei Forschung mit besonders sensiblen Kategorien personenbezogener Daten (z. B. Gesundheitsdaten) sollten Forschende durch **Handreichungen** zur rechtssicheren Einholung von Einwilligungen sowie durch die Förderung und gesetzliche **Anerkennung innovativer Einwilligungsmodelle** unterstützt werden. Zusätzlich zu den weiteren Entwicklungen zur Reichweite des sog. Weiterverarbeitungsprivilegs für die Forschung könnten dazu auch digitale Einwilligungsassistenten oder ein sog. Meta Consent gehören.

19

Die DEK unterstützt prinzipiell die Entwicklung in Richtung eines **„lernenden Gesundheitssystems“**, in dem die Daten aus der alltäglichen Gesundheitsversorgung systematisch und qualitätsgestützt im Sinne der evidenzbasierten Medizin genutzt werden, um die Versorgung kontinuierlich zu verbessern. Allerdings sollte flankierend, beispielsweise durch **Verwertungsverbote**, mehr Schutz vor dem erheblichen Diskriminierungspotenzial sensibler Datenkategorien geschaffen werden.

20

Im Zentrum aller Bemühungen um eine Verbesserung des kontrollierten Zugangs zu (ursprünglich) personenbezogenen Daten steht die Entwicklung von Verfahren und Standards der **Anonymisierung** und **Pseudonymisierung**. Durch rechtliche Vermutungen, dass bei Einhaltung des Standards kein Personenbezug

mehr gegeben ist bzw. dass „geeignete Garantien“ für die Rechte betroffener Personen vorliegen, könnte die Rechtssicherheit deutlich verbessert werden. Diese Maßnahmen sollten flankiert werden durch strafbewehrte Verbote einer De-Anonymisierung (für den Fall, dass bei bisher anonymen Daten, etwa durch die Entwicklung der Technik, ein Personenbezug hergestellt werden kann) bzw. der Aufhebung der Pseudonymisierung jenseits eng definierter Rechtfertigungsgründe. Auch die Forschung im Bereich **synthetischer Daten** ist vielversprechend und sollte weiter gefördert werden.

21

Großes Potenzial sieht die DEK grundsätzlich auch in **innovativen Datenmanagement- und Datentreuhandsystemen**, sofern diese praxisgerecht, robust und datenschutzkonform ausgestaltet sind. Solche Modelle rangieren von rein technischen Dashboards (**Privacy Management Tools,** PMT) bis hin zu umfassenden Dienstleistungen der Daten- und Einwilligungsverwaltung (**Personal Information Management Services,** PIMS). Ziel ist die Befähigung des Einzelnen zur Kontrolle über seine personenbezogenen Daten sowie die Entlastung des Einzelnen von Entscheidungen, die ihn überfordern. Die DEK empfiehlt, Forschung und Entwicklung im Bereich von Datenmanagement- und Datentreuhandsystemen intensiv zu fördern, mahnt aber auch an, dass eine die Rechte und Interessen aller Beteiligten wahrende Entwicklung ohne eine **begleitende europäische Regulierung** nicht zu erwarten ist. Diese Regulierung müsste zentrale Funktionen absichern, ohne die Betreiber solcher Systeme nur sehr eingeschränkt tätig werden können. Andererseits geht es um den Schutz des Einzelnen vor vermeintlichen Interessenwaltern, die in Wahrheit vorrangig wirtschaftliche Eigeninteressen oder Interessen Dritter vertreten. Sofern dieser Schutz auch in der Praxis garantiert werden kann, kann Datentreuhandmodellen die Funktion einer wichtigen Schnittstelle zwischen Belangen des Datenschutzes und der Datenwirtschaft zukommen.

22

In Bezug auf das Recht auf **Datenportabilität** aus Art. 20 DSGVO empfiehlt die DEK die Erarbeitung branchenbezogener Verhaltensregeln und Standards betreffend Datenformate. Soweit Art. 20 DSGVO nicht nur Anbieterwechsel erleichtern, sondern auch den Datenzugang für andere Anbieter verbessern soll, empfiehlt sich eine sorgfältige Evaluierung, wie sich das bestehende Portabilitätsrecht auf den Markt auswirkt und wie eine zunehmende Stärkung der Marktmacht weniger Anbieter verhindert werden kann. Bevor die Ergebnisse einer solchen Evaluierung vorliegen, sollte von einer vorschnellen Erweiterung des Portabilitätsrechts, etwa auf andere als bereitgestellte Daten oder auf Portierung in Echtzeit, abgesehen werden.

23

Eine **Pflicht zur Interoperabilität bzw. Interkonnektivität** in bestimmten Sektoren – etwa bei Messenger-Diensten und sozialen Netzwerken – könnte dazu beitragen, Markteintrittsbarrieren für neue Anbieter zu senken. Für eine solche Pflicht würde sich eine asymmetrische, d. h. nach Marktmacht gestaffelte Regulierung empfehlen. Dies wäre auch eine Voraussetzung dafür, bestimmte Basisdienstleistungen der Informationsgesellschaft in Europa neu aufzubauen bzw. zu stärken.

5. Datenzugangsdebatten jenseits des Personenbezugs

Der Datenwirtschaft kommt für die künftige Wettbewerbsfähigkeit deutscher und europäischer Unternehmen eine Schlüsselrolle zu. Die zunehmende Verbreitung des Internet of Things (IoT) und des Internet of Services (IoS) hat zu einer wachsenden industriellen Bedeutung von Daten geführt, die durch Sensorik automatisch erhoben werden und zur Entwicklung neuer Geschäftsmodelle und Innovationen beitragen können. **Deutschlands Stärke** in vielen IoT/IoS-relevanten Technologien (z. B. Sensortechnologie, Maschinenbau, eingebettete Systeme) und allgemein in der industriellen Produktion, nebst den verbundenen industrienahen digitalen Dienstleistungen, bedeutet hier eine günstige Startposition, die dazu genutzt werden muss, den Wohlstand in Zeiten zunehmenden weltweiten Wettbewerbs zu sichern. Deutschland verfügt mit seiner differenzierten und leistungsfähigen Forschungslandschaft, seiner breit aufgestellten Wirtschaftsstruktur und seiner Technologieführerschaft in wichtigen Industriefeldern, wie der Industrie 4.0, über eine ausgezeichnete Ausgangslage, um die mit der Datenwirtschaft verbundenen Potenziale für die Wertschöpfung der Zukunft zu nutzen.

5.1 Gesamtwirtschaftliche Bedeutung eines angemessenen Datenzugangs

Die DEK sieht einen wesentlichen Faktor zur Gewährleistung einer marktgerechten und gemeinwohlorientierten Datenwirtschaft und zur Stärkung der digitalen Souveränität Deutschlands und Europas in einem angemessenen Datenzugang deutscher und europäischer Unternehmen und in der Auflösung bestehender Abhängigkeiten von wenigen Datenoligarchen. Dabei bedeutet Datenzugang im engeren Sinne zunächst die Frage, inwieweit für ein bestimmtes Geschäftsmodell oder sonstiges Vorhaben erforderliche Daten **faktisch und rechtlich genutzt werden können**. Datenzugang im engeren Sinne nutzt nur Akteuren, die auch über entsprechendes **Bewusstsein über die Bedeutung von Daten** und entsprechende **Datenkompetenz** verfügen, und in ganz überproportionalem Ausmaß denjenigen, bei denen bereits der größte **Ausgangsbestand** an Daten und die besten **Dateninfrastrukturen** vorhanden sind. Die DEK empfiehlt daher, bei der Diskussion um eine Verbesserung des Datenzugangs stets die genannten anderen Faktoren gemäß dem **ASISA-Prinzip** (Awareness – Skills – Infrastructures – Stocks – Access) zu berücksichtigen.

Der Schwerpunkt der Betrachtung liegt in diesem Abschnitt auf nicht-personenbezogenen Daten. Das Potenzial **genuin nicht-personenbezogener Daten** für Wissenschaft, Wirtschaft und Gesellschaft ist hoch und wird oft unterschätzt. Ein Großteil der Wissenschaftsdaten – angefangen bei den Daten der technischen Wissenschaften (z. B. Ingenieur- und Materialwissenschaften) und der Physik (z. B. die Daten der Teilchenbeschleuniger), über die Daten der Biologie (z. B. Pflanzen- und Tierreich), der Geologie und der Chemie, über Umweltdaten, Wetterdaten und Meeresdaten bis hin zu Wirtschaftsdaten (z. B. Daten der Finanzmärkte) – sind nicht-personenbezogen. Sie haben aber einen großen Wert für Wissenschaft, Wirtschaft und Gesellschaft, wenn sie u.a. mit Big Data-Methoden analysiert und zur Entwicklung von Künstlicher Intelligenz (KI) verwendet werden. Dies sollte gezielt gefördert und der Zugang zur Nutzung solcher Daten sollte systematisch erleichtert werden.

Aufgrund der Weite des Begriffs der personenbezogenen Daten unter der DSGVO ist allerdings auch davon auszugehen, dass ein nicht unerheblicher Teil der Datenbestände gemischter Natur ist und auch Daten beinhaltet, die personenbezogen sind oder werden können. Auch sind bestimmte, dem Einzelnen durchaus nützliche oder gemeinwohlorientierte Aktivitäten der Datenwirtschaft nicht ohne Verarbeitung personenbezogener Daten möglich. Daher erscheint es wenig sinnvoll, im Zusammenhang mit Datenzugang ausschließlich nicht-personenbezogene Daten zu betrachten. Ein sachgerechter Ansatz dürfte vielmehr in einem **allgemeinen Datenzugangsregime** liegen, das nur insoweit, als personenbezogene Daten betroffen sind, **vom Datenschutzrecht überlagert** wird, d. h. datenwirtschaftliche Aktivitäten müssen sich dann zwingend im Rahmen der DSGVO bewegen. Zu betonen ist jedoch, dass die DSGVO bereits heute in vielfacher Weise die wirtschaftliche Verwertung personenbezogener Daten erlaubt. So treten neben die Einwilligung (Art. 6 Abs. 1 lit. a DSGVO) fünf weitere Rechtfertigungstatbestände (Art. 6 Abs. 1 lit. b–f), die teils explizit auf wirtschaftliche Interessen und Bedürfnisse zugeschnitten sind.

5.2 Schaffung der erforderlichen Rahmenbedingungen

5.2.1 Bewusstseinsbildung und Datenkompetenz

Die werteorientierte Nutzung von Daten setzt zunächst bei den Akteuren – ob privat- oder gemeinwohlorientiert – ein hinreichendes Bewusstsein für die bestehenden Möglichkeiten und Risiken sowie hinreichende Datenkompetenz – u.a. in technischer, ökonomischer, ethischer und rechtlicher Hinsicht – voraus (→ oben, Teil D, 3.). Bei deutschen **Unternehmen** besteht bislang teilweise noch ungenutztes Potenzial, die eigenen Datenbestände und Datenströme produktiver und gegebenenfalls gemeinwohlorientierter einzusetzen. Die DEK begrüßt Maßnahmen zur Bewusstseinsbildung und zur Förderung digitaler Kompetenzen seitens diverser Akteure (z. B. der Industrie- und Handelskammern, Verbände oder auch berufsbildende Einrichtungen). Es bedarf einer Verbesserung der werteorientierten Datenkompetenz auf breiter Ebene, was etwa durch entsprechende **Bildungs- und Weiterbildungsangebote** erreicht werden kann. Diese müssen stets auch Sensibilität betreffend datenschutzrechtlicher und ethischer Risiken für das Individuum und die Gesellschaft schaffen.

Staatliche Stellen erkennen erst langsam die Bedeutung und Implikationen der von ihnen bereits heute im großen Umfang, etwa zu Zwecken der Statistik, generierten Daten sowie die Vorteile und Risiken, die der Austausch von staatlichen Daten gegenüber Privaten (sog. Government-to-Business Data Sharing, G2B) oder von Betriebsdaten gegenüber staatlichen Stellen (sog. Business-to-Government Data Sharing, B2G), bieten. Angesichts der bisher eher zurückhaltend genutzten Möglichkeiten ist insoweit auf einen weiteren Wandel der Verwaltungskultur hinzuwirken, wie sie etwa bei eGovernment-Vorreitern wie z. B. den skandinavischen Ländern oder Estland zu finden sind. Die DEK empfiehlt der Bundesregierung zusätzlich, entsprechende Aktivitäten einschlägiger Forschungsinstitutionen zu stärken.

5.2.2 Förderung der Infrastrukturen für eine datenbasierte Ökonomie

Deutschland nimmt zwar nach wie vor eine Spitzenposition in der wissenschaftlichen Technologieforschung ein, jedoch sind es derzeit vor allem amerikanische und zunehmend auch chinesische Technologieunternehmen, die wichtige Daten- und Analyseinfrastrukturen für die neue digitale Wirtschaft bereitstellen. Daher liegen viele europäische Daten – sowohl Konsumentendaten als auch Unternehmens- und Forschungsdaten – außerhalb Europas und werden durch Software nichteuropäischer Unternehmen in Drittländern analysiert. Der Entwicklung **eigener Infrastrukturen** für eine datenbasierte Ökonomie kommt daher eine herausgehobene Rolle zu.

Die DEK empfiehlt der Bundesregierung, die folgenden, von der Europäischen Kommission angestoßenen **Maßnahmen auf europäischer Ebene** zu unterstützen:

a) Einrichtung und weiterer Ausbau des Unterstützungszentrums für die gemeinsame Datennutzung;

b) Erarbeitung von Modellverträgen für die Datenwirtschaft;

c) Förderung von Foren und Konsortien zur Entwicklung von offenen Standards für einen rechtssicheren Datenaustausch, insbesondere von für den Datenaustausch geeigneter Formate und Programmierschnittstellen (APIs) und für die Rückverfolgbarkeit von Datenflüssen;

d) Förderung von europäischen Plattformen für den rechtssicheren Datenaustausch und

e) Einrichtung einer European Open Science Cloud (EOSC).

Als wichtige Voraussetzung für die digitale Souveränität Deutschlands ist die **Zugangskontrolle** für sensible Daten und die Möglichkeit einer ausreichenden **Überprüfung** kritischer Datenanalysesoftware, beispielsweise anhand der Offenlegung von Quellcode und Designkriterien, anzusehen. In geographischer Hinsicht sollte die Durchführung von ethisch sensitiven Analysen daher nach Möglichkeit in unserem Rechtsraum stattfinden.

Die DEK **begrüßt ausdrücklich eine Reihe von Initiativen** der Bundesregierung und anderer Akteure, die darauf gerichtet sind, von Deutschland aus sichere internationale Datenräume für verschiedene Anwendungsdomänen zu schaffen, und die Unternehmen und Organisationen verschiedener Branchen und aller Größen die souveräne Bewirtschaftung und den geregelten Austausch ihrer Datenbestände untereinander ermöglichen.

Zur Unterstützung bei der Aushandlung von Datenzugangsvereinbarungen in schwierigen Fällen und zur Vermittlung bei Streitigkeiten ist ferner die Einrichtung einer **Ombudsstelle** auf Bundesebene empfehlenswert. Soweit personenbezogene Daten betroffen sind, sollte diese die zuständigen Datenschutzbehörden beteiligen, wobei zur Vermeidung divergierender Entscheidungen die Entscheidungshoheit letztlich bei den Datenschutzbehörden liegen müsste.

Aufbau von Dateninfrastrukturen

Zu den **Initiativen der Bundesregierung** betreffend den Aufbau von Dateninfrastrukturen gehören:

a) Die Bemühungen der DFG zum Aufbau einer Nationalen Forschungsdateninfrastruktur (NFDI). In der NFDI sollen Datenbestände in einem aus der Wissenschaft getriebenen Prozess systematisch erschlossen, langfristig gesichert und über Disziplinen- und Ländergrenzen hinaus zugänglich gemacht werden;

b) Das vom BMBF geförderte, offene Konsortium International Data Spaces (IDS, vormals Industrial Data Space). Es stellt für die teilnehmenden Unternehmen und Organisationen eine standardisierte Schnittstelle zu einer Datenaustauschplattform dar, die einem föderalen Architekturkonzept folgt;

c) Die Initiative zum Aufbau eines großen Netzwerkes von Big Data- und KI-Zentren mit über ganz Deutschland verteilten Knoten im Sinne eines nationalen, allgemein zugänglichen Ökosystems. Dieses Netzwerk kann nicht nur eine große Vielzahl und Vielfalt an Daten kontinuierlich bereitstellen, sondern bietet gleichzeitig Werkzeuge der gesamten Datenwertschöpfungskette (Aufbereitung, Analyse, Visualisierung, Verwertung) einfach nutzbar an und entwickelt sie aufgrund der Erfahrung bei deren Nutzung stetig weiter.

Neben diesen technischen Plattformen sind auch die von der Bundesregierung gemeinsam mit Verbänden aufgesetzten Plattformen zur Förderung der koordinierten Forschung und Entwicklung sowie der Standardisierung und praktischen Umsetzung von datenintensiven Anwendungen in Form von gesellschaftlich und wirtschaftlich innovativen Zukunftsprojekten zu nennen, wie etwa die Plattformen Industrie 4.0, Smart Service Welt und Lernende Systeme.

Auf europäischer Ebene führt die **Europäische Kommission** vergleichbare Projekte durch (z. B. das Zukunftsprojekt FIWARE). Sie entwickelt derzeit einen kostenlos verfügbaren Baukasten von Open-Source-Softwarekomponenten, mit denen sich innovative Internetdienste rasch konfigurieren lassen. Die Big Data Value-Public-Private-Partnership (BDVA) hat auf europäischer Ebene ein interoperables und datengetriebenes Ökosystem für neue Geschäftsmodelle auf der Basis von Massendaten mit vielen Leuchtturmprojekten hervorgebracht. Auch im Rahmen des European Institute of Innovation and Technology (EIT Digital) ist europaweit ein technisch-wirtschaftliches Ökosystem mit 180 Unternehmen und Forschungseinrichtungen entstanden.

5.2.3 Nachhaltige und strategische Wirtschaftspolitik

Zu den größten Herausforderungen Europas im Bereich der Datenwirtschaft gehört die häufig mangelnde **Nachhaltigkeit** der Förderung von Forschungsprojekten und das Fehlen von hinreichend **Venture Capital**, um entwickelte Ideen zur Vermarktungsreife zu bringen und in einer Weise mit Kapital auszustatten, dass rechtzeitig der Sprung auf eine wettbewerbsfähige Größe gelingt. Der Erfolg der USA im Bereich digitaler Produkte und Dienstleistungen wurde durch die Bereitschaft vieler Kapitalgeber, Milliardenbeträge in hoch riskante Projekte zu investieren und zu einem nicht unerheblichen Teil auch zu verlieren, begünstigt. Zudem ist zu beobachten, dass innovative Unternehmen von ausländischen Firmen **aufgekauft** oder von internationalen Kapitalgebern zur Sitzverlegung in das außereuropäische Ausland gezwungen werden.

Die **Kapitalausstattung** ebenso wie **steuerliche Anreize** sind für deutsche Start-ups zu verbessern, damit Deutschland – über die von der DEK ausdrücklich befürwortete Strategie des „europäischen Weges“ hinaus (→ unten Teil G) – weiter hinreichende Anziehungskraft auf die innovativsten Köpfe und Ideen ausübt.

Bereiche wie Bildung, öffentliche Verwaltung und Medizin sind gekennzeichnet durch ein hohes öffentliches Interesse und eine Wertebindung, die sich in Recht und Berufsethik manifestiert. Gleichzeitig ist das Potenzial für Effizienzgewinne durch Digitalisierung und KI in diesen Bereichen hoch, ohne dass schon globale, dominante Plattformen im gleichen Ausmaß etabliert sind, wie wir sie bereits in anderen Themenbereichen vorfinden. Es empfiehlt sich vor diesem Hintergrund, gerade in diesen drei Bereichen mit öffentlichen Mitteln gezielt Anreize zur **Entwicklung von Plattformen** in Deutschland zu setzen, die unseren Werten entsprechen und zugleich international skalierbar sind.

5.2.4 Verbesserter Leistungsschutz

Die DEK spricht sich zwar auch unter dem Aspekt der Datenwirtschaft **gegen ein neues Ausschließlichkeitsrecht** an Daten aus (oft unter dem Schlagwort „Dateneigentum“ oder „Datenerzeugerrecht“ diskutiert, → vgl. dazu schon oben 3.3.2). Ein solches Recht, das zu den bestehenden Regelungen wie Datenschutzrecht, Persönlichkeitsrecht, Recht des geistigen Eigentums, Geschäftsgeheimnisschutz, Eigentumsrechten am Speichermedium etc. hinzukäme und mit diesen in Einklang zu bringen wäre, würde die ohnehin bestehende Komplexität und Rechtsunsicherheit nur deutlich erhöhen, ohne dass ersichtlich wäre, dass ein solches Recht für die Verkehrsfähigkeit von Daten erforderlich oder auch nur in signifikanter Weise dienlich wäre.

Dennoch hält die DEK das Bedürfnis beispielsweise der Industrie oder auch öffentlicher Stellen für berechtigt, vertraglichen Absprachen (beispielsweise zur Einschränkung der Datenweitergabe oder der Zweckbindung von Datenverwendung) eine **begrenzte Drittwirkung** zu verleihen. Nach derzeit geltender Rechtslage ist eine solche Drittwirkung – sofern kein immaterialgüterrechtlicher Schutz eingreift, einschließlich des sog. sui generis-Schutzes von Datenbanken – allenfalls in Extremfällen gegeben. Hier wäre zu erwägen, in Anlehnung an Art. 4 Abs. 4 der Geschäftsgeheimnis-Richtlinie 2016/943 eine Drittwirkung in weiterem Umfang anzuerkennen.[27] Danach gälte der Erwerb, die Nutzung oder die Weitergabe von Daten als rechtswidrig, wenn eine Person zum Zeitpunkt des Erwerbs, der Nutzung oder der Weitergabe wusste oder unter den gegebenen Umständen hätte wissen müssen, dass die Zwischenperson, von der sie die Daten unmittelbar oder mittelbar erhalten hatte, sie rechtswidrig genutzt oder weitergegeben hat. Dieses Konzept würde der Datenwirtschaft helfen und sich bruchlos in das bestehende, im Wesentlichen auf Verträge fokussierte Modell einpassen.

27 Diese Lösung wird etwa von den Vorentwürfen Nr. 2 (Februar 2019) und Nr. 3 (Oktober 2019) der ALI-ELI Principles for a Data Economy (oben Fn. 1) verfolgt.

5.2.5 Datenpartnerschaften

Auch im Bereich des **Kartellrechts** hält die DEK es für sachgerecht, den geltenden Rechtsrahmen behutsam fortzuentwickeln. Die dynamische Entwicklung der Datenwirtschaft stellt das Kartellrecht vor neue Herausforderungen, und das Kartellrecht bringt umgekehrt genauso neue Herausforderungen für digitale Unternehmen mit sich. Die DEK empfiehlt der Bundesregierung, insbesondere die Chancen und Risiken von **Datenpartnerschaften** zu prüfen. Zu erwägen wäre hierbei auch eine Pflicht zur vertraulichen Anzeige von Datenpartnerschaften an die Kartellbehörden, sowie – im Hinblick auf personenbezogene Daten – an die datenschutzrechtlichen Aufsichtsbehörden. Die DEK verweist im Übrigen auf die Vorschläge, die die Kommission Wettbewerbsrecht 4.0 zu diesen Themen unter den Begriffen „Datenaustausch" und „Datenpooling" vorgelegt hat.

5.3 Datenzugang in bestehenden Wertschöpfungssystemen

5.3.1 Problemstellung

In modernen Wertschöpfungssystemen kommt dem Aspekt eines **fairen und effizienten Datenzugangs** erhebliche Bedeutung zu. Die faire und effiziente Regelung des Datenzugangs verschiedener Akteure im Wirtschaftsverkehr wird primär durch das **Vertragsrecht** gewährleistet. In ihm kommt die Autonomie privater Akteure, die sog. Privatautonomie, am deutlichsten zum Tragen. Zugleich besteht eine allgemeine Vermutung, dass durch frei ausgehandelte Vereinbarungen – jenseits von Fällen des Marktversagens – eine effiziente Ressourcenallokation erreicht und damit die allgemeine Wohlfahrt gesteigert wird.

Aufgrund von Machtungleichgewichten und Informationsasymmetrien kann es allerdings auch zu **unfairen und ineffizienten vertraglichen Regelungen** kommen. Dies trifft insbesondere auf Aspekte des Datenzugangs zu, welche typischerweise in der Verhandlungsphase unterschätzt und dementsprechend vergessen oder nicht hinreichend durchdacht werden. Angesichts des dynamischen Charakters datenbezogener Interessenslagen und der dementsprechend dynamischen Bewertung von Datenrechten und Datenpflichten (→ oben 2.1) ist es auch vielfach schwierig für die Parteien, für die gesamte Vertragsdauer vorauszusehen, wie ein faires und effizientes Datenzugangsregime genau zu gestalten ist. Dadurch kommt es in der Praxis nicht selten später zu nicht vorhergesehenen Verschiebungen und Ungleichgewichten, die das ursprünglich vereinbarte Gefüge von Rechten und Pflichten empfindlich stören. Da typischerweise eine der Parteien von solchen Verschiebungen profitiert, kommt es aber vielfach nicht zu Neuverhandlungen und zu einer sachgerechten und effizienten Regelung.

Gerade in komplexen Wertschöpfungssystemen sind die Zugang begehrende Partei und die die Daten faktisch kontrollierende Partei oftmals auch gar **nicht unmittelbar vertraglich miteinander verbunden** (z. B. weil ein weiteres Glied in der Vertriebskette zwischengeschaltet ist), während sie aus Gründen der Fairness und Effizienz durch ein Datenzugangsregime miteinander verbunden sein sollten. Eingriffe in die vertragliche Abschlussfreiheit in Gestalt eines sog. Kontrahierungszwangs folgen derzeit im Verhältnis zwischen zwei Unternehmen (sog. Business-to-Business-Bereich, B2B) fast ausschließlich aus dem Kartellrecht, bei lebenswichtigen Gütern und monopolartigen Stellungen teilweise auch aus allgemeinen Vorschriften, und sind insgesamt auf wenige extreme Situationen beschränkt.

5.3.2 Situation bei Bestehen eines Vertragsverhältnisses

Nach Auffassung der DEK bedarf es zur Gewährleistung fairer und effizienter vertraglicher Regelungen des Datenzugangs zunächst Maßnahmen zur **Bewusstseinsbildung und zur Förderung digitaler Kompetenzen** (→ oben) sowie praktische Unterstützung in Form der Bereitstellung von **Modellverträgen**, die eine gerechte Verteilung des Datenzugangs vorsehen, sowie von **Infrastrukturen und Intermediären**, die eine geteilte Datennutzung ermöglichen, ohne beispielsweise Geschäftsgeheimnisse offenbaren zu müssen (→ oben).

Soweit ein vertragliches Rechtsverhältnis bereits besteht, kann den Prinzipien eines fairen Datenzugangs vor allem im Wege der (gegebenenfalls ergänzenden) **Vertragsauslegung** – etwa durch Annahme entsprechender vertraglicher Nebenpflichten – sowie im Wege der **Kontrolle Allgemeiner Geschäftsbedingungen (AGB)** nach § 307 BGB (sog. Inhaltskontrolle) Rechnung getragen werden. Ein Problem bei der Vornahme der Inhaltskontrolle stellt allerdings das weitgehende Fehlen dispositiver Regelungen dar, die als Maßstab der Inhaltskontrolle dienen könnten. Daher könnten einzelne Tatbestände als ausdrücklich verbotene Vertragsklauseln (sog. **Klauselverbote)** formuliert werden (→ zur entsprechenden Forderung bei Verträgen zwischen Unternehmern und Verbrauchern, sog. B2C-Verträge, siehe schon oben 3.2.3). Daneben kommt bei einer wesentlichen Änderung der Verhältnisse ein Rückgriff auf die Regelungen zur **Störung der Geschäftsgrundlage** nach § 313 BGB in Betracht.

Die DEK bekräftigt in diesem Zusammenhang die von der **Europäischen Kommission** in ihrer Mitteilung vom April 2018 zum „Aufbau eines gemeinsamen europäischen Datenraums“ entwickelten **allgemeinen Grundprinzipien für einen Datenaustausch zwischen Unternehmen (B2B-Bereich)**.[28] Diese Grundprinzipien sehen vor:

a) Transparenz von Zugangsrechten und Zwecken der Datennutzung;

b) Anerkennung von Beiträgen anderer Beteiligter zur Wertschöpfung;

c) Gegenseitige Achtung der Geschäftsinteressen aller Beteiligten;

d) Gewährleistung eines unverfälschten Wettbewerbs; und

e) Minimierung der Datenabhängigkeit von einem Anbieter (Daten-Lock-in).

28 Europäische Kommission: Aufbau eines gemeinsamen europäischen Datenraums, COM(2018) 232 final, 25.4.2018, S. 12 (abrufbar unter: *https://ec.europa.eu/transparency/regdoc/rep/1/2018/DE/COM-2018-232-F1-DE-MAIN-PART-1.PDF*).

Darüber hinaus kommen – insbesondere für potenziell gemischte Datenbestände mit personenbezogenen Daten – eine Ergänzung um das informationelle Selbstbestimmungsrecht der Betroffenen und das Nichtschadensprinzip in Betracht.

5.3.3 Situation bei Fehlen eines Vertragsverhältnisses

Soweit Teilnehmer eines Wertschöpfungssystems trotz aller unterstützenden Maßnahmen nicht unmittelbar vertraglich miteinander verbunden sind, greift mangels eines Vertrages weder die Vertragsauslegung noch die Inhaltskontrolle von AGB noch kann auf die Grundsätze von der Störung der Geschäftsgrundlage rekurriert werden. Nach Auffassung der DEK begründet allerdings bereits der bloße Umstand, dass eine Zugang begehrende Partei zur Generierung von Daten beigetragen hat – und zwar umso mehr, wenn dies innerhalb eines prinzipiell durch Verträge geprägten Wertschöpfungssystems erfolgt – eine rechtliche Sonderbeziehung zu der die Daten faktisch kontrollierenden Partei (→ oben). Aus dieser rechtlichen Sonderbeziehung können gewisse Schutz- und Treuepflichten einschließlich der **Pflicht zur Aufnahme von Vertragsverhandlungen über ein faires und effizientes Datenzugangsregime** erwachsen. Dies sollte von der Rechtsordnung künftig explizit anerkannt werden.

Die DEK empfiehlt daher eine **Ergänzung von § 311 BGB** um einen weiteren Absatz, welcher diese Sonderbeziehung bei Beteiligten eines Wertschöpfungssystems (z. B. als Zulieferbetrieb, Hersteller, Händler oder Endnutzer) zum Ausdruck bringt und entsprechende Pflichten nach sich zieht. Die Bedeutung von Daten für den allgemeinen Rechts- und Wirtschaftsverkehr rechtfertigt es, dies nicht länger unter die Generalklausel der „ähnlichen geschäftlichen Kontakte“ zu fassen, sondern mit einem eigenen Absatz im Gesetz zu bedenken. Dieser würde weder eine eigene Rechtsgrundlage für die Verarbeitung personenbezogener Daten darstellen noch könnte er datenschutzrechtliche Positionen beschränken.

Darüber hinausgehend könnte erwogen werden, ein an den oben (→ unter 2) genannten Prinzipien orientiertes **dispositives Datenschuldrecht** zur Lückenfüllung und als Maßstab für die Inhaltskontrolle von AGB zu schaffen.[29] Ein solches Datenschuldrecht könnte Bedingungen definieren, unter denen insbesondere Ansprüche auf Zugang zu Daten und/oder Ansprüche auf Unterlassung eines Datenzugangs oder einer Datennutzung und/oder Ansprüche auf Korrektur von Daten bestehen. Allerdings hat die DEK auch Bedenken, dass durch speziell normierte (wenngleich dispositive) Ansprüche zusätzliche Streitigkeiten provoziert werden könnten.

5.3.4 Sektorspezifische Datenzugangsrechte

Was darüber hinausgehende Datenzugangsrechte in bestehenden Wertschöpfungssystemen anbelangt, wird zunächst an sektorspezifische Lösungen zu denken sein. Die DEK empfiehlt der Bundesregierung insofern, bei Erlass und/oder Überarbeitung sektorspezifischer Regelungen Fragen des Datenzugangs verstärkte Aufmerksamkeit zu widmen.

29 Für personenbezogene Daten siehe Louisa Specht: Datenrechte – Eine Rechts- und Sozialwissenschaftliche Analyse im Vergleich Deutschland – USA, Teil 1: Rechtevergleichende Analyse des zivilrechtlichen Umgangs mit Daten in den Rechtsordnungen Deutschlands und der USA, ABIDA-Gutachten, 2017, S. 89 ff. (abrufbar unter: *http://www.abida.de/sites/default/files/ABIDA_Gutachten_Datenrechte.pdf*); für nicht-personenbezogene Daten siehe ALI-ELI Principles for a Data Economy (oben Fn. 1).

5.4 Offene Daten des öffentlichen Sektors

5.4.1 Vorüberlegungen

Die Öffnung von Daten des öffentlichen Sektors durch sog. Open-Government-Data-Konzepte hat mit der jüngst überarbeiteten Richtlinie (EU) 2019/1024 über Offene Daten und Informationen des öffentlichen Sektors (PSI-Richtlinie) sowie auf nationaler Ebene mit dem Informationsweiterverwendungsgesetz (IWG), dem E-Government-Gesetz (EGovG) und weiteren Spezialgesetzen eine feste gesetzliche Grundlage. Open Government Data beruht auf der Überlegung, dass Bürger und Unternehmen für die Generierung dieser Daten bereits **mit Steuergeldern bezahlt** haben und daher an den Daten partizipieren und nicht etwa doppelt finanziell belastet werden sollten. Die Öffnung von Daten des öffentlichen Sektors zur Weiterverwendung durch die Privatwirtschaft kommt zudem der europäischen Datenwirtschaft zugute. Da den Daten des öffentlichen Sektors vielfach ein **großes Wertschöpfungspotenzial für privatwirtschaftliche Unternehmen** zukommt, können diese Unternehmen damit neue innovative Produkte und Dienstleistungen entwickeln und so auch zur allgemeinen Wohlfahrtssteigerung beitragen.

Über die Wirtschaft hinaus ist der Zugang zu staatlichen Daten auch wichtig für die **Demokratie und einen offenen Diskurs** in der Gesellschaft, denn er erhöht die Verwaltungstransparenz, erleichtert Partizipation und fördert eine auf Fakten gestützte öffentliche Diskussion und Kontrolle. Darüber hinaus können Daten des öffentlichen Sektors in vielfältiger Form für gesellschaftliche Initiativen und Innovationen genutzt werden, etwa zu sozialen oder ökologischen Zwecken.

Die DEK unterstützt daher grundsätzlich die auf dem G8-Gipfel 2013 beschlossene **Open-Data-Charta**. Diese definiert zentrale Prinzipien für den Umgang mit Verwaltungsdaten:

a) Standardmäßig offene Daten (Förderung der Erwartung, dass Verwaltungsdaten bei Beibehaltung des Schutzes der Privatsphäre öffentlich gemacht werden);

b) Qualität und Quantität (Freigabe qualitativ hochwertiger, aktueller und gut beschriebener offener Daten);

c) Von allen verwendbar (Freigabe so vieler Daten wie möglich in so vielen offenen Formaten wie möglich);

d) Freigabe von Daten für verbessertes verantwortungsbewusstes staatliches Handeln (Weitergabe von Expertise und Herstellung von Transparenz betreffend Datensammlung, Standards und Veröffentlichungsverfahren);

e) Freigabe von Daten für Innovation (Nutzer-Konsultationen und Unterstützung künftiger Generationen von Ideengebern).

Geben öffentliche Stellen Daten unentgeltlich an kommerzielle Akteure weiter statt sie gewinnbringend zu veräußern oder sonst wirtschaftlich zu verwerten, sollte dies aus ethischer Sicht allerdings bei pauschalierender Betrachtung durch entsprechende gesamtgesellschaftliche **Wohlfahrtsgewinne** gerechtfertigt sein.

Ferner weist die DEK auf ein mögliches **Spannungsverhältnis** zwischen Forderungen nach Privacy-by-Default einerseits und Open-by-Default andererseits sowie ganz generell zwischen dem **Diskurs um Datenschutz und dem Diskurs um Open Government Data** hin. Soweit im Rahmen von Open-Data-Konzepten personenbezogene Daten in rechtlich zulässiger Weise öffentlich gemacht werden, ist nicht gesichert, dass die zur Wahrung des Schutzes der informationellen Selbststimmung getroffenen Sicherungsmechanismen in Form ausdrücklicher oder impliziter Weiterverwendungsbeschränkungen sowie in Form technischer und organisatorischer Schutzmaßnahmen gewahrt bleiben. Gleiches gilt für die allgemeinen datenschutzrechtlichen Bestimmungen für die Weiterverwendung. Da Art. 30 DSGVO zudem nur die Dokumentierung der „Kategorien von Empfängern“ verlangt und staatliche Stellen die Einhaltung „geeigneter Garantien“ im Sinne des Art. 89 DSGVO so gut wie nicht überwachen können, geht von der Offenlegung von Daten, welche personenbezogen sind oder werden können, für die Betroffenen ein besonderes Gefährdungspotenzial aus.

Vor diesem Hintergrund ist im Zusammenhang mit Open-Government-Data-Konzepten stets eine besonders sorgfältige Abwägung des grundrechtlich verankerten Rechts auf informationelle Selbstbestimmung mit den durch Open Government Data verfolgten Gemeinwohlbelangen und dem – ebenfalls grundrechtlich verankerten – Recht auf Informationsfreiheit und mit der Berufsfreiheit der durch Open Government Data Begünstigten vorzunehmen. Nach Auffassung der DEK muss diese Abwägung **in Zweifelsfällen** für den staatlichen Schutzauftrag ausfallen. Dies gilt umso mehr, als der Einzelne teilweise nicht frei bestimmen kann, welche Daten er staatlichen Akteuren anvertraut bzw. er **in besonderem Maße darauf vertraut**, dass staatliche Akteure personenbezogene Daten nicht an Dritte weiterleiten.

5.4.2 Rechtsrahmen und Infrastrukturen

Die DEK begrüßt den Nationalen Aktionsplan der Bundesregierung zur Umsetzung der G8 Open-Data-Charta und die Bemühungen der Bundes- und Landesregierungen um die Digitalisierung der Verwaltung unter Einschluss von Open-Government-Data-Konzepten. Sie empfiehlt der Bundesregierung, darauf hinzuwirken, dass die in § 12a Abs. 1 S. 1 EGovG bereits für die Behörden der unmittelbaren Bundesverwaltung normierte **Pflicht zur Veröffentlichung strukturierter, unbearbeiteter Daten (Open-by-Default)** und zur grundsätzlich unentgeltlichen Bereitstellung dieser Daten zur uneingeschränkten Nutzung umfassend implementiert wird. Im Lichte des oben beschriebenen möglichen Spannungsverhältnisses von Open Government Data und Datenschutz wird die von § 12a EGovG vorgesehene Bereitstellung zur entgeltfreien und uneingeschränkten Weiterverwendung der Daten durch jedermann allerdings nur für bestimmte Datenarten (insbesondere effektiv anonymisierte Daten) in Frage kommen.

Der von der DEK begrüßte Versuch des Gesetzgebers, einen Kulturwandel der Verwaltung im Umgang mit Daten zu initiieren, wird allerdings dadurch erschwert, dass die gegenwärtige **Rechtslage sehr zersplittert** ist. Sowohl für die Behörden wie auch für potenzielle Nutzer der Daten öffentlicher Stellen ist das Zusammenspiel der unterschiedlichen Rechtsregime aus allgemeinen und speziellen Informationszugangs-, Informationsweiterverwendungs- und E-Government-Regelungen je auf Bundes- und auf Landesebene nur schwer durchschaubar. Hinzu kommt das in der Praxis vielfach schwierige Zusammenspiel dieser Regelungen mit dem Datenschutzrecht und dem Schutz des geistigen Eigentums, insbesondere dem Urheberrecht. Die DEK empfiehlt diesbezüglich eine **Zusammenführung** und **Synchronisierung** der verschiedenen Rechtsgrundlagen in Deutschland sowie sachgerechte **Klarstellungen** zur Abgrenzung der Rechtsmaterien.

Der erforderliche Kulturwandel wird auch dadurch erschwert, dass sich derzeit kaum verbindlich überprüfen lässt, ob die Behörden ihren schon bestehenden Verpflichtungen zur Datenbereitstellung tatsächlich nachkommen. So sieht etwa § 12a Abs. 1 EGovG zwar eine Pflicht der Behörden der unmittelbaren Bundesverwaltung vor, Daten zum öffentlichen Abruf bereitzustellen, gewährt der Zugang suchenden Person aber ausdrücklich **keinen einklagbaren Anspruch auf Bereitstellung**. Damit fehlt es zugangssuchenden Unternehmen an wirksamen Mechanismen, um eine Durchsetzung der gesetzlichen Bereitstellungspflicht (Open-by-Default) zu erzwingen. Aus Sicht der DEK kann die Schaffung eines **subjektiven Rechts auf Bereitstellung** dazu beitragen, die Bereitschaft der Verwaltung zur proaktiven Bereitstellung offener Daten – im Rahmen der vom EGovG bzw. IWG für die Bereitstellungspflicht statuierten Grenzen – zu fördern.

Zudem stellt die geltende Rechtslage nicht hinreichend sicher, dass die von der öffentlichen Hand zur Verfügung gestellten Daten eine **ausreichende Datenqualität aufweisen**. Insbesondere beschränkt sich die Bereitstellungspflicht nach dem EGovG auf unbearbeitete Daten. Dabei ist eine problemlose Weiterverwendung von Daten, wie sie den Zielen von Open Government Data entspricht, nur möglich, wenn eine hohe Datenqualität gewährleistet ist.

Neben den rechtlichen sind zudem die **infrastrukturellen Grundlagen** (z. B. Open-Government-Data-Portale wie GovData) zu schaffen bzw. auszubauen, auch und gerade, was beispielsweise kommunale Plattformen betrifft. Dies gilt auch für die Investition in hinreichende Qualitätssicherungsmaßnahmen.

5.4.3 Schutzauftrag des Staates

Im Hinblick auf den Schutzauftrag des Staates bezüglich aller ihm anvertrauten Daten muss durch **entsprechende Vorkehrungen** gewährleistet sein, dass der Schutz wichtiger Individualinteressen (z. B. bei personenbezogenen Daten, Betriebs- und Geschäftsgeheimnissen oder sonstigen schutzbedürftigen Daten, wie etwa vertraulichen Informationen im Rahmen von Vergabeverfahren der öffentlichen Hand) ebenso vollumfänglich garantiert ist wie der Schutz wichtiger Allgemeininteressen (wie etwa Sicherheitsinteressen oder Interessen der nationalen Souveränität). Die dem Open-Government-Data-Konzept zugrundeliegende ethische Überlegung, dass die Bürger und Unternehmen für die Daten bereits mit ihren Steuergeldern bezahlt haben, bedeutet auch gewisse **Einschränkungen der Weiterverwendung**. Insbesondere ist Sorge zu tragen, dass die Daten nicht zur privatwirtschaftlichen Entwicklung von Diensten und Produkten verwendet werden, welche die Freiheit der Bürger und Unternehmen letztlich einschränken und/oder ihnen schließlich zu unfairen Konditionen angeboten werden.

Der Bundesregierung ist daher zu empfehlen, von der in Art. 8 der neugefassten PSI-Richtlinie eröffneten Möglichkeit Gebrauch zu machen, in Standardlizenzen **Modellkonditionen** einschließlich Zweckbindungsvereinbarungen und Bedingungen für die Weitergabe an Dritte zu entwickeln bzw. auf deren Entwicklung auf europäischer Ebene hinzuwirken. Die DEK empfiehlt, die Verwendung solcher Modellkonditionen – mindestens sektorspezifisch – sogar **bindend vorzuschreiben**. Dabei sollten sie sich u.a. an folgenden Eckpunkten orientieren:

a) Gemäß Art. 8 Abs. 1 PSI-Richtlinie müssen die Bedingungen objektiv, verhältnismäßig, nichtdiskriminierend und durch ein im Allgemeininteresse liegendes Ziel gerechtfertigt sein; sie dürfen die Möglichkeiten der Weiterverwendung nicht unnötig einschränken und nicht der Behinderung des Wettbewerbs dienen;

b) Unternehmen sollten sich Richtlinien unterwerfen, die klar definierte Garantien für die Rechte betroffener Dritter enthalten sowie Mechanismen für deren Überprüfbarkeit vorsehen;

c) Mithilfe der Daten entwickeltes geistiges Eigentum darf nicht dazu genutzt werden, Aktivitäten, die öffentliche Stellen im Rahmen der Erfüllung ihrer öffentlichen Aufgaben verfolgen, zu untersagen bzw. nur noch gegen Zahlung einer Lizenzgebühr zuzulassen;

d) Wird mithilfe der Daten ein Produkt oder eine Dienstleistung entwickelt, sollte dieses Produkt bzw. diese Dienstleistung öffentlichen Stellen unter Vorzugsbedingungen anzubieten sein;

e) Marktstarke Unternehmen sollten eine Reziprozitätsverpflichtung in dem Sinne eingehen, dass sie ihrerseits unter gleichen Bedingungen Betriebsdaten zur Verfügung stellen; und

f) Daten sollten nur für unternehmerische Aktivitäten in der EU verwendet werden, oder bei denen zumindest die Entwicklung des Produkts oder der Dienstleistung in der EU erfolgt.

Bei jedem Transfer von Daten, bei dem der Empfänger eine Kopie der Daten auf einer von ihm kontrollierten Infrastruktur erhält, lässt sich die Einhaltung vereinbarter Garantien und Zweckbeschränkungen im Prinzip nicht mehr zuverlässig kontrollieren. Im Bestreben, seinem Schutzauftrag bei Daten, die – gegebenenfalls auch nur im Fall einer De-Anonymisierung oder Verknüpfung mit anderen Datensätzen – zum Schaden Dritter oder der Allgemeinheit verwendet werden könnten, gerecht zu werden, werden staatliche Stellen insbesondere erwägen müssen, ausschließlich den überwachten Zugang und die **überwachte Verarbeitung** auf einer von der staatlichen Stelle kontrollierten Infrastruktur zuzulassen. Die dafür anfallenden Kosten wären auf die Zugang suchenden Unternehmen umzulegen.

5.5 Offene Daten des privaten Sektors

5.5.1 Plattformen und Datennutzung

In der deutschen Wirtschaft fallen im Geschäftsbetrieb des Unternehmens sog. Betriebsdaten an. Diese Daten haben einen großen Wert für Innovationen, insbesondere, wenn sie mit den Daten anderer Teilnehmer der Wertschöpfungskette verknüpft werden. Zum Zwecke einer derartigen Verknüpfung hat die deutsche Wirtschaft sektorspezifische Plattformen geschaffen.

Beispiele für die verschiedenen Plattformtypen sind: (1) Zusammenschluss verschiedener Unternehmen in einer GmbH; (2) Eigenbetrieb eines Unternehmens mit Anbindung von Partnern; (3) Begründung einer unternehmenseigenen Plattform als Serviceplattform für Dritte.

Neben den Plattformen verständigen sich die verschiedenen Branchen zunehmend auf gemeinsame Regelungskonzepte zur Nutzung der Daten.

Die DEK geht davon aus, dass die Wirtschaft weiterhin branchenspezifisch ihre Datennutzung innerhalb der Wertschöpfungssysteme selbst organisiert und dabei auch die für Innovation notwendige Offenheit für neue Marktteilnehmer und Start-ups zeigt. Es liegt nämlich im Interesse der Marktteilnehmer, in Zusammenarbeit mit innovativen Start-ups, digitale Sprunginnovationen zu entwickeln und in diesem Zusammenhang ihre Daten zu teilen. Die Selbstorganisation in unterschiedlichen Typen von Plattformen stärkt das in Europa bestehende industrielle Knowhow und gewährleistet eine höhere Qualität der Datennutzung (einschließlich Datenschutz und Informationssicherheit). Die DEK regt an, die positive **Entwicklung der privatwirtschaftlich organisierten Plattformen zu fördern**, um die erforderliche Marktgröße und Skaleneffekte zu erreichen und damit gemeinsam international wettbewerbsfähig zu sein.

5.5.2 Anreize zum weitergehenden freiwilligen Teilen

Bereits gegenwärtig existieren viele Geschäftsmodelle, die auf einer freiwilligen Gewährung eines Datenzugangs für die Allgemeinheit seitens privater Anbieter beruhen.

Beispiel 14
Dies ist etwa beim sog. Geobusiness der Fall, bei dem (z.T. aus behördlichen Quellen stammende) Geobasisdaten mit weiteren Informationen angereichert werden, sodass für die verschiedensten Zwecke Geofachdaten bereitgestellt werden. Zu denken ist hier nicht nur an Kartendienste, wie Open Street Map oder Google Maps, die über die reinen topographischen und administrativen Informationen mit einer Vielzahl zusätzlicher Informationen versehen werden, sondern auch spezifische Angebote wie Vorhersagen hinsichtlich Wetter oder Verkehrsbedingungen.

Die DEK empfiehlt eine Förderung solcher Offenlegung auf freiwilliger Basis. Hierfür sind neben den empfohlenen **Maßnahmen praktischer Unterstützung** (→ oben 5.2) auch **weitere Anreize** zum freiwilligen Teilen von Daten in Erwägung zu ziehen, etwa eine positive Berücksichtigung von Daten(frei)gaben und Open-Access-Strategien

- im Steuerrecht;
- im Rahmen des Vergaberechts;
- bei der Vergabe von Fördermitteln (auch außerhalb des Forschungsbereichs) oder
- bei der Durchführung von Genehmigungsverfahren.

Freiwilliges Teilen, Daten(frei)gaben und Open-Access-Strategien kommen in den vorgenannten Bereichen allerdings nur in Betracht, soweit damit keine Geheimhaltungserfordernisse aufgrund des Vergaberechts oder aufgrund von Betriebs- und Geschäftsgeheimnissen verletzt und keine Regelungen des Datenschutzrechts missachtet werden.

5.5.3 Gesetzliche Datenzugangsrechte

Im Gegensatz zum freiwilligen Teilen von Daten steht bei gesetzlichen Datenzugangsrechten der Gedanke im Mittelpunkt, dass bei großen Datenbeständen, soweit sie durch das Zusammenwirken vieler Mitglieder der Gesellschaft akkumuliert wurden – etwa durch soziale Netzwerke – der Gesellschaft auch etwas zurückzugeben ist. Dieser Gedanke könnte – in Verbindung mit dem grundlegenden Wert gesellschaftlicher Solidarität sowie mit im konkreten Fall einschlägigen Gemeinwohlinteressen – **weitergehende Zugangsgewährungs- und Offenlegungspflichten** Privater begründen.[30]

Zur Verbesserung des allgemeinen Zugangs zu privat gehaltenen Daten wird zunächst diskutiert, ein an Art. 20 DSGVO angelehntes, **allgemeines Portabilitätsrecht** für nicht-personenbezogene Daten zu schaffen. Das würde beispielsweise bedeuten, dass auch eine juristische Person, auf die sich bestimmte Daten beziehen oder auf die Daten bezogen werden können, gegenüber jedem Akteur, der solche Daten in seinem Besitz hat, verlangen kann, dass ihr diese Daten in einem gängigen und maschinenlesbaren Format übermittelt werden oder dass sie direkt auf einen dritten Akteur übertragen werden. Aus im Wesentlichen ähnlichen Gründen, wie sie bereits gegen eine Erweiterung von Art. 20 DSGVO angeführt wurden (→ oben), empfiehlt die DEK der Bundesregierung, die Entwicklungen hinsichtlich **Nutzung und Auslegung des Art. 20 DSGVO zunächst abzuwarten**. Hinzu kommt als besondere Herausforderung, dass sich bei nicht-personenbezogenen Daten die Frage der Zuordnung, d. h. die Frage nach dem Inhaber des Portabilitätsrechts, ganz neu stellen würde.

Zur Verbesserung des allgemeinen Zugangs zu privat gehaltenen Daten werden auch eine Reihe weiterer Maßnahmen diskutiert, die im Ergebnis auf gesetzliche Datenzugangsrechte hinauslaufen. Dazu gehören als **denkbare Modi der Ausgestaltung** eine gesetzliche Pflicht zur Bereitstellung bestimmter intern erstellter Datenanalysen für die Öffentlichkeit, die Einräumung individueller Zugangsrechte (z. B. Pflicht zur Lizensierung unter FRAND-Bedingungen[31] und/oder nach Anwendung des urheberrechtlichen Drei- bzw. Vier-Stufen-Tests[32]) oder auch die Offenlegung von Daten gegenüber der Allgemeinheit (Open Access), welche sowohl marktanteilsbezogen als auch allgemein ausgestaltet sein kann.

Bei all diesen Maßnahmen sind nach Auffassung der DEK zunächst mindestens die folgenden Faktoren zu berücksichtigen:

30 Dazu u.a. Viktor Mayer-Schönberger / Thomas Ramge: Das Digital, 2017, S. 195 ff.
31 FRAND = Fair, Reasonable and Non-Discriminatory.
32 Der „Drei-Stufen-Test“ bezeichnet einen in mehreren internationalen Verträgen vorgesehenen dreistufigen Test, mit dem geprüft wird, ob eine Ausnahmeregelung (sog. Schrankenbestimmung) einen akzeptablen Eingriff in die Rechte des Urhebers darstellt. Solche Ausnahmen dürfen laut Test nur (i) in Sonderfällen zur Anwendung kommen, welche (ii) die normale, kommerzielle Verwertung nicht beeinträchtigen und auch (iii) die berechtigten Interessen des Rechteinhabers nicht ungebührlich verletzen. Es wird verstärkt gefordert, auch (iv) Drittinteressen sowie Allgemeininteressen zwingend in den Test einzubeziehen.

a) Der Schutz der von der Zugangsgewährung oder Offenlegung betroffenen personenbezogenen Daten sowie von Betriebs- und Geschäftsgeheimnissen muss gewährleistet sein;

b) Die Anforderungen an die Verhältnismäßigkeit des Eingriffs in die Grundrechte der von der Zugangsgewährungs- oder Offenlegungspflicht betroffenen Privaten müssen gewahrt sein; dies betrifft insbesondere die Berufsfreiheit;

c) Negative Folgen für den Wettbewerb durch den Zugang oder die Offenlegung, etwa aufgrund strategischer Nutzung durch – gegebenenfalls selbst nicht offenlegungspflichtige – Mitbewerber sind zu vermeiden;

d) Anreize, in Geschäftsmodelle der Datenwirtschaft zu investieren, dürfen nicht genommen werden; und

e) Der Schutz strategischer Interessen deutscher bzw. europäischer Unternehmen gegenüber globalen Wettbewerbern ist zu berücksichtigen. Dies betrifft insbesondere Konsequenzen für die Stellung der deutschen bzw. europäischen Wirtschaft im globalen Wettbewerb, wenn gerade deutsche bzw. europäische Unternehmen zur Offenlegung ihrer Datenbestände gezwungen wären und diese Datenbestände in die Hände derjenigen Akteure gelangen würden, bei denen bereits jetzt die größte Datenkompetenz, die besten Dateninfrastrukturen und vor allem die größten Datenbestände liegen.

Vor diesem Hintergrund empfiehlt die DEK primär ein **sektorspezifisches Vorgehen**. Im Kontext raumbezogener Informationen stehen mit der INSPIRE-Richtlinie und deren Umsetzung in nationales Recht bereits sektorspezifische Zugangsregelungen zur Verfügung, die allerdings nur öffentliche Stellen zur Bereitstellung verpflichten. Einen ersten Anwendungsfall für ein sektorspezifisches Datenzugangsrecht zu Daten im privaten Sektor gibt es im Bereich der Zahlungsdienstleistungen. Die DEK regt an, Bedarf und Implementierungsoptionen in einer Reihe weiterer ausgewählter Sektoren zu prüfen, wobei beispielsweise der **Nachrichten-, Mobilitäts- oder Energiesektor** infrage käme.

5.5.4 Rolle des Wettbewerbsrechts

Wenngleich das geltende Wettbewerbsrecht kaum datenspezifische Regelungen enthält, sind doch die allgemeinen Regelungen auch auf die Datenwirtschaft anwendbar. Die **Essential Facility Doctrine** (EFD) kann – gegebenenfalls in leicht modifizierter Form – etwa eingreifen, wenn ein marktbeherrschendes Unternehmen exklusiv eine Ressource (z. B. Netz/Infrastruktur) kontrolliert, die für den Wettbewerb auf einem angrenzenden Markt unerlässlich ist. Die **Aftermarket-Doktrin** betrifft den Fall, dass der Nachfrager eines Primärprodukts infolge von Lock-in in der Ausübung seiner Wahlfreiheit auf einem Sekundärmarkt (z. B. Markt für Reparaturen/Ersatzteile) beschränkt wird bzw. ein Drittanbieter auf einem solchen Sekundärmarkt in wettbewerbswidriger Weise behindert wird.[33] Allerdings stellt die Missbrauchsaufsicht derzeit infolge der Unklarheit der Rechtslage, hoher Anforderungen sowie der Dauer und Kosten von Verfahren keine allgemeine Lösung etwaiger Probleme des Datenzugangs dar. Das geltende Wettbewerbsrecht oder einzelne seiner Elemente könnten jedoch zu einem zentralen Baustein eines neuen, **digitalen Wirtschaftsrechts** werden, das wesentlich auch Probleme des Datenzugangs adressiert. Diesbezüglich sind die Ergebnisse der Kommission Wettbewerbsrecht 4.0 zu berücksichtigen.[34]

33 Jacques Crémer / Yves-Alexandre de Montjoye / Heike Schweitzer: Competition policy for the digital era, Special Advisers' Report for the European Commission, S. 87 ff (abrufbar unter: *https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf*).

34 Ein neuer Wettbewerbsrahmen für die Digitalwirtschaft, Bericht der Kommission Wettbewerbsrecht 4.0, 2019 (abrufbar unter: *https://www.bmwi.de/Redaktion/DE/Publikationen/Wirtschaft/bericht-der-kommission-wettbewerbsrecht-4-0.pdf?__blob=publicationFile&v=4*).

5.6 Datenzugang zugunsten von öffentlichen Stellen (B2G) und gemeinwohlorientierten Zwecken

Zu erwägen ist, inwieweit eine Pflicht zur Zugangsgewährung definierter Teilmengen von Daten zugunsten bestimmter **öffentlicher Stellen** oder bestimmter **gemeinwohlorientierter Zwecke** in Betracht kommt. Eine besondere Bedeutung können Zugangsrechte zu den Daten Privater bzw. Offenlegungspflichten im Bereich der **Forschung** haben. Hier könnte ein erleichterter Zugang – im Falle einer angemessenen Ausgestaltung, die den Rechten der Betroffenen vollumfassend Rechnung trägt – zum allgemeinen Erkenntnisfortschritt beitragen. Entsprechende Zugangsrechte zu Daten des Privatsektors können zudem Nicht-Regierungs-Organisationen, Medien und ähnlichen Stellen die Erfüllung ihrer gesellschaftlichen Funktionen erleichtern und so zur Sicherung des **demokratischen Gemeinwesens** beitragen. Eine besonders herausgehobene Stellung muss auch stets dem Zweck der **Gefahrenabwehr** (z. B. Unwetterwarnung) zukommen.

Dabei empfiehlt sich aus Sicht der DEK erneut primär ein **sektorspezifisches Vorgehen**, das die Ausgestaltung von Datenzugangs- und Offenlegungspflichten an die konkret betroffenen verfassungsrechtlichen Vorgaben einerseits und die praktischen Gegebenheiten des Sachbereichs andererseits anpasst. Hohe Priorität besteht insbesondere im **Gesundheitssektor**, im **Mobilitätssektor** und im **Energiesektor**. Eine allgemeinere Pflicht zur Datenbereitstellung – etwa generell zu gemeinwohlorientierten Forschungszwecken – bedürfte dagegen nach Auffassung der DEK erst einer breiten gesellschaftlichen Debatte, zu welcher die DEK an dieser Stelle einladen möchte.

Die DEK bekräftigt die von der Europäischen Kommission in ihrer Mitteilung vom 25. April 2018 zum „Aufbau eines gemeinsamen europäischen Datenraums“ formulierten **Grundprinzipien für einen Datenaustausch zwischen privaten Unternehmen und dem öffentlichen Sektor (sog. Business-to-Government-Konstellation, B2G)**:[35]

a) Verhältnismäßigkeit (d. h. Zweckdienlichkeit für ein klares und nachweisbares öffentliches Interesse und Angemessenheit im Hinblick auf Detailliertheit, Relevanz und Datenschutz);

b) Zweckbindung (d. h. eindeutige Beschränkung auf einen oder mehrere Zwecke und Zusicherung der Nichtverwendung in Verwaltungs- oder Gerichtsverfahren);

c) Schadensvermeidung (d. h. Schutz berechtigter Interessen wie dem informationellen Selbstbestimmungsrecht betroffener Personen, Geschäftsgeheimnissen, vertraulichen Geschäftsinformationen und Verwertungsinteressen);

d) Berücksichtigung des öffentlichen Interesses bei den Vertragsbedingungen (Vorzugsbedingungen für öffentliche Stellen, Gleichbehandlung öffentlicher Stellen, Verringerung der Gesamtbelastung für Bürger und Unternehmen);

e) Datenqualitätsmanagement (zumutbare Unterstützung bei der Qualitätsbewertung, aber normalerweise keine Pflicht zur Qualitätsverbesserung);

f) Transparenz und Einbeziehung der Öffentlichkeit bezüglich Vertragsparteien, Zielen, erlangten Erkenntnissen und bewährten Verfahren.

Diese Grundprinzipien könnten einen **guten Ausgangspunkt** darstellen, und zwar nicht nur für die Bedingungen frei vereinbarter Verträge zum Datenaustausch, sondern auch als mögliche Bedingungen etwaiger weitergehender, sektorspezifischer gesetzlicher Maßnahmen zur Verbesserung eines Datenzugangs.

35 Europäische Kommission, Aufbau eines gemeinsamen europäischen Datenraums, COM(2018) 232 final, 25.4.2018, S. 15 f (abrufbar unter: *https://ec.europa.eu/transparency/regdoc/rep/1/2018/DE/COM-2018-232-F1-DE-MAIN-PART-1.PDF*).

Zusammenfassung der wichtigsten Handlungsempfehlungen

Datenzugangsdebatten jenseits des Personenbezugs

24

Für die Entwicklung der europäischen Datenwirtschaft sieht die DEK einen zentralen Faktor im Zugang europäischer Unternehmen zu geeigneten nicht-personenbezogenen Daten in geeigneter Qualität. **Datenzugang** nutzt allerdings nur Akteuren, die ein entsprechendes Bewusstsein für die Bedeutung von Daten haben und über entsprechende Datenkompetenz verfügen, und in ganz überproportionalem Ausmaß denjenigen, bei denen bereits der größte Ausgangsbestand an Daten und die besten Dateninfrastrukturen vorhanden sind. Die DEK empfiehlt daher, bei der Diskussion um eine Verbesserung des Datenzugangs stets die genannten Faktoren gemäß dem **ASISA-Prinzip** *(Awareness – Skills – Infrastructures – Stocks – Access)* mit zu berücksichtigen.

25

Daher unterstützt die DEK die bereits auf europäischer Ebene begonnenen Maßnahmen zur Förderung von **Dateninfrastrukturen** im weitesten Sinne (z. B. Plattformen, Standards für Programmierschnittstellen und weitere Elemente, Modellverträge, EU-Unterstützungszentrum) und empfiehlt der Bundesregierung, diese weiterhin durch entsprechende Bemühungen auf nationaler Ebene zu flankieren. In diesem Zusammenhang bietet sich die Einrichtung einer Ombudsstelle auf Bundesebene an, welche bei Aushandlung von Datenzugangsvereinbarungen und bei Streitigkeiten hilft und vermittelt.

26

Die DEK sieht einen Schlüsselfaktor in einer holistisch gedachten, nachhaltigen und strategischen **Wirtschaftspolitik**, welche der Abwanderung innovativer europäischer Unternehmen bzw. deren Aufkauf durch Akteure aus Drittstaaten ebenso effektiv entgegenwirkt wie der übermäßigen Abhängigkeit von Infrastrukturen (z. B. Serverkapazitäten) in Drittstaaten. Dabei ist die richtige Balance zu finden zwischen gewollter internationaler Kooperation und Vernetzung einerseits und andererseits der entschlossenen Übernahme von Verantwortung für nachhaltige Sicherheit und Wohlfahrt in Europa vor dem Hintergrund sich wandelnder globaler Machtverhältnisse.

27

Die DEK sieht auch unter dem Blickwinkel einer Förderung der Datenwirtschaft keinen Bedarf nach der Einführung neuer Ausschließlichkeitsrechte („Dateneigentum“, „Datenerzeugerrecht“), sondern empfiehlt stattdessen eine **beschränkte Drittwirkung vertraglicher Vereinbarungen** (z. B. betreffend Beschränkungen der Nutzung und Weitergabe von Daten) nach dem Vorbild des neuen europäischen Regimes zum Schutz von Geschäftsgeheimnissen. Ferner wäre es wünschenswert, wenn gesetzlich Wege aufgezeigt würden, wie europäische Unternehmen – etwa unter Einschaltung von Treuhändern – unter voller Wahrung kartellrechtlicher Belange bei der Datennutzung kooperieren können **(„Datenpartnerschaften“)**.

28

In bestehenden Wertschöpfungssystemen (z. B. Produktions- und Vertriebsketten) fallen vielfach Daten an, die innerhalb wie außerhalb des Wertschöpfungssystems von enormer wirtschaftlicher Bedeutung sind. Die zwischen den einzelnen Teilnehmern eines Wertschöpfungssystems bestehenden Verträge enthalten aber häufig entweder keine bzw. eine unfaire und/oder ineffiziente Regelung des Datenzugangs, oder es fehlt ganz an einer vertraglichen Vereinbarung. Weit über die klassische „Datenwirtschaft" hinaus ist daher **Bewusstseinsbildung bei Wirtschaftstreibenden** erforderlich, die durch praktische Hilfestellungen (z. B. Modellverträge) ergänzt werden sollte.

29

Darüber hinaus regt die DEK eine **behutsame Ergänzung des geltenden Rechtsrahmens** an. Dabei sollte ein erster Schritt darin liegen, die Sonderbeziehung zwischen einer Partei, welche zur Generierung von Daten in einem Wertschöpfungssystem beigetragen hat, und der Partei, welche die Daten faktisch kontrolliert, in § 311 BGB explizit anzuführen. Unter anderem sollte die Aufnahme von Vertragsverhandlungen über ein faires und effizientes Datenzugangsregime Bestandteil einer solchen allgemeinen Treuepflicht sein. Im Übrigen sollte geprüft werden, ob darüber hinaus Maßnahmen erforderlich sind, welche von punktuellen Klauselverboten in B2B-Geschäften über ein dispositives Datenschuldrecht bis zu sektorspezifischen Datenzugangsrechten rangieren könnten.

30

Die DEK sieht großes Potenzial in **Konzepten offener Daten des öffentlichen Sektors** (Open Government Data, OGD) und empfiehlt, solche Konzepte auszubauen und zu fördern. Sie empfiehlt eine Reihe von Maßnahmen, die einen teilweise noch nicht ganz vollzogenen **Bewusstseinswandel öffentlicher Stellen** befördern und das Teilen von Daten im Rahmen von OGD-Konzepten praktisch erleichtern könnten. Dazu gehört neben der Etablierung entsprechender **Infrastrukturen** (z. B. Plattformen) auch eine Harmonisierung und punktuelle Ergänzung des derzeit zersplitterten und nicht in jeder Hinsicht konsistenten **Rechtsrahmens**.

31

Allerdings sieht die DEK auch ein schwer zu lösendes Spannungsverhältnis zwischen der Diskussion um OGD (mit Prinzipien wie „offen by default" und „offen für alle Zwecke") einerseits und um besseren Schutz von Geschäftsgeheimnissen und personenbezogenen Daten (mit gesetzlichen Vorgaben wie „Datenschutz by default") andererseits. Sie plädiert dafür, in Zweifelsfällen zugunsten des staatlichen Schutzauftrags zu entscheiden, der in Bezug auf Daten, welche Einzelne oder Unternehmen dem Staat – oft nicht freiwillig – anvertraut haben (z. B. Steuerdaten), besteht. Diesem **staatlichen Schutzauftrag** ist durch eine Reihe von Maßnahmen nachzukommen, die auch technische und rechtliche Schutzvorkehrungen gegen Missbrauch umfassen.

32

In diesem Zusammenhang wird insbesondere empfohlen, für das Teilen von Daten durch den öffentlichen Sektor **Standardlizenzen und Modellkonditionen** zu entwickeln und – mindestens sektorspezifisch – deren Verwendung bindend vorzuschreiben. Diese sollten klar definierte Garantien für die Rechte betroffener Dritter enthalten. Ferner sollten sie Mechanismen vorsehen, die geeignet sind, eine gemeinwohlschädigende Nutzung der Daten ebenso zu verhindern wie eine wettbewerbsrechtlich unerwünschte Verstärkung bestehender Marktmacht oder eine Doppelbelastung des Steuerzahlers.

33

Betreffend **Konzepte offener Daten im privaten Sektor** sollte in erster Linie auf die **Ermutigung und Förderung eines freiwilligen Teilens** von Daten gesetzt werden. Dabei ist nicht nur an Infrastrukturen (z. B. Plattformen) zu denken, sondern auch an eine breite Palette möglicher Anreizstrukturen, etwa bei der Besteuerung, bei öffentlichen Ausschreibungen, bei Förderprogrammen oder bei Genehmigungsverfahren. Gesetzliche Datenzugangsrechte und korrespondierende Zugangsgewährungspflichten sollten dagegen erst in zweiter Linie in Betracht gezogen werden.

34

Insgesamt rät die DEK bei allgemeinen gesetzlichen Datenzugangsrechten zu einem behutsamen Vorgehen, idealerweise **zunächst in ausgewählten Sektoren**. Beispielsweise könnte ein Bedarf im Nachrichten-, Mobilitäts- oder Energiesektor geprüft werden. Dabei sind jeweils alle möglichen Konsequenzen einer Zugangsgewährungs- oder gar Offenlegungspflicht sorgsam zu bedenken und gegeneinander abzuwägen, angefangen von möglichen Implikationen für den Datenschutz und Schutz von Geschäftsgeheimnissen, über Folgen für Investitionsentscheidungen und die Verteilung von Marktmacht bis hin zu den strategischen Interessen deutscher und europäischer Unternehmen im Verhältnis zu Unternehmen in Drittstaaten.

35

Die DEK empfiehlt, Zugangsgewährungspflichten privater Unternehmen **zugunsten gemeinwohlorientierter Zwecke und des öffentlichen Sektors** (Business-to-Government, B2G) in Erwägung zu ziehen. Auch diesbezüglich dürfte indessen ein behutsames und sektorspezifisches Vorgehen anzuraten sein.

Teil F

Algorithmische Systeme

1. Charakteristika algorithmischer Systeme

Zahlreiche Produkte und Anwendungen – von der Sprachassistenz über die automatisierte Kreditvergabe bis hin zum „autonomen“ Fahrzeug – basieren heute auf mehr oder weniger „intelligenten“ Algorithmen. Gerade aufgrund der Vielfalt der Erscheinungsformen derartiger Techniksysteme empfiehlt es sich aus Sicht der DEK, beim ethischen und rechtlichen Zugriff auf die Materie vom **allgemeinen Begriff des algorithmischen Systems** auszugehen (→ oben Teil C, 2.2.5). Die Leitfragen der Bundesregierung zu den Themenbereichen „Algorithmenbasierte Prognose- und Entscheidungsprozesse“ und zur „Künstlichen Intelligenz“ werden daher im Folgenden gemeinsam als Fragen des Umgangs mit algorithmischen Systemen diskutiert.

Bei der ethischen und rechtlichen **Bewertung einzelner algorithmischer Systeme** müssen allerdings insbesondere **folgende Differenzierungen** berücksichtigt werden:

- In **technischer Hinsicht** weisen algorithmische Systeme unterschiedliche Eigenschaften auf. Das Spektrum reicht von Systemen, die vollständig deterministisch operieren, bis hin zu Systemen, die im Wege maschinellen Lernens eigenständig Handlungspläne entwickeln, um das vom Betreiber des algorithmischen Systems vorgegebene Ziel zu erreichen.

- Im algorithmischen System als sozioinformatisches System können ethisch und rechtlich relevante Vorgänge auf **unterschiedlichen Systemebenen** angesiedelt sein, d. h. auf der Ebene der Datenbasis, des Algorithmus im technischen Sinne bis hin zur Ebene der an der Entwicklung, Implementierung, Bewertung oder Korrektur des Systems beteiligten Menschen.

- **Zweck und Folgen** des Einsatzes algorithmischer Systeme differieren erheblich. Soweit algorithmische Systeme menschliche Entscheidungen und Prognosen unterstützen oder ersetzen, wirken sie oft unmittelbar auf die Rechte und Interessen von Individuen ein. Als Beispiele können die automatisierte Kreditvergabe und der automatisierte Verwaltungsakt dienen. Algorithmische Systeme finden aber auch dort Verwendung, wo sich ein derartiger Bezug zu menschlichen Entscheidungen allenfalls mittelbar herstellen lässt. Letzteres ist etwa bei verschiedenen für das „autonome“ Fahren konstitutiven Prozessen oder bei sog. Predictive Maintenance im Maschinenbau der Fall.

- Je nach Einsatzkontext berühren algorithmische Systeme unterschiedliche **ethische und rechtliche Prinzipien**. So wirft bei „autonom“ agierenden cyber-physischen Systemen üblicherweise das äußerlich sicht- und spürbare „Verhalten“ der Systeme Fragen auf. Dieser Aspekt steht etwa bei der Diskussion um den Einsatz von Robotik in der Pflege im Vordergrund. Für die Beurteilung dieser Systeme sind in erster Linie Prinzipien wie der Grundsatz menschenzentrierten Designs maßgeblich. Dort, wo algorithmische Systeme nicht in ähnlicher Form „verkörperlicht“ sind, ist es hingegen vielfach der äußerlich nicht sichtbare Weg zur „Entscheidung“ des Systems, dem die Aufmerksamkeit gilt. Diskutiert wird dabei etwa über die Transparenz der Systeme oder um den Grundsatz menschlicher Letztentscheidung gemäß Art. 22 DSGVO. Ein Beispielsfall hierfür ist die automatisierte Kreditwürdigkeitsprüfung. Die Unterscheidung von „verhaltens-“ und „entscheidungs-“orientierter Perspektive relativiert sich allerdings bei näherer Betrachtung. Denn jedem sichtbaren „Verhalten“ eines Systems ist zu irgendeinem Zeitpunkt eine menschliche „Entscheidung“ vorgelagert, etwa bei der Konstruktion des Systems, und jede „Entscheidung“ findet ihre Brisanz gerade darin, dass eine andere Systemkomponente (einschließlich eines menschlichen Akteurs) ihr „Verhalten“ daran ausrichtet.

Insbesondere dort, wo algorithmische Systeme eng in menschliche **Entscheidungsprozesse eingebunden** sind, bietet es sich aus Sicht der DEK an, **weitere Differenzierungen** vorzunehmen. Ein Algorithmus selbst kann keine Entscheidung im ethisch gehaltvollen Sinne treffen, da er aus sich heraus keine wertebasierten Präferenzen hat. Je nach der konkreten Aufgabenverteilung zwischen menschlichen Akteuren und Maschine lassen sich drei verschiedene Stufen des Einbezugs von algorithmischen Systemen in menschliche Entscheidungen unterscheiden:

- **Algorithmen*basierte*** Entscheidungen sind menschliche Entscheidungen, die sich auf algorithmisch berechnete (Teil-)Informationen stützen. Beispiele sind klinische Entscheidungsunterstützungssysteme, die anhand von Patientendaten aus der elektronischen Patientenakte und auf der Grundlage einer Auswertung der wissenschaftlichen Literatur dem Arzt Behandlungsempfehlungen geben. Der Arzt trifft dann unter Berücksichtigung dieser Empfehlung mit dem Patienten gemeinsam die Entscheidung, welche Behandlung letztlich gewählt wird. Algorithmenbasierte Entscheidungen können gleichwohl auf subtile Weise menschliche Entscheidungen im Ergebnis signifikant beeinflussen, etwa wenn das algorithmische System Informationen über Menschen/Objekte/Verfahrensweisen zusammenstellt, die eine Wertung enthalten, die dem Anwender nicht bewusst sein muss.

- **Algorithmen*getriebene*** Entscheidungen sind menschliche Entscheidungen, die durch die Ergebnisse algorithmischer Systeme in einer Weise geprägt werden, dass der tatsächliche Entscheidungsspielraum und damit die Selbstbestimmung des Menschen eingeschränkt werden, insbesondere, weil sich die Entscheidung nur in algorithmisch ermittelten und vorgegebenen Bahnen bewegen kann. Als Beispiel kann eine Anwendung aus dem Bereich Industrie 4.0 dienen, bei denen in der Mensch-Maschine-Interaktion ein robotisches System dem am Fertigungsprozess beteiligten Menschen nur begrenzte Handlungsspielräume eröffnet.

- **Algorithmen*determinierte* und damit vollständig automatisierte** Entscheidungen erfolgen *prima facie* unabhängig von einem menschlichen Akteur. Vielmehr führen die Ergebnisse eines algorithmischen Systems automatisiert zu Konsequenzen, so dass keine ausdrückliche menschliche Entscheidung mehr erfolgt. Anwendungsbeispiele reichen von Preisdifferenzierungen im Online-Handel über den voll-automatisierten Verwaltungsakt bis hin zu sog. autonomen Waffensystemen. Menschliche Entscheidungen sind gleichwohl involviert, da Menschen darüber entschieden haben, die algorithmischen Systeme zu diesen Zwecken und in dieser Weise einzusetzen.

Beispiel 1

Anhand eines algorithmischen Systems im Rahmen der Auswahl von Bewerbern für einen Arbeitsplatz können die Unterschiede veranschaulicht werden: Im Falle eines algorithmischen Systems, das dem auswählenden Arbeitgeber lediglich Informationen zu den einzelnen Bewerbern zusammenstellt, auf deren Grundlage dieser dann seine Entscheidungen trifft, handelt es sich um ein algorithmenbasiertes Entscheidungsverfahren. Das System führt zu algorithmengetriebenen Entscheidungen, sobald die dem Arbeitgeber übermittelten Informationen eine Bewertung der einzelnen Bewerber (etwa ein Ranking) enthalten, da dieses die Auswahlwahrscheinlichkeiten für einzelne Bewerber signifikant beeinflussen kann. Noch deutlicher wird die faktische Beschränkung der Entscheidungsmöglichkeiten des Arbeitgebers, wenn das System bereits eine Vorauswahl unter den Bewerbern trifft, so dass der Arbeitgeber einzelne Bewerbungen gar nicht mehr zur Kenntnis nimmt. Bei einem algorithmendeterminierten Auswahlprozess würde die Nachricht über die Annahme oder Ablehnung einer Bewerbung automatisiert durch das algorithmische System erfolgen, ohne dass ein Mensch die Auswahl noch einmal überprüft.

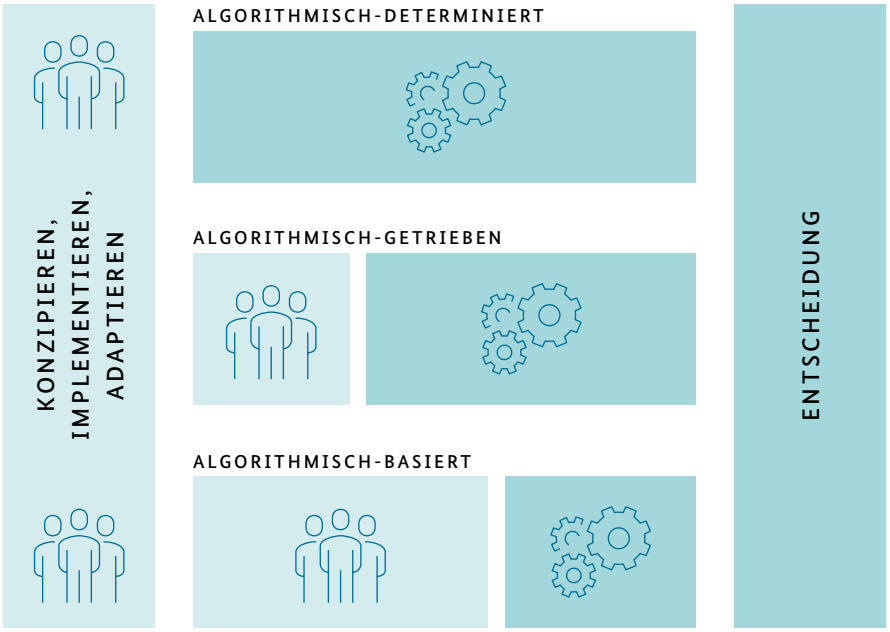


Abbildung 7:
Charakteristika algorithmischer Systeme

Die Zuordnung eines algorithmischen Systems zu einer der drei Formen ist vielfach schwierig, und es sind **Mischformen** innerhalb einer komplexen Softwarearchitektur möglich. Auch kann je nach Wirkweise des Systems der Determinierungsgrad für menschliche Akteure im selben Punkt unterschiedlich hoch sein. So ist im obigen Beispiel ein Entscheidungsprozess, in dem ein algorithmisches System vorab einzelne Bewerber ausfiltert und diesen absagt, aus Sicht dieser aussortierten Bewerber algorithmendeterminiert, für alle verbleibenden Bewerber hingegen algorithmengetrieben.

Hinzu kommt, dass es aufgrund von sog. **Automation Bias- und Default-Effekten** in der praktischen Handhabung der Systeme zu **Überschneidungen** kommen kann. Selbst im Fall algorithmenbasierter Entscheidungen, bei denen der Mensch die volle Entscheidungshoheit hat, kann er dazu tendieren, ohne ausreichend kritische Prüfung der Empfehlung des algorithmischen Systems einfach zu folgen, da er sich ansonsten einem unbequemen Rechtfertigungszwang ausgesetzt fühlt und subjektiv den Eindruck hat, dass sich das Risiko der Vorwerfbarkeit einer Fehlentscheidung erhöht. Gleichwohl ist die grundsätzliche Unterscheidung für die Zuordnung von Verantwortung für eine Risikobestimmung und damit auch für eine Regulierung relevant.

2. Allgemeine Anforderungen an algorithmische Systeme

Maßstab für die Gestaltung und den Einsatz algorithmischer Systeme sind die **allgemeinen ethischen und rechtlichen Grundsätze und Prinzipien**, zuvörderst die Würde des Menschen (→ oben Teil B, 3). Im Sinne des **Grundsatzes vorausschauender Verantwortung** sind bei der Bewertung konkreter algorithmischer Systeme die beabsichtigten und unbeabsichtigten Auswirkungen auf die Nutzer sowie auf die vom Einsatz eines algorithmischen Systems betroffenen Personen zu bedenken. Insbesondere mit Blick auf die Netzwerk-, Skalen- und Verbundeffekte sind, je nach Einsatzzweck und Anwendungskontext, auch gesellschaftliche Folgewirkungen zu reflektieren und vorausschauend zu berücksichtigen. Diese reichen von den positiven Effekten sozialer Innovationen bis hin zu (teilweise subtilen) negativen Effekten etwa auf Vielfalt und Kultur der gesellschaftlichen Debatte als wesentlicher Bedingung für eine funktionierende Demokratie. Hieraus lassen sich nach Auffassung der DEK die folgenden, für die Gestaltung und den Einsatz algorithmischer Systeme zentralen Anforderungen ableiten, die – im Sinne der hier eingenommenen **Governance-Perspektive** – im Zusammenspiel insbesondere von Entwicklern, Unternehmen, Nutzern und staatlichen Stellen umzusetzen sind.

2.1 Menschenzentriertes Design

Im Mittelpunkt steht das Gebot, ein **menschenzentriertes und werteorientiertes Design** algorithmischer Systeme anzustreben, das die grundlegenden Rechte und Freiheiten berücksichtigt. Die Zentrierung auf den Menschen hat nach Ansicht der DEK den gesamten Design-Prozess zu durchdringen. Sie ist durch eine breite Palette unterschiedlicher Maßnahmen sicherzustellen, zu denen auch und gerade die **inklusive und partizipative Entwicklung** von algorithmischen Systemen gehören kann.

Menschenzentriertes Design verlangt insbesondere Veränderungen der Selbstwahrnehmung und Selbstgestaltung infolge der Konfrontation des Einzelnen mit algorithmischen Systemen Rechnung zu tragen. Dabei sind etwa auch Kompetenzgewinne und -verluste im Umgang mit den Systemen, Auswirkungen auf die eigene Lebensweise und die Urteilsbildung sowie auf das körperliche Wohlbefinden schon bei der Entwicklung der Systeme zu berücksichtigen.

Augenmerk ist aber nicht zuletzt auch auf das **emotionale Wohlbefinden** betroffener Personen zu richten, die bei Einsatz menschlicher Akteure und herkömmlicher Technologie anders (niedriger oder auch höher) sein mag als bei Einsatz algorithmischer Systeme. Dies ist nicht nur für die von einer Entscheidung betroffene Person, sondern auch auf der Anwenderseite bedeutsam. Dabei ist u. a. zu berücksichtigen, dass unmittelbare zwischenmenschliche Interaktion eine Vielzahl von Funktionen erfüllt, die weit über das Fällen „guter Entscheidungen“ hinausgehen.

Beispiel 2

Bei der Unterstützung medizinischer Diagnosen durch algorithmische Systeme ist zuvörderst die Treffsicherheit der Diagnose als Einsatzzweck zu identifizieren. Allerdings ist auch das vielfach ausgeprägte Bedürfnis nach menschlicher Zuwendung im Therapiegespräch (mit entsprechender Bedeutung für den Therapieerfolg) nicht außer Acht zu lassen und ebenso das Bedürfnis des Arztes, die eigene ärztliche Erfahrung einbringen zu können. Umgekehrt mag es in bestimmten Situationen – etwa bei schambesetzten Symptomen – für Patienten sogar angenehmer sein, sich nicht primär einem menschlichen Gegenüber anvertrauen zu müssen.

Zu diesen Funktionen gehören etwa: die Befriedigung eines menschlichen Grundbedürfnisses nach **Kommunikation**; das Gefühl, das Gegenüber prinzipiell in seinen Denk- und Reaktionsweisen einschätzen zu können und vom Gegenüber verstanden zu werden, die Chance, das Gegenüber vom eigenen Standpunkt noch überzeugen zu können, sowie der gewisse Kontrolleffekt, der dadurch entsteht, dass das menschliche Gegenüber unmittelbar mit der Reaktion des von einer Entscheidung Betroffenen konfrontiert wird.

Beispiel 3
Emotionale Aspekte spielen auch beim Einsatz algorithmischer Systeme in der Mensch-Maschine-Interaktion eine wichtige Rolle. Beispielsweise kann ein an sich zur Unterstützung der Beschäftigten vorgesehenes System von diesen als invasiv oder bevormundend wahrgenommen werden, weil damit das Verhalten von Beschäftigen analysiert wird, ihnen bestimmte lieb gewonnene Tätigkeiten abgenommen werden oder ihnen suggeriert wird, dass die eigene Leistungsfähigkeit im Vergleich zum „Kollegen Roboter" unterlegen ist.

Das Wohlbefinden aller von einer Technologie Betroffenen, so etwa beim Einsatz von Robotik in der Pflege, ist ein zentraler Leitwert, der bei ethischer Technikgestaltung unbedingt berücksichtigt werden muss. Wichtig ist hierbei, dass Wohlbefinden höchst subjektiv und nicht statisch ist, sondern sich in Abhängigkeit des Kontexts und im Verlauf der Zeit verändern kann und daher einer **ständigen Neubewertung** bedarf.

2.2 Vereinbarkeit mit gesellschaftlichen Grundwerten

Je nach Einsatzgebiet können die Auswirkungen algorithmischer Systeme gesamtgesellschaftliche Relevanz haben, etwa auf die **demokratische Willensbildung**, die **Bürgernähe** staatlichen Handelns, auf den **Wettbewerb**, auf die **Zukunft der Arbeit** und auch auf die **digitale Souveränität** Deutschlands und Europas.

Beispiel 4
Bei der Entwicklung intelligenter Systeme haben diejenigen Anbieter eine privilegierte Startposition, die ihre Geschäftsmodelle auf großen Datenmengen aufbauen können, da viele Anwendungen algorithmischer Systeme auf eben solche Datenmengen angewiesen sind. Je mehr Daten durchforstet werden können, um so eher lassen sich Zusammenhänge und Erkenntnisse generieren. Zusammengenommen mit den für Plattformmärkte typischen Netzwerk-, Skalen- und Verbundeffekten beginnt sich ab einer gewissen Schwelle die Marktmacht von Unternehmen zu verfestigen, und es bilden sich Monopole. Dies versetzt Unternehmen schließlich in die Lage, den Marktzutritt neuer Akteure zu behindern und die marktregulierenden Kräfte des Wettbewerbs zu beeinträchtigen. Je nach Anwendungsbereich können Unternehmen dann gesellschaftliche Meinungsbildungsprozesse und Marktverhalten steuern. Um dem entgegenzuwirken und Rahmenbedingungen für einen fairen Wettbewerb zu schaffen, müssen die wettbewerbsrechtlichen Kontrollmechanismen neu justiert und gegebenenfalls nachgeschärft werden.

Diese überindividuellen Folgewirkungen lassen sich nach Auffassung der DEK regelmäßig nicht allein durch staatliche Stellen und mit den Mitteln des Rechts in den Griff bekommen. Sie müssen vielmehr in allen Phasen der Gestaltung und des Einsatzes algorithmischer Systeme mitbedacht werden. **Entwickler, Unternehmen und Nutzer haben insoweit eine gesellschaftliche (Mit-)Verantwortung**. Insbesondere dort, wo entsprechende Folgewirkungen naheliegen, etwa im Falle algorithmischer Systeme, die die demokratierelevante Kommunikation zwischen Menschen berühren, bedarf es bereits im Gestaltungsprozess sorgfältiger Abschätzungen der Zwecke und der nicht-intendierten Nebenfolgen des Systems und die Prüfung der Frage, in wie weit die Funktion des Systems die Funktion der Demokratie, Grundrechte, das Sekundärrecht oder die Grundregeln des Rechtsstaats berühren kann. Soweit möglich, sollte sich bei der Technikgestaltung eine Kultur des „Einbaus" der Grundprinzipien von Demokratie, Rechtsstaatlichkeit und Grundrechten in die Systemarchitektur etablieren.

Vieles im Zusammenspiel von Technik und Gesellschaft ist bisher freilich noch im Unklaren. Aus Sicht der DEK bedarf es daher vermehrter Forschungsanstrengungen, um die gesellschaftlichen Auswirkungen algorithmischer Systeme aufzuhellen und entsprechende Strategien zur Einhegung negativer Folgen zu entwickeln.

2.3 Nachhaltigkeit bei Gestaltung und Einsatz algorithmischer Systeme

Die Bewertung der individuellen und gesellschaftlichen Folgen algorithmischer Systeme muss auch eine zeitlich übergreifende und globale Perspektive einnehmen. Bei der Entscheidung über den Einsatz und die Gestaltung algorithmischer Systeme sind daher insbesondere auch Aspekte der **Nachhaltigkeit** und des **menschlichen Kompetenzerhalts** zu berücksichtigen. Diese sind wichtig für verbleibende menschliche Kontrollfunktionen (z. B. sog. Human-in-the-Loop-Prinzip), für den Ausfall algorithmischer Systeme in Ausnahmesituationen (z. B. im Katastrophenfall oder bei Cyberangriffen) und für die Innovationskraft künftiger Generationen (z. B. Entwicklung neuer digitaler Technologien). Es ist dabei in erster Linie eine Frage der Aus- und Fortbildung sowie der Bildung im Sinne eines lebenslangen Lernens, für entsprechende generelle Kompetenzen auch künftiger Generationen zu sorgen und schon die Ausbildung nicht auf die reine Anwenderperspektive zu beschränken.

Bildung und Förderung digitaler Kompetenzen fördern auch **soziale Nachhaltigkeit**. Gesellschaftliche Rahmenbedingungen etwa im Sinne von Institutionen und Verfahren sind so auszurichten, dass eine partizipative und inklusive Gestaltung algorithmischer Systeme und ihr dem Gemeinwohl dienender Einsatz gefördert werden.

Der Aspekt der nachhaltigen Entwicklung erfasst darüber hinaus die **ökologische Dimension**. Ungeachtet des positiven Beitrags, den algorithmische Systeme zum Umweltschutz leisten können, ist eine Minimierung des Bedarfs an elektrischer Energie und an bestimmten Ressourcen wie etwa „seltenen Erden“ sowie ihr effizienter Einsatz eine zentrale ethische Forderung.

Ökonomische Nachhaltigkeit erfordert eine Pespektive, die über ausschließlich kurzfristige wirtschaftliche Gewinne hinausweist und auch die langfristigen Auswirkungen berücksichtigt. So kann kurzfristiger kommerzieller Erfolg langfristig zu katastrophalen Auswirkungen führen, wie etwa die Weltfinanzkrise vor einigen Jahren gezeigt hat. Dies soll die Freiheit wirtschaftlicher Betätigung nicht einschränken, aber das Augenmerk auf die Verantwortung lenken, die im Rahmen einer sozialen Marktwirtschaft mit wirtschaftlichem Handeln verbunden ist.

Das Prinzip vorausschauender Verantwortung sowie Erwägungen der Gerechtigkeit und Solidarität sind im Hinblick auf Nachhaltigkeit bei der Gestaltung und dem Einsatz algorithmischer Systeme besonders zu berücksichtigen. Ebenso wie im Hinblick auf den Umgang mit Daten hat die **Risikofolgenabschätzung** für die ökologische, ökonomische und soziale Nachhaltigkeit bei der Gestaltung und dem Einsatz algorithmischer Systeme eine unverzichtbare Bedeutung.

2.4 Hohes Maß an Qualität und Leistungsfähigkeit

Algorithmische Systeme müssen gut und zuverlässig funktionieren, um die mit ihrer Hilfe verfolgten Zwecke zu erreichen. Dienen die Systeme dazu, ethisch wertvolle Zwecke zu befördern, kommt technischen und rechtlichen Vorgaben, die die **Hebung, Fortentwicklung und Sicherung des Stands der Technik** anstreben, eine ethische Qualität zu. Dort, wo die Systeme menschliche Aktivitäten unterstützen oder ersetzen, verbindet sich mit ihnen die Perspektive, auf diese Weise – unbeschadet des Eigenwerts menschlichen Handelns – ethische Grundsätze besser als bisher zu verwirklichen.

Beispiel 5
Ein ethisch vertretbarer Einsatz algorithmischer Systeme im medizinischen Bereich setzt zunächst eine entsprechende medizinische Qualität der Technologien voraus, d. h. die Richtigkeit der Befunderhebung, die Treffsicherheit der Diagnose, die Erfolgswahrscheinlichkeit der empfohlenen Therapie oder die Erfolgsquote bei einem medizinischen Eingriff etc. müssen beim Einsatz des Systems grundsätzlich mindestens gleich gut und – angesichts des sensiblen Einsatzkontextes – idealerweise besser sein als beim Einsatz herkömmlicher Technologien und menschlicher Akteure.

Die Steigerung von Qualität und Leistungsfähigkeit kann durch ganz unterschiedliche Maßnahmen erfolgen. Dazu gehören beispielsweise adäquate Risikomodelle, eine möglichst inklusive und partizipative Standardentwicklung, systemische Management- und Kontrollansätze sowie ein Prozessdesign, das auf stetige Verbesserung des Gesamtsystems hin ausgerichtet ist. Die Rolle jener menschlichen Akteure, die Teil des als sozioinformatisches Ensemble verstandenen algorithmischen Systems sind (→ oben 1), muss in diesem Kontext stets mitbedacht werden. Denn nach wie vor entfalten etliche algorithmische Systeme ihre Leistungsfähigkeit gerade im Zusammenspiel mit kritischen und fachkundigen Menschen. Teil einer an Qualität orientierten Systemgestaltung sind daher auch Mechanismen, die zur **Steigerung der menschlichen Fähigkeiten** beitragen und einem Abbau von Kompetenzen und kritischer Reflektionsfähigkeit und -bereitschaft, etwa im Zusammenhang mit einem Automation Bias, vorbeugen bzw. entgegenwirken. Beispiele für ein produktives und kompetenzerhaltendes Zusammenspiel von Mensch und Maschine finden sich etwa bei der algorithmengestützen bildgebenden Diagnostik im medizinischen Bereich.

2.5 Gewährleistung von Robustheit und Sicherheit

Algorithmische Systeme müssen robust und sicher sein, sonst lassen sich die mit ihnen verfolgten legitimen Zwecke nicht oder nur unter Inkaufnahme potenzieller Schäden an ethischen und rechtlichen Gütern und schutzwürdigen Interessen erreichen. Aus ethischer Sicht partizipiert das Postulat robuster und sicherer Systemgestaltung und eines entsprechenden Systemeinsatzes daher an der Wertigkeit der jeweiligen Systemzwecke sowie am Schutzbedarf der vom System verwendeten Daten. Aus diesem Grund sind allerdings auch die Anforderungen an die Robustheit und Sicherheit nicht für alle Systeme identisch. Die spezifischen Anforderungen können vielmehr je nach dem **konkreten Schutzbedarf und dem Einsatzkontext** verschieden ausgeprägt sein.

Beispiel 6
Nicht belastbare oder unsichere Systeme, die in Steueranlagen eingesetzt werden, können unmittelbar Personen oder die Umwelt bedrohen, etwa wenn sie den Schadstoffausstoß von Industrieanlagen regeln, Roboter steuern oder autonome Fahrzeuge im Verkehr lenken. Ein Fehlversagen kann hier sogar zu Schäden für wichtige Rechtsgüter wie Leib und Leben führen. Um dies zu verhindern, gilt es, Prozesse zu initialisieren, die den gegenwärtigen Stand der Technik definieren, Rechtsnormen zu erlassen, die die Orientierung am Stand der Technik verbindlich machen, und Maßnahmen zu implementieren, die die effektive Durchsetzung des Standards garantieren.

Robuste und sichere Systemgestaltung umfasst sowohl die **Sicherheit des Systems** gegen Einflüsse von außen (z. B. durch Verschlüsselung, Anonymisierung etc.) als auch den **Schutz der Menschen und der Umwelt vor negativen Einflüssen durch das System** (insbesondere durch einen systematischen Risikomanagementansatz, z. B. auf der Grundlage einer Risikofolgenabschätzung). Sie muss zudem alle Phasen der Datenverarbeitung und alle technischen und organisatorischen Komponenten einbeziehen. Risiken können sich dabei nicht nur aus der technischen Gestaltung, sondern auch aus Fehlern ergeben, die menschliche Entscheidungen im Umgang mit algorithmischen Systemen mit sich bringen. Da algorithmische Systeme und ihre Einbettung in die sonstige Informationstechnik einer Organisation nicht statisch sind, wird zudem ein **Managementsystem** benötigt, das die Wirksamkeit der Maßnahmen angesichts veränderter Bedingungen, beispielsweise neu bekannt gewordener Risiken, überprüft und sicherstellt.

2.6 Minimierung von Bias und Diskriminierung als Vorbedingung gerechter Entscheidungen

Ein wesentliches Ziel der Regulierung algorithmischer Systeme besteht darin, sicherzustellen, dass die den algorithmischen Systemen zu Grunde liegenden Entscheidungsmuster keine systematischen Verzerrungen (Biases) aufweisen, die zu diskriminierenden und ungerechten Entscheidungen führen. Dabei ist zunächst festzuhalten, dass verzerrte, diskriminierende oder ungerechte Entscheidungen auch bei Einsatz herkömmlicher Technologien und menschlicher Akteure zu beobachten sind. Im Gegensatz zu vorurteilsbehafteten Entscheidungen einzelner Menschen besteht bei algorithmischen Systemen aber die Gefahr, dass der einem System inhärente Effekt über eine skalenmäßig große Anwendung des Systems eine Breitenwirkung entfaltet, die einzelne menschliche Entscheider nie erreichen könnten. Vor diesem Hintergrund ist die Diskussion um Bias und Diskriminierung durch algorithmische Systeme nach Auffassung der DEK **auch als Chance zu begreifen**, in bestehenden Entscheidungskontexten bereits bestehende Probleme aufzudecken und ganz allgemein zu besseren Entscheidungsprozessen zu gelangen.

Beispiel 7

Ein zur Erkennung von Hautkrebs eingesetztes algorithmisches System wurde vorwiegend an Patienten weißer Hautfarbe trainiert und die Wahrscheinlichkeit einer korrekten Erkennung von Hautkrebs ist bei Patienten mit weißer Hautfarbe daher signifikant höher als bei Patienten mit anderer Hautfarbe. Als Medizinprodukt würde ein solches System nur für die Anwendung an weißhäutigen Patienten zugelassen werden. Der gleiche Effekt wäre freilich zu verzeichnen, wenn ein Dermatologe seine Ausbildung und klinische Praxis allein in einem bestimmten Kulturkreis erworben hat. Letztlich ist in beiden Fällen darauf zu achten, dass alle Patienten unabhängig von ihrer Hautfarbe medizinisch gut versorgt werden.

Auch in Fällen, in denen bei der Entwicklung algorithmischer Systeme keine unmittelbare Diskriminierungsabsicht vorliegt, kann es zu diskriminierenden Entscheidungen kommen, also zu solchen, die bestimmte Gruppen ungerechtfertigterweise systematisch benachteiligen. Insbesondere bei Maschinellem Lernen rührt das Problem vielmehr daher, dass die Systeme anhand vorhandener Daten Modelle erlernen. Die daraus resultierenden Prognosen und Empfehlungen **schreiben die Vergangenheit in die Zukunft fort**, wodurch bestehende gesellschaftliche Ungerechtigkeiten durch den Einbau in scheinbar neutrale Technologien verschleiert und potenziell verstärkt werden können.

Beispiel 8

Ein zur Bewertung von Bewerbungen um eine Führungsposition eingesetztes algorithmisches System wurde mit den Daten derjenigen Führungskräfte trainiert, die sich im betreffenden Unternehmen in den letzten Jahrzehnten bewährt haben. Da in den letzten Jahrzehnten vorwiegend männliche Führungskräfte eingestellt wurden, bewertet das System, das mit diesem Datensatz trainiert wurde, männliche Bewerber durchgehend besser als gleich qualifizierte Bewerberinnen.

Unter dem englischen Stichwort **Bias** versammeln sich eine **Vielzahl systematischer Verzerrungen**, die unterschiedlicher Natur sind und unterschiedliche Ursachen haben. Bei menschlichen Akteuren geht es sowohl um kognitive Verzerrungen, als auch um gesellschaftliche Vorannahmen, Vorurteile oder Stereotypen, welche Entscheidungsfindungen negativ beeinflussen können. In Bezug auf algorithmische Systeme kann sich Bias auf die technische Abbildung eben jener gesellschaftlichen Vorannahmen, Vorurteile oder Stereotypen beziehen. Diese Abbildung kann v.a. im Kontext von Maschinellem Lernen an mehreren Stellen erfolgen. Häufig führt eine ungenügende Repräsentativität oder eine geringe Fallzahl einer gesellschaftlichen Gruppe in den Trainingsdaten zu Verzerrungen, indem die Spezifika dieser Gruppe im Rahmen der Entwicklung nicht ausreichend erkannt und damit berücksichtigt werden. Jenseits der verwendeten Trainingsdaten können auch andere technisch-methodische Entscheidungen, z. B. bzgl. der Zielvariablen oder Labels, zu diskriminierenden Modellen und dadurch ungerechten Entscheidungen führen. Zuletzt können sich auch erst im Einsatz von Systemen Probleme ergeben, z. B. wenn algorithmische Systeme unter veränderten gesellschaftlichen Rahmenbedingungen oder in nicht vorhergesehenen Einsatzkontexten genutzt werden.

Besonders kritisch sind unter dem Gesichtspunkt der Diskriminierung algorithmische Systeme, die rechtlich als besonders **sensibel anerkannte Kategorien von Daten** wie Geschlecht oder Herkunft **direkt** verwenden. Eine direkte Verwendung sensibler Informationen kann, je nach Anwendungsgebiet, wichtig für eine korrekte Datenverarbeitung sein und ist – im Rahmen der rechtlichen Grenzen – vielfach auch zulässig.

Beispiel 9
Viele Systeme zur Krankheitsdiagnose kennen und berücksichtigen das Geschlecht oder das Alter eines Patienten. Auch für die Umsetzung von Geschäftsstrategien, etwa dem Ausbau des Geschäftes in einer Altersgruppe, Berufsgruppe oder einer Region, können sensible Merkmale im Rahmen einer Geschäftsentscheidung Verwendung finden, wenn sie beispielsweise ein Kundensegment definieren, für das vereinfachte Annahmekriterien gelten.

Ebenfalls kritisch kann aber auch die Verwendung von Informationen sein, die sensible Kategorien **indirekt** kodieren.

Beispiel 10
Im Rahmen der Schätzung der Kreditwürdigkeit wird das Haushaltseinkommen als Information verwendet. Dieses fällt in Deutschland für die Geschlechter im Mittel unterschiedlich aus. In der Folge kann ein algorithmisches System, welches das Haushaltseinkommen verwendet, zu unterschiedlichen Verteilungen der Schätzungen für die Kreditwürdigkeit von Männern und Frauen gelangen.

Diskriminierung vollständig zu verhindern, ist selbst hinsichtlich rechtlich anerkannter Kategorien wie Geschlecht oder Herkunft im Kontext von algorithmischen Systemen schwierig. Darüber hinaus kann der Einsatz algorithmischer Systeme dazu führen, dass **ganz neue nach mehr oder weniger zufälligen Merkmalen zusammengewürfelte Gruppen** mit einer gewissen Systematik und ohne rechtfertigenden Grund von gesellschaftlichen Gütern ausgeschlossen werden oder mit sonstigen negativen Folgen konfrontiert werden. Vor diesem Hintergrund ist eine Sensibilisierung für komplex bedingte diskriminierende Effekte für alle an der Entwicklung und dem Einsatz eines solchen Systems Beteiligten erforderlich, damit sie solche so weit wie möglich vermeiden oder ihnen gegensteuern können (→ siehe unten).

Allerdings haben technische Maßnahmen zur Minimierung von Diskriminierung selbst bei der Anwendung ständiger Verbesserungsprozesse Grenzen, u. a. weil sich unterschiedliche technische Fairnessziele nicht gleichzeitig erfüllen lassen. Welche Kriterien für Nicht-Diskriminierung und Gerechtigkeit in welchem Kontext angemessen sind, ist keine technische, sondern eine gesellschaftliche und politische Frage. Daher dürfen diese Entscheidungen auch nicht allein den Technik-Entwicklern überlassen werden. Stattdessen muss sie Bestandteil einer künftigen Regulierung algorithmischer Systeme werden und sich in den Betreiberpflichten der Verantwortlichen manifestieren. Bedingung dafür ist, dass die **Kriterien kontextspezifisch und demokratisch ausgehandelt** werden.

Die genaue Analyse algorithmischer Systeme ist schwierig. Um Diskriminierungen erkennen und vermeiden zu können, müssen Verantwortliche und Kontrollstellen die Möglichkeiten haben, sich ein Bild des algorithmischen Systems sowohl im Rahmen seiner Entwicklung als auch im Zuge seines produktiven Einsatzes über eventuell auftretende ungewollte Diskriminierungseffekte zu machen. Durch Verfahren wie **Risikofolgenabschätzung und Output-Analysen** können solchen Effekte identifiziert werden.

Es besteht ein Spannungsverhältnis zwischen den Vorgaben zur Einschränkung in der Erhebung und Speicherung diskriminierender Merkmale und dem Anliegen, dass es möglich bleibt, etwaige diskriminierende Effekte festzustellen oder eine Nicht-Diskriminierung belegen zu können. Diese verschiedenen Anforderungen müssen im Einzelfall in einen Ausgleich gebracht werden, was auch Einfluss auf Tests in verschiedenen Phasen des Lebenszyklus der Systementwicklung haben kann; ein standardmäßiges Mitsammeln von allen potenziell diskriminierenden und damit sensiblen Informationen nur zum Zwecke eines Nachweises, dass aufgrunddessen keine Diskriminierung stattfindet, wäre nicht gerechtfertigt. Hier bedarf es verstärkter Anstrengungen, eine **praktische Konkordanz von Anti-Diskriminierungsrecht und Datenschutzrecht** herzustellen.

2.7 Transparenz, Erklärbarkeit und Nachvollziehbarkeit

Für eine belastbare ethische und rechtliche Bewertung algorithmischer Systeme ist es essenziell, dass ausreichend Informationen über dessen Reichweite, Funktionsweise, Datengrundlage und Datenauswertung zur Verfügung stehen. **Nur ein im Ansatz transparentes System lässt sich darauf überprüfen, ob es einen legitimen Einsatzzweck verfolgt.** Je nach Art und Adressat möglicher Transparenzverpflichtungen kommen dem Transparenzgrundsatz weitere zentrale Funktionen zu. In Bezug auf die Öffentlichkeit muss hinreichend Transparenz hergestellt werden, um eine ausreichende Informationsgrundlage für einen gesellschaftspolitischen Diskurs über algorithmische Systeme führen zu können. Aufsichtsbehörden oder sonstige Kontrollstellen müssen in der Lage sein, entscheiden zu können, ob die rechtlichen und technischen Vorgaben beim Einsatz algorithmischer Systeme eingehalten werden bzw. wurden. Einzelne Bürger müssen informierte und souveräne Entscheidungen bezüglich der Verwendung algorithmischer Systeme treffen können und im Falle von negativen Auswirkungen auf ihre Freiheiten und Rechte beurteilen können, ob und inwiefern Sie von ihren Rechten Gebrauch machen wollen. Auch das ist eine Konsequenz des ethischen Prinzips der digitalen Selbstbestimmung.

Angesichts immer komplexerer Systeme ist die Forderung nach Transparenz in der Praxis allerdings damit konfrontiert, dass es selbst für Fachleute oft kaum mehr möglich ist, alle Einzelkomponenten eines Systems und ihr Zusammenspiel vollständig zu durchdringen und in angemessener Zeit **nachzuvollziehen**. Insbesondere bei einzelnen Methoden des Maschinellen Lernens ist es beim heutigen Stand von Wissenschaft und Technik schwierig, anzugeben, welche Eingabe zu einer spezifischen Ausgabe des Systems geführt hat. Hinzu kommt, dass selbst technisch einfache algorithmische Systeme oftmals in komplexe sozioinformatische Ökosysteme eingebunden sind, d. h. informations- und arbeitsteilige Prozesse, in denen eine Vielzahl von Herstellern und Betreibern mitwirkt.

Beispiel 11
Die Anzeige einer individualisierten Online-Werbung ist das Ergebnis komplexer Prozesse, in denen die Auslieferung und Bezahlung der Werbung auf der Basis von verhaltensbasierter Analyse und Segmentierung erfolgt. Hierzu werden insbesondere sog. Analytics-Dienste genutzt, die webseitenübergreifend durch Einbinden des entsprechenden Programmcodes (wie beispielsweise JavaScript-Code zum Tracking) von den Seitenanbietern eingesetzt werden. Die Komponenten solcher Systeme sind auch nicht statisch, sondern können sich verändern, z. B. wenn Hersteller neue Versionen bereitstellen oder wenn es sich um adaptierende bzw. selbstlernende Systeme handelt.

Auch rechtliche Gesichtspunkte können bestimmten Formen der Offenlegung von Informationen über algorithmische Systeme **Grenzen** ziehen. Quellcodes und Hardware-Designs sind oftmals als Geschäftsgeheimnisse geschützt. Betreiber haben zudem vielfach ein legitimes Interesse daran, Manipulationen an ihren Systemen zu verhindern. Sofern algorithmische Systeme personenbezogene Daten verarbeiten, kann auch das Datenschutzrecht dem Informationsinteresse der Öffentlichkeit oder betroffener Bürger Grenzen ziehen. Sofern es allerdings bei der Transparenzanforderung an die Systeme um die Offenlegung des Quellcodes geht, der als solcher keine personenbezogenen Daten enthält, steht das Datenschutzrecht der Offenlegung nicht entgegen.

Die allgegenwärtige Komplexität kann jedoch das Ziel, algorithmische Systeme transparent zu gestalten, nicht widerlegen oder Intransparenz rechtfertigen. Ebenso wie die erwähnten Rechtsgründe sind sie gleichwohl bei der Ausgestaltung etwaiger Informationsrechte und Transparenzpflichten zu berücksichtigen, die sich am rechtlich und tatsächlich Möglichen orientieren müssen. **Transparenz als Prinzip** verlangt dabei auch, die Technik so fortzuentwickeln, dass eine Offenlegung von Informationen einfacher wird – etwa durch Verwendung von Open-Source-Software und Open-Hardware – und Ansätze zu entwickeln, die Komplexität reduzieren. Hier ist auch die Forschung gefordert. Unter dem Stichwort „Explainable AI“ arbeiten Forscherinnen und Forscher mit wachsendem Erfolg daran, aussagekräftige Erkenntnisse über die internen Prozesse algorithmischer Systeme zu generieren.

Die Forderung nach Transparenz muss stets die **unterschiedlichen Kompetenzniveaus** der potenziell an Transparenz Interessierten berücksichtigen. So kann die Offenlegung des Computercodes Aufsichtsbehörden, die entsprechende Kontrollen vornehmen, ein Verständnis des Systems entscheidend erleichtern. Laien haben hingegen vielfach eher ein Bedürfnis nach klar verständlich aufbereiteten Informationen über grundlegende Eigenschaften des Systems, die es ihnen ermöglichen, eine alltagstaugliche Risikoeinschätzung durchzuführen. Zugleich beschränkt sich ihr Interesse selten auf das System „an sich“. Schon um in Zukunft etwaige negative Entscheidungen zu vermeiden, wird zusätzlich eine **Erklärung** verlangt, wie die sie konkret betreffende Entscheidung zustande gekommen ist und welche Faktoren dabei welches Gewicht entfaltet haben. Die spezifische Ausgestaltung der Vorgaben für Transparenz und Erklärbarkeit sollte sich dabei am Verständnishorizont der Betroffenen orientieren und für diese stets **nachvollziehbar** sein. In diesem Sinne sichern Vorgaben zu Transparenz und Erklärbarkeit die Handlungsfähigkeit und Selbstbestimmung der Bürger.

2.8 Klare Rechenschaftsstrukturen

Ebenso wie die Herrschaft über Daten die Pflicht begründet, für diese Macht Rechenschaft abzulegen, muss auch die Möglichkeit, über algorithmische Systeme zu verfügen, mit der Bereitschaft einhergehen, **für das eigene Handeln Rede und Antwort zu stehen**, d. h. gegebenenfalls auch zu **haften**.

Erneut ist es die Komplexität algorithmischer Systeme, die in der Praxis Verantwortungszuschreibungen erschweren kann. Hersteller der Hard- oder der Software, Datenzulieferer, Algorithmenentwickler, Betreiber einzelner Komponenten, Auftraggeber, Anwender – jeweils als Organisation oder als darin konkret Beschäftigte – leisten ihren Beitrag zum System. Vielfach werden Komponenten verwendet, die sich ohne Kenntnis oder Kontrolle des Einsetzenden verändern können, etwa durch wichtige Updates für die Informationssicherheit. Oftmals sind die Beteiligten zudem an verschiedenen Orten auf der ganzen Welt ansässig. Es bedarf Anstrengungen auf allen Ebenen, um einer Diffusion der Verantwortung entgegenzuwirken und **Rechenschaftsstrukturen zu etablieren**, beginnend bei der technischen Gestaltung der Systeme bis hin zu rechtlichen Vorgaben, etwa in Form des aus dem Datenschutzrecht bekannten Instituts der „gemeinsamen Verantwortlichkeit“ (Artikel 26 DSGVO).

2.9 Ergebnis: Verantwortungsgeleitete Abwägung

Die Bewertung algorithmischer Systeme in ethischer Hinsicht ist **in der Praxis überaus komplex**. Dies ist bedingt durch die Vielzahl der zu berücksichtigenden Faktoren sowie durch die Tatsache, dass in einem konkreten Anwendungsbereich unterschiedliche Individuen jeweils „besser“ und „schlechter“ gestellt werden können. Entsprechendes gilt für gesellschaftliche Folgewirkungen und Nachhaltigkeitsaspekte, die sich selten eindeutig als „positiv“ oder „negativ“ klassifizieren lassen werden. Das bedeutet jedoch nicht, dass der Mensch seine Urteilskraft aufgeben darf. Dort, wo Abwägungen schwierig werden, sind vielmehr alle bei ihren Wertungen und Entscheidungen zur besonderen Sorgfalt angehalten. Dort, wo algorithmische Anwendungen (perspektivisch) eine so überragend große Leistungsfähigkeit und Reichweite entwickeln, dass Fragen über die Zukunft der Menschheit entstehen, geraten Abwägungen der Chancen und Risiken zunehmend an Grenzen und erfordern grundsätzlichere anthropologische und ethische Auseinandesetzungen. Gerade hier ist das Prinzip der vorausschauenden Verantwortung von grundlegender Bedeutung.

Bei alldem stellt der **demokratische Prozess** Mittel und Wege bereit, um einander widersprechende Überzeugungen zum Ausgleich zu bringen – idealerweise unterstützt durch besondere **deliberative Prozesse und Institutionen**, in denen sich die Gesellschaft in einer möglichst inklusiv und partizipativ ausgestalteten Form über den Umgang mit den Herausforderungen durch algorithmische Systeme vergewissern kann.

Nur selten dürfte die Situation gegeben sein, dass eine Abwägung zwischen menschlichem Handeln und dem Einsatz eines algorithmischen Systems verzichtbar ist, weil dieses in allen ethisch relevanten Belangen ein „besseres“ Ergebnis erzielt als menschliche Akteure, die herkömmliche Technologien nutzen. Dort, wo dies der Fall ist, gilt nach Auffassung der DEK allerdings, dass der Einsatz algorithmischer Systeme **ethisch geboten ist**, denn ein genereller ethischer Vorzug menschlichen Handelns vor dem Einsatz von Maschinen zulasten des Schutzes wichtiger Rechtsgüter ist nach Auffassung der DEK nicht gerechtfertigt. Regelmäßig werden bei der Frage, ob menschliches oder maschinelles Handeln zu bevorzugen ist (→ s. dazu auch Teil B, 1), jedoch weitere Faktoren wie etwa emotionales Wohlbefinden von Menschen, menschlicher Kompetenzerhalt und nachhaltige Entwicklung zu berücksichtigen sein, die letztlich doch wieder eine Abwägung erforderlich machen. Diese Abwägung kann zulasten, aber auch zugunsten des algorithmischen Systems ausgehen.

Beispiel 12

Führt der Einsatz eines diagnostischen algorithmischen Systems in einem bestimmten klinischen Bereich dazu, dass nur 2 % der Patienten versterben, während infolge menschlicher Fehldiagnosen 10 % aller Patienten versterben würden, wäre der Einsatz des Systems – je nach den Umständen des Einzelfalls – ethisch geboten, auch wenn dadurch leichte, aber verschmerzbare Einbußen beim emotionalen Wohlbefinden der Patienten einträten und zusätzliche Maßnahmen zum menschlichen Kompetenzerhalt ergriffen werden müssten.

Sofern hingegen nach Berücksichtigung aller Umstände durch den Einsatz eines algorithmischen Systems zulasten wichtiger Rechtsgüter ein schlechteres Ergebnis erzielt wird als bei dem Einsatz herkömmlicher Technologien und menschlicher Akteure – etwa, weil mehr Fehlentscheidungen getroffen werden – und bloß ein Gewinn an Effizienz oder Bequemlichkeit entsteht, ist der Einsatz algorithmischer Systeme im Grundsatz **ethisch abzulehnen**. Ethisch vertretbare Ausnahmen können in diesem Fall aus ökonomischen Erwägungen heraus allerdings ausnahmsweise hinzunehmen sein, wenn einer nur minimalen Beeinträchtigung ein außergewöhnlich hohes Einsparungspotenzial gegenübersteht, das dem Wohle der Allgemeinheit zugute kommt.

3. Empfehlung eines risikoadaptierten Regulierungsansatzes

Aus regulatorischer Sicht legt die Tatsache, dass algorithmische Systeme je nach Einsatzzweck, Leistungsfähigkeit, Robustheit und Sicherheit sowie mit Blick auf ihre Wirkungen ethisch sehr unterschiedlich zu bewerten sind, einen **risikoadaptierten Regulierungsansatz**[1] nahe. Dieser folgt dem Prinzip , dass ein **steigendes Schädigungspotenzial** algorithmischer Systeme mit **wachsenden Eingriffstiefen** der regulatorischen Instrumente einhergeht. Das Risiko-Spektrum algorithmischer Systeme reicht dabei von solchen, deren Anwendung allenfalls ein geringes Risiko birgt, bis hin zu Systemen, die zu irreversiblen Schäden für Individuen und Gesellschaft führen können. Ursache für die Risiken können etwa nicht adäquate Modelle, eine ungeeignete Datengrundlage insbesondere bei selbstlernenden Systemen oder unpassende Grundannahmen und Gewichtungen sein (→ oben 2.3 und 2.6).

Mögliche **Schäden** durch algorithmische Systeme können unterschiedlicher Natur sein, etwa finanziell, immateriell oder physisch. So können einzelne Anwendungen potenziell schwerwiegende finanzielle Schäden verursachen (etwa Kredit- oder Versicherungskonditionen), Chancen der Teilhabe beeinflussen (etwa Diskriminierung bei Stellenvergaben) sowie Grundrechtsverletzungen und Risiken für Leben und Gesundheit von Verbrauchern nach sich ziehen (beispielsweise bei Pflegerobotern oder Mobilitätsanwendungen).

Übergreifendes Ziel rechtlicher Regulierung des Einsatzes algorithmischer Systeme ist es, schädliche Effekte auf individueller und überindividueller Ebene zu vermeiden. Insbesondere dort, wo algorithmische Systeme grundrechtssensible Sachverhalte berühren, bedarf es dazu auch gesetzlicher Vorgaben für die Gestaltung der Systeme. Anzustreben ist dabei eine Regulierung, die so viel wie nötig und zugleich so wenig wie möglich vorschreibt, um Innovation und Kreativität nicht zu behindern, gleichzeitig aber den Schutz grundlegender Rechte, Freiheiten und Werte sichert. Eine **effiziente und sachgerechte Regulierung** kann dazu beitragen, das Vertrauen der Bevölkerung hinsichtlich des Einsatzes algorithmischer Systeme zu stärken. In der öffentlichen Wahrnehmung gelten insbesondere selbstlernende Systeme als nicht kontrollierbar, was zu einer entsprechenden Skepsis vor der Technologie als solcher beiträgt.[2]

Primäre Adressaten der rechtlichen Regulierung sind nach Auffassung der DEK die **Hersteller** und **Betreiber** algorithmischer Systeme. Aufgrund der unmittelbaren Grundrechtsbindung des Staates ist bei der näheren Ausgestaltung der Regulierung allerdings zwischen **privatem und staatlichem Einsatz** algorithmischer Systeme zu differenzieren (→ dazu insb. unten 7). Angesichts des Modell- und Vorbildcharakters staatlichen Handelns wird der Bundesregierung empfohlen, beim Einsatz algorithmischer Systeme für staatliche Zwecke besondere Sorgfalt walten zu lassen.

3.1 Systemkritikalität und Systemanforderungen

Ein risikoadaptierter Regulierungsansatz kann durch die Orientierung an dem Modell der Kritikalität eines algorithmischen Systems konkretisiert werden. Die **Systemkritikalität** setzt am Schädigungspotenzial des Systems an. Dieses bestimmt sich aus der Schwere und der Eintrittswahrscheinlichkeit des zu befürchtenden Schadens.

1 Vgl. hierzu insbesondere Tobias Krafft / Katharina Zweig - Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse, Studie im Auftrag des Verbraucherzentrale Bundesverband e.V. (vzbv), 22.01.2019, S. 18 ff. (abrufbar unter: *https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf*).

2 Sarah Fischer / Thomas Petersen: Was Deutschland über Algorithmen weiß und denkt – Ergebnisse einer repräsentativen Bevölkerungsumfrage, Bertelsmann Stiftung, 2018 (abrufbar unter: *https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/was-deutschland-ueber-algorithmen-weiss-und-denkt/*).

Die **Schwere** zu befürchtender Schäden, etwa im Falle einer Fehlentscheidung, bezieht sich u. a. auf die Wertigkeit der betroffenen Rechtsgüter und Interessen (insbesondere z. B. das Recht auf informationelle Selbstbestimmung sowie freie Entfaltung der Persönlichkeit, das Grundrecht auf Leben und körperliche Unversehrtheit sowie auf Gleichbehandlung) und die Tiefe des potenziellen Schadens durch eine Rechtsverletzung. Zudem ist für die Bestimmung der Schwere des potenziellen Schadens die besondere Sensibilität der verwendeten Daten, die Höhe eines möglichen Schadens für Einzelne oder Gruppen (einschließlich immaterieller Schäden bzw. monetär schwer zu beziffernder Nutzeneinbußen), die Zahl der Betroffenen, die Summe der potenziellen Schäden und der gesamtgesellschaftliche Schaden, der über eine reine Summierung von Einzelschäden weit hinausgehen kann, zu berücksichtigen. Dabei sollen die Auswirkungen des Einsatzes des algorithmischen Systems, je nach Anwendungsbereich, hinsichtlich ihrer ökologischen, sozialen, psychologischen, kulturellen, ökonomischen und juristischen Dimensionen betrachtet werden. Maßstabsetzend für die Wertigkeit sind dabei die allgemeinen ethischen Werte und Prinzipien (→ oben Teil B).

Die **Wahrscheinlichkeit** eines Schadeneintritts hängt auch von den nachfolgenden Systemeigenschaften und Faktoren ab:

- Rolle algorithmischer Berechnungen im Entscheidungsprozess (von der bloßen Inspiration menschlicher Akteure ohne Richtigkeitsanspruch bis hin zur algorithmendeterminierten Entscheidung, → siehe oben 1);
- Komplexität der Entscheidung (vom schlichten deterministischen Abbild der Realität über eine probabilistische Einschätzung der Realität bis hin zur multifaktoriellen und nicht-determinierten Prognose einer künftigen Realität);
- Wirkungen der Entscheidung (von einem bloß abstrakt denkbaren Handlungskontext über einen konkreten Handlungskontext bis hin zur unmittelbaren Implementierung); und
- Reversibilität der Wirkungen (von voller Reversibilität bis hin zur Irreversibilität).

Schwere und Wahrscheinlichkeit zu befürchtender Schäden können auch abhängig sein vom **staatlichen oder privaten Charakter** des Handelnden und – gerade in wirtschaftlichen Zusammenhängen – von der **Marktmacht** desjenigen Akteurs, der sich des algorithmischen Systems bedient, weil der staatliche oder private Charakter sowie die Marktmacht nicht nur für Grundrechtsbindung und gesamtgesellschaftlichen Schaden relevant ist, sondern auch über allfällige Ausweichmöglichkeiten betroffener Personen entscheidet. Mit der **Abhängigkeit der betroffenen Personen** von einem algorithmischen System, etwa hinsichtlich des Zugangs zu Märkten, Gütern und Dienstleistungen, steigt dessen Kritikalität. Die Beschränkung der Auswahlmöglichkeiten kann auf verschiedene Ursachen zurückzuführen sein. Zu nennen sind etwa Netzwerk-, Skalen- und Verbundeffekte, die sich wiederum in Marktmacht und (fehlenden) äquivalenten Alternativangeboten niederschlagen können.

Je höher die Systemkritikalität eines Systems ist, desto höher sind die **Anforderungen**, die aus regulatorischer Sicht an dieses System zu stellen sind. Diese Anforderungen werden insbesondere durch

a) Korrektur- und Kontrollinstrumente;

b) Vorgaben für die Transparenz algorithmischer Systeme und die Erklärbarkeit und Nachvollziehbarkeit der Ergebnisse; und

c) Regelungen zur Zuordnung von Verantwortlichkeit und Haftung im Zusammenhang mit Entwicklung und Einsatz algorithmischer Systeme

ausgestaltet (→ siehe unten 4, 5 und 8).

Die Vielfalt, Komplexität und Dynamik algorithmischer Systeme stellt die Regulierung vor große Herausforderungen. Sie kann sich nicht auf einen beschränkten Instrumentenkasten stützen, sondern muss, je nach Kritikalität des Systems, **auf unterschiedlichen Regulierungsebenen ganz unterschiedliche Korrektur- und Kontrollinstrumente** in Stellung bringen, um die Ziele der Regulierung zu erreichen und die Risiken der Systeme beherrschbar zu machen. Das Spektrum möglicher Instrumente reicht dabei vom Verzicht auf spezialgesetzliche Vorgaben und „weiche" Anreize für Selbstregulierung über behördliche Kontrollrechte bis zum Vorbehalt der menschlichen Letztentscheidung oder dem Verbot bestimmter Einsatzzwecke und -kontexte algorithmischer Systeme.

Zentrale Bausteine eines Korrektur- und Kontrollregimes für algorithmische Systeme sind Vorgaben für die **Transparenz** der Systeme und die **Erklärbarkeit** sowie **Nachvollziehbarkeit** ihrer Ergebnisse (→ oben 2.7). Auch insoweit bestimmt die Kritikalität des Systems die Reichweite etwaiger Informationsrechte und -pflichten. Wie die geforderten Informationen nachvollziehbar kommuniziert werden können, unterscheidet sich je nach Adressatenkreis der Systeme und damit auch nach Einsatzzweck und -kontext.

Aus ethischer und rechtlicher Sicht ist für den Umgang mit algorithmischen Systemen entscheidend, dass zu jedem Zeitpunkt eine klare Zuordnung von **Verantwortung** ihrer Auswirkungen zu menschlichen Entscheidungsträgern gewährleistet ist. Hierbei kommt insbesondere auch Regelungen zur **Haftung** eine zentrale Bedeutung zu, wobei die Frage nach der angemessenen Ausgestaltung eines Haftungsregimes für bestimmte digitale Produkte, Inhalte und/ oder Dienstleistungen erneut auch mit Blick auf die Kritikalität des Systems erfolgen muss (→ unten 8).

An der **Konkretisierung und Ausgestaltung** dieser differenzierten **Regulierungsanforderungen** müssen im Sinne der von der DEK eingenommenen Governance-Perspektive **alle relevanten Akteure** – Staat, Unternehmen, Entwickler und die Bevölkerung – partizipieren. Die DEK weist darauf hin, dass auch ohne spezielle Regulierung der Einsatz algorithmischer Systeme an den allgemeinen Rechtsnormen zu messen ist. Hierzu gehört insbesondere das zivilrechtliche Haftungsrecht, das bei Handlungen, die rechtlich geschützte Interessen verletzen, grundsätzlich zum Schadensersatz verpflichtet. Auch finden die Regelungen des Gesetzes gegen den unlauteren Wettbewerb Anwendung, etwa im Falle von Irreführungen von Verbrauchern, sowie das Strafrecht, wenn mithilfe algorithmischer Systeme Straftaten begangen werden. Bei der Prüfung der Voraussetzung dieser Normen kommt der Kritikalität der Systeme und den daraus abzuleitenden Systemanforderungen auch nach allgemeinen Maßstäben rechtliche Bedeutung zu.

Algorithmische Systeme kommen zum Einsatz, um spezifische Funktionen zu erfüllen. Um die Systemkritikalität zu bewerten, ist daher auch die **ethische Bewertung dieses Zwecks** von ausschlaggebender Bedeutung. Ist der Einsatzzweck ethisch unvertretbar, etwa weil er grundlegende Rechte und Freiheiten verletzt oder gegen die freiheitlich-demokratische Grundordnung verstößt, ergeben sich „rote Linien" oder „absolute Grenzen" – für algorithmische Systeme ebenso wie für Menschen. So ist beispielsweise ein der politischen Manipulation, dem Betrug oder der kollusiven Preisabsprache dienendes algorithmisches System per se als ethisch verwerflich anzusehen.

Dabei sind Einsatzzwecke oft vielschichtig, und einzelne ihrer Facetten – insbesondere was Nebenzwecke betrifft – können ethisch jeweils unterschiedlich zu bewerten sein. Die Herausarbeitung eines für die Bewertung maßgeblichen Einsatzzwecks setzt insofern oft schwierige **Wertungsentscheidungen** voraus. Die Bewertung des Einsatzzwecks algorithmischer Systeme wird bei digitalen Produkten dadurch erschwert, dass sich die Phasen der Entwicklung und der Implementierung im Markt zunehmend überschneiden; auch kann die Zweckbestimmung eines Produkts nach seiner Implementierung im Markt durch Updates oder den Einsatz in anderen Anwendungskontexten verändert werden.

Komplexe Zweckbestimmung bei Medienintermediären

Manche Medienintermediäre, wie Suchmaschinen, sind im Zeitalter des Internets unverzichtbar, weil sie den Zugang zu Informationen im Netz ermöglichen, die Informationsflut kanalisieren und dem Einzelnen die Nutzung des Internet faktisch überhaupt erst ermöglichen. Insoweit sind ihre Zwecke wünschenswert und ethisch unkritisch. Medienintermediäre können aber in ihrer konkreten Ausgestaltung ethisch problematisch sein. Ihre Systeme stellen für Nutzer eine personalisierte Auswahl an Informationen bereit. Dies führt zu einer Auswahlentscheidung über die angezeigten Inhalte. Da damit aber die überwiegende Mehrzahl der Inhalte nicht oder nur nachrangig angezeigt wird, verengt sich das Wahrnehmungsspektrum des Einzelnen. In der Konsequenz entscheidet der Intermediär im Wege der Programmierung über den Kopf des Nutzers hinweg darüber, was dieser wahrnimmt. Soweit die Geschäftsmodelle der Medienintermediäre werbegetrieben sind, wie dies etwa in großen sozialen Netzwerken der Fall ist, besteht das Risiko, dass Betreiber ein wirtschaftliches Interesse daran haben, auch ethisch fragwürdige oder gar extremistische Inhalte zu verbreiten, weil diese eine höhere Verweildauer der Nutzer auf der Plattform versprechen, wodurch Werbeeinnahmen steigen. Es besteht durch das Zusammenspiel von Sortierung und Verengung des Wahrnehmbaren und der zusätzlichen Gefahr der Einflussnahme auf den Nutzer durch intransparente Drittinteressen die Möglichkeit der intransparenten Einflussnahme etwa auf die politische Willensbildung bis hin zu einer politischen Manipulation. Das ist eine erhebliche Gefahr für die freie Meinungsbildung als Grundlage der Demokratie.

3.2 Kritikalitätspyramide

Die DEK empfiehlt, den Kritikalitätsgrad algorithmischer Systeme einheitlich anhand eines **übergreifenden Modells** zu bestimmen. Der Kritikalitätsgrad soll Gesetzgeber und Gesellschaft bei der Suche nach geeigneten Regulierungsschwellen und -instrumenten anleiten, kann aber auch Entwicklern und Betreibern bei der Selbsteinschätzung ihrer Produkte und Systeme Orientierung bieten und schließlich in Aus-, Fort- und Weiterbildung für die **Sensibilisierung und Schulung** unterschiedlicher Akteure eingesetzt werden. Die DEK unterscheidet insoweit mit Blick auf das Schädigungspotenzial algorithmischer Systeme – für private wie für staatliche Betreiber – **fünf Kritikalitäts-Stufen**:

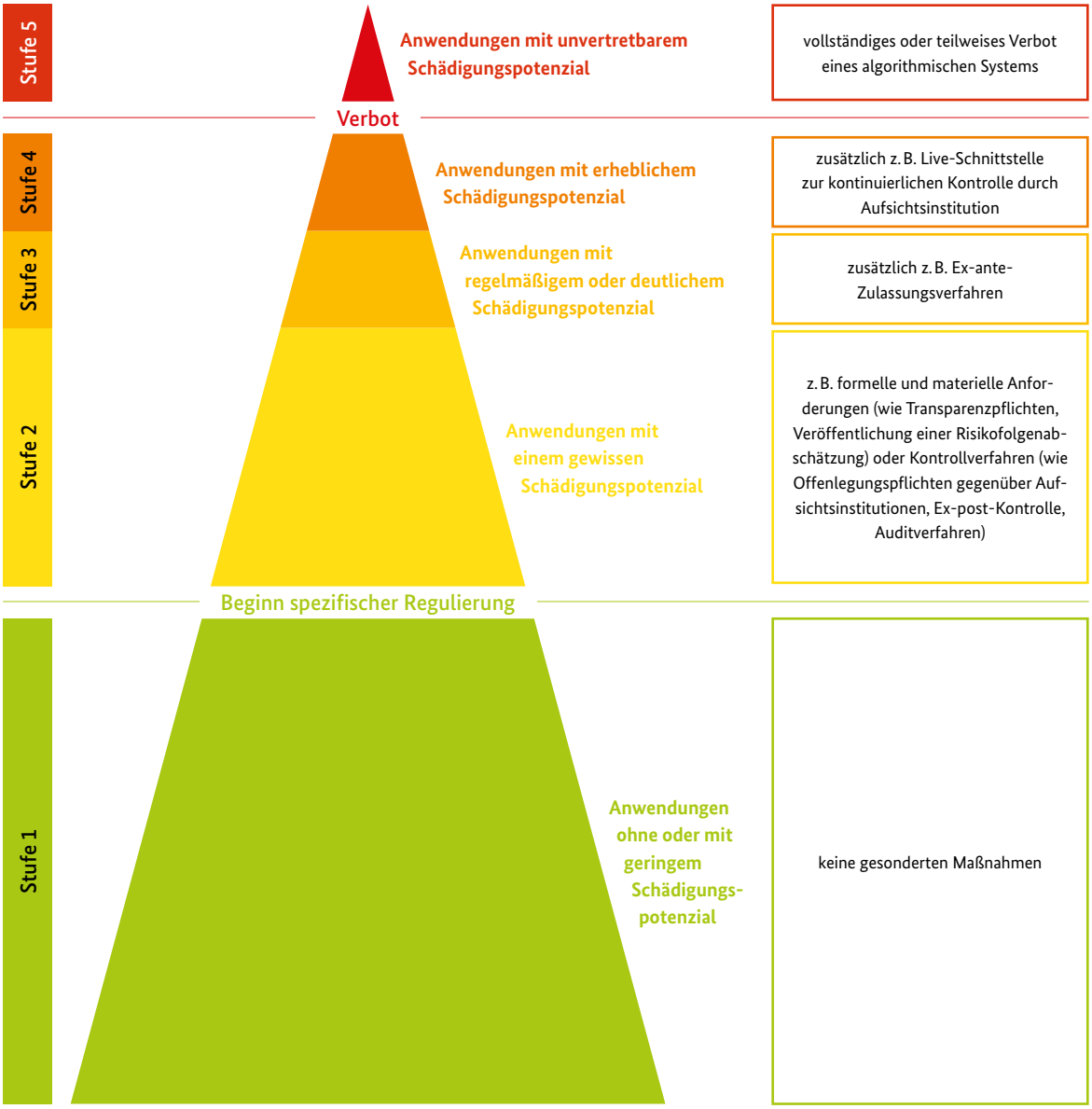


Abbildung 8:
Kritikalitätspyramide und risikoadaptiertes Regulierungssystem für den Einsatz algorithmischer Systeme

In unproblematischen Anwendungskontexten wird es in der Regel nicht erforderlich sein, von den Entwicklern, Auftraggebern oder Betreibern zu verlangen, bestimmte Verfahren ethisch-rechtlicher Kontrolle zu durchlaufen. So sieht auch die DEK in der Vielzahl von **Anwendungen ohne oder mit nur sehr geringem Schädigungspotenzial** – also auf der untersten Stufe („Stufe 1") der Kritikalitätspyramide – keine Notwendigkeit einer besonderen Kontrolle, welche über die allgemeinen Qualitätsanforderungen, die auch für Produkte ohne algorithmische Elemente gelten, hinausginge.

Beispiel 13
Die in einem Getränkeautomaten zum Einsatz gelangenden Algorithmen haben zwar auch ein gewisses Schädigungspotenzial, weil ein Nutzer z. B. keine Ware erhalten und sein Geld verlieren könnte. Dieses Schädigungspotenzial überschreitet aber nicht die Schwelle zu einem besonderen Schädigungspotenzial im Algorithmenkontext. Es ist ausreichend, hier auf die allgemeinen Mechanismen zu vertrauen, welche Vertragspartner zur Erbringung ihrer vertraglich geschuldeten Leistung oder Hersteller zur Produktion funktionierender Geräte verpflichten.

Bei **Anwendungen mit einem gewissen Schädigungspotenzial** – also auf Stufe 2 in der Kritikalitätspyramide – kann und soll Regulierung einsetzen. Allerdings sind die hier erforderlichen Maßnahmen in ihrer Reichweite beschränkt. Mit Blick auf die niedrige Kritikalität gilt es hier besonders, eine übermäßige Belastung der Hersteller und Betreiber zu vermeiden, um technologische und soziale Innovationen sowie die Marktentwicklung nicht übermäßig zu behindern. Maßnahmen, die sich auf Stufe 2 anbieten können, umfassen etwa Ex-post-Kontrollen (beispielsweise in Form einer Input-Output-Kontrolle), insb. wenn ein begründeter Verdacht auf Fehlverhalten der Systeme besteht. Darüber hinaus sollte die Pflicht zur Erstellung und Veröffentlichung einer angemessenen Risikofolgenabschätzung bestehen (→ unten). Sektorspezifisch können ferner Offenlegungspflichten gegenüber Aufsichtsinstitutionen (einschließlich Einrichtung einer Schnittstelle zur Durchführung von Input-Output-Kontrollen durch eine Aufsichtsinstitution), gesteigerte Transparenzpflichten sowie Auskunftsrechte für Betroffene (→ dazu im Einzelnen) sinnvoll sein. Zu denken ist auch an Codes of Conduct, welche branchenspezifisch erarbeitet und dann von den zuständigen Aufsichtsbehörden genehmigt werden. Die Einhaltung wäre dann durch Stichproben sowie anlassbezogen durch die Aufsichtsbehörden zu prüfen (→ unten).

Kritikalität bei Smart Mobility-Anwendungen

Ein Anbieter von Smart Mobility-Anwendungen greift auf einen über alle Fahrzeug- und Mobilitätsdaten generierten Datenpool zu. Sofern diese Daten ausschließlich zur Stauvorhersage genutzt werden, ist die Kritikalität als gering einzustufen. Durch den Einsatz von Smart Mobility ist aber auch der Verkehrsfluss steuerbar. Können Algorithmen etwa anhand der aus Fahrzeugdaten in Echtzeit ermittelten Gesamtauslastung des Mobilitätssystems aus Straße, Schiene, Wasser und Luft erkennen, welche Wegführung für eine Fortbewegung von A nach B optimal ist, so kann dem Nutzer ein entsprechender Weg nach seinen Vorlieben (z. B. schnellste/umweltfreundlichste/günstigste etc. Route) vorgeschlagen werden. Es stellt sich aber auch die Frage, ob der Staat bestimmte Routen unter Berücksichtigung staatlich vorgegebener Kriterien für den Nutzer festlegen kann. Hier läge angesichts des veränderten Schädigungspotentials die Kritikalität höher und würde daher einer strengeren, kritikalitätsangemessenen Regulierung bedürfen.

Beispiel 14
Dynamische Preissetzung (etwa nach den Kriterien von Angebot und Nachfrage) im Online-Handel, die aber keine Personalisierung von Preisen beinhaltet, hat ein meist geringes, aber doch die Relevanzschwelle überschreitendes Schädigungspotenzial, etwa betreffend einer versteckten Diskriminierung.

Beispiel 15
Preisalgorithmen zur Festsetzung personalisierter Preise (d.h. Festsetzung des Preises nach auf den einzelnen Kunden zugeschnittenen, i.d.R. die maximale individuelle Zahlungsbereitschaft abschätzenden Kriterien) bringen ein deutliches Schädigungspotenzial mit sich, beispielsweise betreffend die Diskriminierung besonders vulnerabler Gruppen. Sie sollten allenfalls nach Durchlaufen eines Zulassungsverfahrens zum Einsatz gelangen können.

Bei **Anwendungen mit regelmäßigem oder deutlichem Schädigungspotenzial** auf Stufe 3 der Kritikalitätspyramide, kann in spezifischen Fällen zusätzlich zu den bereits bei Stufe 2 zu fordernden Mechanismen eine Ex-ante-Kontrolle in der Form eines Zulassungsverfahrens gerechtfertigt sein (→ unten). Aufgrund der hohen Dynamik mancher algorithmischer Systeme ist bei erteilter Zulassung eine regelmäßige Überprüfung erforderlich.

Das Gleiche, was für Stufen 2 und 3 gilt, hat auch für **Anwendungen mit erheblichem Schädigungspotenzial** auf Stufe 4 zu gelten. Allerdings sind hier zusätzliche Kontroll- und Transparenzpflichten bis hin zu einer weitergehenden Veröffentlichung der in die algorithmische Berechnung einfließenden Faktoren und deren Gewichtung, der Datengrundlage sowie des algorithmischen Entscheidungsmodells in nachvollziehbarer Form zu fordern oder auch die kontinuierliche Kontrolle durch eine Live-Schnittstelle vorzusehen. Auch weitergehende Schutzmaßnahmen zur Schadensvermeidung sind erforderlich.

Differenzierte Kritikalität bei Medienintermediären

Medienintermediäre verarbeiten und vermitteln mithilfe ihrer algorithmischen Filtersysteme sowohl meinungsrelevante Inhalte, die für die demokratische Willensbildung relevant sind, als auch Inhalte, die der Werbung, Kaufempfehlung oder Unterhaltung dienen. Sie stehen geradezu paradigmatisch für Konstellationen, in denen der Einsatz desselben algorithmischen Systems unterschiedliche Gefährdungspotenziale hat. Wenn es um Nutzerinteraktion im Konsumgüterbereich (insbes. Werbung oder Kaufempfehlungen) geht, besteht – in Abhängigkeit von dem verwendeten Personalisierungsmodell – ein geringes bis hohes Gefährdungspotenzial. Sobald aus übergeordneten Interessen zur Erhaltung der freiheitlichen Ordnung ausgewogene Vielfalt erzeugt werden muss (insbesondere bei meinungsrelevanten Themen), ist das Gefährdungspotenzial bereits durch den Inhalt von vorne herein höher. Damit verändern sich zugleich die Regulierungsanforderungen. Bei Konsum- und Unterhaltungsangeboten muss, je nach verwendeten Personalisierungskriterien, Anwendungskontexten oder zu erwartenden Wohlfahrtseffekten, eine mehr oder weniger strenge Regulierung erfolgen.

Beispiel 16
Auf Stufe 4 wären etwa algorithmische Systeme von Akteuren mit massiver Marktmacht einzustufen, die der Ermittlung der Kreditwürdigkeit eines individuellen Verbrauchers oder Unternehmers dienen. Ob eine Person einen Kredit erhält oder nicht, kann für ein individuelles Schicksal entscheidend sein. Die hohe Systemkritikalität wird auch begründet durch die Marktkonzentration auf wenige Anbieter und die Tendenz, dass sich ein Kreditgeber auf das Urteil eines bestimmten Akteurs verlässt.

Mit Blick auf die Kriterien für die Systemkritikalität kann für **Anwendungen mit unvertretbarem Schädigungspotenzial** (Stufe 5) schließlich ein vollständiges oder teilweises ex-ante-**Verbot** des Einsatzes eines algorithmischen Systems infrage kommen. Zudem kann ein Verbot ex post als Sanktion für Verstöße gegen geltendes Recht oder die Nichteinhaltung der für die konkrete Systemkritikalität erforderlichen Systemanforderungen folgen.

Beispiel 17
Autonome Waffensysteme (Lethal Autonomous Weapons) werden vielfach als „rote Linie" angesehen, weil die Tötung von Menschen nicht Maschinen überlassen werden dürfe. Das kann allerdings wohl nur gelten, soweit man von algorithmendeterminierten Tötungen ausgeht. Soweit autonome Waffensysteme menschliche Soldaten lediglich bei der Objekterkennung unterstützen oder sofern sie lediglich dazu dienen, einen Flugkörper trotz Seitenwinds in der Bahn zu halten, ist eine ethische „rote Linie" nicht überschritten.

Die Einordnung eines algorithmischen Systems in die Kritikalitätspyramide muss – unter Berücksichtigung der dynamischen Natur dieser Systeme – gegebenenfalls **regelmäßig überprüft** werden.

3.3 Regulierung algorithmischer Systeme durch horizontale Vorgaben im Recht der Europäischen Union und sektorale Konkretisierung

Algorithmische Systeme erfassen immer mehr Bereiche unseres individuellen und gesellschaftlichen Lebens. Die Zwecke algorithmischer Systeme und die möglichen Einsatzfelder sind dabei nicht fest definiert. So kann ein für die Gesichtserkennung bei Privatfotos entwickeltes System auch von staatlichen Ermittlungsbehörden für Zwecke der Strafverfolgung oder Gefahrenabwehr genutzt werden. Das legt nahe, den Herausforderungen algorithmischer Systeme nach dem Vorbild des Datenschutzrechts in Form **horizontaler Regulierung** zu begegnen, d. h. durch einen Rechtsakt, dessen sachlicher Anwendungsbereich allgemein algorithmische Systeme erfasst und der in personeller Hinsicht für **private und öffentliche Akteure** gleichermaßen gilt. Neben der hohen Symbolkraft spräche für eine horizontale Regulierung auch die Tatsache, dass Schutzlücken ausgeschlossen wären und gegenwärtig noch gar nicht absehbare Gefährdungskonstellationen erfasst wären. Eines der wichtigsten Argumente für eine derart übergreifende Regelung, die Grundprinzipien für alle algorithmischen Systeme festlegt, ist zudem, dass die Bürger so in allen Bereichen Erwartungsklarheit erhalten und der (europäische) Gesetzgeber diese Aufgabe in einem überschaubaren Zeitraum leisten kann.

Vor diesem Hintergrund **empfiehlt die DEK** der Bundesregierung, auf europäischer Ebene auf die Erarbeitung einer horizonalen Grundregelung im Form einer **EU-Verordnung für Algorithmische Systeme (EUVAS)** hinzuwirken. Der horizontale Rechtsakt sollte neben den zentralen Grundprinzipien für algorithmische Systeme, wie sie hier als Anforderungen an algorithmische Systeme entwickelt wurden, allgemeine materielle Regelungen zur Zulässigkeit und Gestaltung algorithmischer Systeme im Sinne der Systemkritikalität, zur Transparenz, zu Betroffenenrechten, zu organisatorischen und technischen Absicherungen und zu den Institutionen und Strukturen der Aufsicht bündeln.

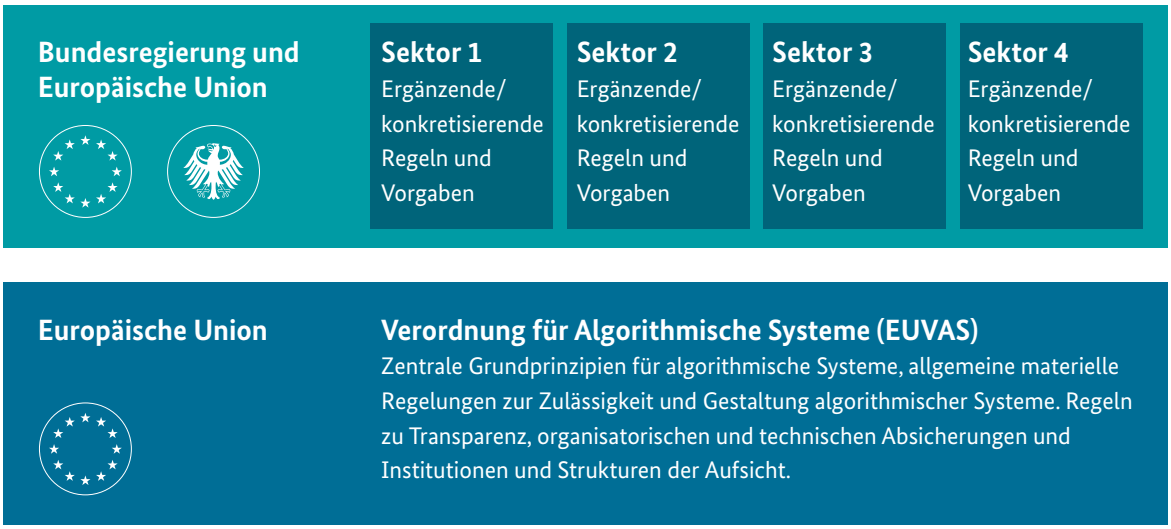


Abbildung 9:
Regulierung algorithmischer Systeme durch horizontale Vorgaben im Recht der Europäischen Union und sektorale Konkretisierung

Zugleich empfiehlt die DEK der Bundesregierung, sich auf europäischer Ebene auch für **sektorale Regeln** einzusetzen und außerhalb der Kompetenzen der EU selbst im Rahmen der ihr zustehenden Gesetzgebungs- und Verwaltungskompetenzen entsprechende sektorale Rechtsakte zu erlassen, die am Gedanken der Systemkritikalität orientiert sind. (Abb. 9).

Eine **übergreifende EUVAS** wird sich auf wenige **Grundprinzipien** beschränken müssen, da anderenfalls der europäische Gesetzgeber überfordert würde. Bei einer zu detaillierten Regelung wäre er insbesondere mit der Frage konfrontiert, wie der kaum mehr überschaubaren Vielzahl der Systeme und der hochdynamischen Entwicklung der Technologie in einem allgemeinen Rechtsakt gerecht zu werden ist. Aus Sicht der Betroffenen tragen allgemeine Rechtsakte zudem das Risiko in sich, dass das administrative Pflichtenprogramm auch in Fällen Anwendung findet, in denen an sich kein hinreichendes Schädigungspotenzial besteht, weil die Differenzierungen zwischen riskanten und weniger riskanten Einsatzzielen – ebenso wie mögliche Ausnahmekonstellationen – in einem horizontalen Rechtsakt nicht derart feingranular vorgenommen werden können, wie sie sich in der Wirklichkeit darstellen.

Mit Blick auf beide Punkte wirkt der **ergänzende** Rückgriff auf in ihrem Anwendungsbereich beschränkte, dadurch aber leichter zu präzisierende **sektorale Normen** entlastend. Ein ergänzender, sektoral differenzierender Zugriff muss zudem die nach geltendem Recht zwischen EU, Bund und Ländern verteilten Gesetzgebungs- und Verwaltungszuständigkeiten berücksichtigen. Hinzu kommt, dass gerade mit Blick auf die Institutionen und Strukturen der behördlichen Kontrolle und Aufsicht aus unterschiedlichen Gründen die Zusammenführung der „Gesamtaufgabe“ in einer Behörde nicht in Frage kommt (→ unten).

Neben der EUVAS ist daher der Erlass mehrerer **Rechtsakte mit spezifischen Vorgaben für einzelne Sektoren oder Gefährdungskonstellationen** erforderlich. Die Kombination einer allgemeinen Grundlagen-Regulierung mit weiteren sektorspezifisch-konkretisierenden Rechtsakten hat nach Auffassung der DEK den großen Vorteil, dass er den zwischen einzelnen Systemen und Einsatzkontexten differierenden Schutzbedarf differenziert abbilden kann. Dies entspricht dem Grundgedanken risikoadaptierter Regulierung, wonach die regulatorischen Anforderungen an algorithmische Systeme in Abhängigkeit von der spezifischen Systemkritikalität festzulegen sind. Auch im Datenschutzrecht gibt es im öffentlichen Bereich zahlreiche Spezialgesetze, die die allgemeinen Vorgaben der DSGVO sektoral ergänzen. Zwar ist es der Grundgedanke des Datenschutzrechts, dass es unter den Bedingungen der automatisierten Datenverarbeitung kein „belangloses“ Datum mehr gibt, weshalb Differenzierungen bei personenbezogenen Daten nach dem Grad der Schutzwürdigkeit oder Kritikalität kaum mehr unter Verzicht auf gemeinsame Grundregeln sinnvoll möglich sind. Es ist aber auch richtig, dass ein erhöhtes Schutzniveau in den verschiedensten Bereichen staatlichen Tätigwerdens durch eine Vielzahl von Spezialregelungen abgesichert wird. Einen ähnlichen Bedarf nach ergänzenden sektoralen Vorgaben gibt es auch für algorithmische Systeme. Deren Anwendung muss auch nicht daran scheitern, dass ihr Zweck und ihr Einsatzkontext wechseln können. Denn zum einen stößt eine solche Änderung gerade bei komplexeren Systemen an Grenzen. Zum anderen lässt sich ihnen regulatorisch dadurch begegnen, dass die Rechtsakte sachlich nicht an die ursprüngliche Zwecksetzung bzw. an den ursprünglichen Einsatzkontext, sondern an die **aktuelle Funktionalität des Systems** bzw. den **beabsichtigten neuen Einsatzzweck** des Systems anknüpfen. Zweck- und Kontextänderungen führen auf diese Weise ggf. zur Anwendung eines neuen regulatorischen Rahmens.

Von diesen primär pragmatischen Erwägungen unberührt ist allerdings die Forderung an den bzw. die Normgeber, bei ihren jeweiligen Vorhaben so weitgehend wie möglich auf **rechtsaktsübergreifende Kohärenz** zu achten. Dies gilt nicht nur für die hier entwickelten Regelungsansätze, d. h. insbesondere den Gedanken der Systemkritikalität, und die Betroffenenrechte. Auch die regulatorischen Infrastrukturen und Prozesse sollten so weit wie möglich einheitlich ausgestaltet sein.

Zusammenfassung der wichtigsten Handlungsempfehlungen

Empfehlung eines risikoadaptierten Regulierungsansatzes

36

Die DEK empfiehlt einen **risikoadaptierten Regulierungsansatz** für algorithmische Systeme. Er sollte auf dem Grundsatz aufbauen, dass ein steigendes Schädigungspotenzial mit wachsenden Anforderungen und Eingriffstiefen der regulatorischen Instrumente einhergeht. Für die Beurteilung kommt es jeweils auf das **gesamte sozio-technische System** an, also alle Komponenten einer algorithmischen Anwendung einschließlich aller menschlichen Akteure, von der Entwicklungsphase (z. B. hinsichtlich der verwendeten Trainingsdaten) bis hin zur Implementierung in einer Anwendungsumgebung und zur Phase von Bewertung und Korrektur.

37

Die DEK empfiehlt, die Bestimmung des Schädigungspotenzials algorithmischer Systeme für Einzelne und/oder die Gesellschaft anhand eines **übergreifenden Modells** einheitlich vorzunehmen. Dafür sollte der Gesetzgeber mit Hilfe von **Kriterien** ein Prüfschema definieren, nach welchem die Kritikalität algorithmischer Systeme auf der Grundlage der von der DEK vorgestellten allgemeinen ethischen und rechtlichen Grundsätze und Prinzipien zu bestimmen ist.

38

Regulatorische Instrumente und **Anforderungen** an algorithmische Systeme sollten u. a. Korrektur- und Kontrollinstrumente, Vorgaben für die Transparenz, die Erklärbarkeit und die Nachvollziehbarkeit der Ergebnisse sowie Regelungen zur Zuordnung von Verantwortlichkeit und Haftung für den Einsatz umfassen.

39

Die DEK erachtet es als sinnvoll, mit Blick auf das Schädigungspotenzial algorithmischer Systeme in einem ersten Schritt **fünf Kritikalitäts-Stufen** zu unterscheiden. Auf der untersten Stufe (Stufe 1) von Anwendungen ohne oder mit geringem Schädigungspotenzial besteht keine Notwendigkeit einer besonderen Kontrolle oder von Anforderungen, die über die allgemeinen Qualitätsanforderungen, welche auch für Produkte ohne algorithmische Elemente gelten, hinausgehen.

40

Bei Anwendungen mit einem **gewissen Schädigungspotenzial** (Stufe 2) kann und soll bedarfsgerechte Regulierung einsetzen, wie etwa Ex-post-Kontrollen, die Pflicht zur Erstellung und Veröffentlichung einer angemessenen Risikofolgenabschätzung, Offenlegungspflichten gegenüber Aufsichtsinstitutionen oder auch gesteigerte Transparenzpflichten sowie Auskunftsrechte für Betroffene.

41

Bei Anwendungen mit **regelmäßigem** oder **deutlichem Schädigungspotenzial** (Stufe 3) können zusätzlich Zulassungsverfahren gerechtfertigt sein. Bei Anwendungen mit **erheblichem Schädigungspotenzial** (Stufe 4) fordert die DEK darüber hinaus verschärfte Kontroll- und Transparenzpflichten bis hin zu einer Veröffentlichung der in die algorithmische Berechnung einfließenden Faktoren und deren Gewichtung, der Datengrundlage und des algorithmischen Entscheidungsmodells sowie die Möglichkeit einer kontinuierlichen behördlichen Kontrolle über eine Live-Schnittstelle zum System.

42

Bei **Anwendungen mit unvertretbarem Schädigungspotenzial** (Stufe 5) ist schließlich ein vollständiges oder teilweises **Verbot** auszusprechen.

43

Zur Umsetzung der durch die DEK vorgeschlagenen Maßnahmen empfiehlt die DEK eine Regulierung algorithmischer Systeme durch allgemeine **horizontale Vorgaben im Recht** der Europäischen Union **(Verordnung für Algorithmische Systeme, EUVAS)**. Dieser horizontale Rechtsakt sollte die zentralen Grundprinzipien für algorithmische Systeme enthalten, wie sie die DEK als Anforderungen an algorithmische Systeme entwickelt hat. Insbesondere sollte er im Lichte der Systemkritikalität allgemeine materielle Regelungen zur Zulässigkeit und Gestaltung algorithmischer Systeme, zur Transparenz, zu Betroffenenrechten, zu organisatorischen und technischen Absicherungen und zu den Institutionen und Strukturen der Aufsicht bündeln. Der horizontale Rechtsakt sollte auf der Ebene der EU und der Mitgliedstaaten eine **sektorale Konkretisierung erfahren**, die wiederum am Gedanken der Systemkritikalität orientiert ist.

44

Im Zuge der hier empfohlenen Entwicklung einer EUVAS sollte die Aufgabenverteilung zwischen dieser Regulierung und der **DSGVO** überdacht werden. Dabei ist zum einen zu berücksichtigen, dass sich spezifische Risiken algorithmischer Systeme für den Einzelnen und für Gruppen auch dann manifestieren können, wenn keine personenbezogenen Daten verarbeitet werden, und dass die Risiken nicht unbedingt solche des Datenschutzes sind, wenn sie etwa das Vermögen, Eigentum, körperliche Integrität oder Diskriminierung betreffen. Zum anderen ist zu bedenken, dass für eine künftige horizontale Regulierung algorithmischer Systeme ein flexibleres, stärker risikoadaptiertes Regulierungsregime als für den Datenschutz in Betracht gezogen werden sollte.

4. Instrumente: Pflichten des Verantwortlichen und Rechte Betroffener

Um dem Einzelnen, aber auch Gruppen, wirksamen Schutz gegen die Gefahren algorithmischer Systeme angedeihen zu lassen, hält die DEK sowohl Transparenzanforderungen (→ s. im Folgenden 4.1) als auch weitere Vorgaben für algorithmische Systeme im Sinne eines wirksamen Schutzes gegen inhaltlich unangemessene oder unfaire Entscheidungen (→ 4.2.) für geboten.

4.1 Transparenzanforderungen

4.1.1 Kennzeichnungspflichten („Ob“)

Ein zentrales Instrument, um Transparenz herzustellen, ist eine **Kennzeichnungspflicht.** Da der Grad der Informationsdichte einer Kennzeichnungspflicht gering ist, sind auch die Eingriffe in die Grundrechte der Betreiber, insbesondere ihrer Geschäftsgeheimnisse, weniger schwer als bei Auskunftsrechten. Dies rechtfertigt es nach Ansicht der DEK, eine Kennzeichnung bei kritischen Systemen (ab Stufe 2) als flächendeckende Pflicht für die Betreiber, und nicht als antragsabhängiges Recht einzelner Betroffener auszugestalten.

Die DEK hält die bestehenden Kennzeichnungspflichten der DSGVO[3] aufgrund des verhältnismäßig engen Anwendungsbereichs des Art. 22 DSGVO (mit seiner Anknüpfung an eine ausschließlich auf einer automatisierten Verarbeitung […] beruhenden Entscheidung), auf den die Informationspflichten rekurrieren, für **nicht ausreichend**. Auch unterhalb der Schwelle des Art. 22 DSGVO können sich nämlich signifikante Auswirkungen für Betroffene einstellen. Das gilt für algorithmenbasierte und algorithmengetriebene Entscheidungen, also Konstellationen, in denen menschliche Entscheidungen Gefahr laufen, algorithmische Informationen und Entscheidungsvorschläge (insbesondere in Bereichen, in denen ein menschliches Abwägen erwartet wird) unreflektiert und standardmäßig zu übernehmen, oder sich nur in algorithmisch ermittelten und vorgegebenen Bahnen zu bewegen.

Da die DEK die Authentizität zwischenmenschlicher Kommunikation als Grundbedingung für einen vertrauensvollen Umgang miteinander in der Gesellschaft ansieht, sollte eine Kennzeichnungspflicht im Falle einer **Verwechselungsgefahr** zwischen Mensch und Maschine immer und somit unabhängig von der Systemkritikalität gelten. Dies gilt etwa für digitale Sprachassistenten und Chatbots, die bisweilen kaum mehr als solche zu erkennen sind. Die Kennzeichnung bei Sprachassistenten kann beispielsweise sowohl durch die regelmäßige Offenlegung der maschinellen Natur (auch während einer laufenden Kommunikation) als auch durch die Verwendung einer maschinell klingenden Stimme erfolgen. Keine Verwechslungsgefahr (und daher auch kein Erfordernis nach einer Kennzeichnungspflicht) besteht nach Ansicht der DEK hingegen in Bereichen, in denen die Natur der Information irrelevant ist oder der Rezipient ohnehin eine maschinelle Stimme erwartet, wie beispielsweise bei Lautsprecheransagen an einem Bahnhof.

4.1.2 Informationspflichten, Erklärungspflicht und Informationszugang („Wie“ und „Was“)

Während Kennzeichnungspflichten den Betreibern Transparenz darüber abverlangen, wann und in welchem Umfang („ob“) algorithmische Systeme überhaupt zum Einsatz kommen, richten sich Informationspflichten und **Auskunftsrechte** regelmäßig auf vertiefte Informationen zum Entscheidungsmechanismus („wie“) und den zugrundeliegenden Daten („was“) des algorithmischen Systems.

3 Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g und Art. 15 Abs. 1 lit. h i.V.m. Art. 22 DSGVO.

Informationspflichten und Auskunftsrechte über das Verhalten algorithmischer Systeme und den Weg der systeminternen Entscheidungsfindung sind aus der Sicht der Bürger wichtig, um Entscheidungen nachvollziehen und individuell überprüfen bzw. überprüfen lassen zu können. Erst mit ihrer Hilfe können betroffene Personen ihre Rechte wahrnehmen und eine Entscheidung fundiert angreifen. Die folgenden Transparenzanforderungen gelten für private und hoheitliche Betreiber algorithmischer Systeme gleichermaßen. Auf besondere Anforderungen, die an die Tranzparenz hoheitlich genutzter Systeme zu stellen sind, wird unten unter 7. näher eingegangen.

4.1.2.1 Informationspflichten und Auskunftsrechte

Dort, wo personenbezogene Daten verarbeitet werden, sehen Art. 13, 14 und 15 DSGVO bereits Informationspflichten und Auskunftsrechte vor. Im Falle einer automatisierten Entscheidung im Sinne des Art. 22 DSGVO verleiht die DSGVO den betroffenen Personen einen Anspruch auf „aussagekräftige" Informationen über die „involvierte Logik", die „Tragweite" und die „angestrebte Auswirkung" der Verarbeitung.[4]

Nach Auffassung der DEK sollte der Rechtsgedanke dieser Normen – ebenso wie im Falle der Kennzeichnungspflichten (→ oben) – auch außerhalb des engen Anwendungsbereichs des Art. 22 Abs. 1 DSGVO Anwendung finden und fester Bestandteil der hier vorgeschlagenen EUVAS (→ oben) werden. Dabei hängt es von der **Kritikalität des Systems** ab, welchen Umfang eine derartige Informationspflicht hat. Bei Anwendungen mit einem geringen Schädigungspotential werden kurze Stellungnahmen zur Entscheidungslogik genügen, etwa zur verwendeten Datengrundlage oder allgemeinen Gewichtung bestimmter Faktoren mit Blick auf das Ergebnis. Je risikoträchtiger das System ist, desto weiter reichen grundsätzlich die Offenlegungspflichten.

Je persönlichkeitssensibler die Entscheidung ist, desto eher ist eine auf den Einzelfall bezogene Detailauskunft angezeigt. Es ist dabei allerdings auch zu bedenken, dass die Erteilung detaillierter Informationen über die Faktoren und ihre Gewichtung auch ethisch möglicherweise bedenkliche Steuerungseffekte für die private Lebensführung des Betroffenen mit sich bringen können. Darüber hinaus könnten die erlangten Informationen vom Betroffenen dazu genutzt werden, ein algorithmisches System, das eine wichtige Aufgabe erfüllt, zu unterlaufen.

Die **technischen und organisatorischen Anforderungen**, die zu erfüllen sind, um diesen weitgehenden Informationspflichten nachkommen zu können, müssen von Anfang an in die Konzeption von algorithmischen Systemen einfließen. Denn deren rechtmäßiger Betrieb lässt sich nur sicherstellen, wenn die entsprechend notwendigen „aussagekräftigen" Informationen beim Einsatz des Systems auch erteilt werden können.

Bei der Ausgestaltung von Informationspflichten und Auskunftsrechten, um die Transparenz algorithmischer Systeme zu stärken, ist zu beachten, dass bei Verbrauchern keine speziellen technischen Fähigkeiten und Kenntnisse vorausgesetzt werden dürfen. Daher gilt es bei jeder Ausweitung der Auskunftsrechte zu bedenken, dass dies aus Sicht der Betroffenen nur dann die Transparenz steigert, wenn die Informationen **adressatengerecht** aufbereitet sind.

4 Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g und Art. 15 Abs. 1 lit. h DSGVO.

4.1.2.2 Erklärungspflichten

Jedenfalls in bestimmten Bereichen komplexer algorithmischer Systeme kann es sachgerecht sein, dem System zusätzlich zur allgemeinen Erläuterung der Logik und Tragweite des Systems eine Erläuterung der konkreten Gründe für das Zustandekommen einer Empfehlung oder Entscheidung des Systems abzuverlangen. Einer derartigen individuellen Erklärung bedarf es vor allem dann, wenn die Entscheidung persönlichkeitssensible Bereiche betrifft oder sonst eine besondere grundrechtliche oder sozioökonomische Tragweite hat. Wesentlich ist dabei, dass betroffene Personen verständlich, relevant und konkret informiert werden. Die DEK begrüßt daher die technischen Bemühungen, die Erklärbarkeit algorithmischer (insbesondere selbstlernender) Systeme (explainable oder explicable AI) zu stärken, und fordert die Bundesregierung auf, derartige Projekte zu fördern.

In bestimmten Situationen ist nach Auffassung der DEK ein Anspruch auf „**kontrafaktische Erklärungen**“ (counterfactual explanations) erwägenswert, wie er teilweise in der Literatur diskutiert wird.[5] Danach werden betroffenen Personen jene Faktoren der Entscheidungsfindung mitgeteilt, die mit Blick auf eine für sie negative Entscheidung den positiven Unterschied gemacht hätten, also zum eigentlich gewünschten Ergebnis geführt hätten. Im Falle der auf der Nutzung eines algorithmischen Systems basierenden Ablehnung eines Antrags auf Kreditgewährung hätte der Betroffene etwa einen Anspruch darauf, vom Betreiber zu erfahren, welche der vom System berücksichtigten Faktoren wie hätten anders sein müssen, damit der Antrag positiv ausgefallen wäre. Die DEK weist allerdings darauf hin, dass dieser Ansatz gegenüber komplexeren Systemen rasch an seine Grenzen gelangt, müssten doch dem Betroffenen hier sehr viele verschiedene „kontrafaktische“ Szenarien mitgeteilt werden, um ihm ein einigermaßen vollständiges Bild zu vermitteln; anderenfalls droht eher Desinformation, bedenkliche Steuerung oder gar Manipulation, indem aus strategischen oder erzieherischen Gründen bestimmte Aspekte in den Vordergrund gerückt werden.

Als allgemeiner Baustein einer Regulierung algorithmischer Systeme eignet sich das Konzept der „kontrafaktischen Erklärung“ daher nach Auffassung der DEK bei dem aktuellen Stand der technischen Entwicklung nicht; für spezielle Verarbeitungssituationen wäre ein Einsatz jedoch denkbar.

4.1.2.3 Informationszugang für nicht unmittelbar betroffene Personen

Zusätzlich empfiehlt es sich nach Auffassung der DEK, in bestimmten Sektoren, in denen nicht nur individuelle, sondern in besonderem Maße auch gesellschaftliche Interessen betroffen sind, auch nicht unmittelbar betroffenen Personen ein Recht auf Zugang zu den Informationen über die algorithmischen Systeme einzuräumen. Das gilt insbesondere, wenn ihr Einsatz **Relevanz für die öffentliche Meinungsbildung** entfaltet oder große **Wohlfahrtseffekte** für die Bevölkerung zur Folge hat. Entsprechende Rechte werden in erster Linie für journalistische und Forschungszwecke in Frage kommen und sind zudem mit Blick auf die betroffenen Interessen der Betreiber durch hinreichende Schutzmaßnahmen zu flankieren.

Unter Umständen, insbesondere beim staatlichen Einsatz von Systemen mit einem erheblichen Schädigungspotential, sind nach Ansicht der DEK darüber hinaus auch **voraussetzungslose Informationszugangsansprüche** und **Veröffentlichungspflichten** vorstellbar.

5 Sandra Wachter / Brent Mittelstadtr / Chris Russel: Harvard Journal of Law & Technology, 2018 (31:2), S. 841, 841 ff.

4.1.2.4 Anforderungen an die Ausgestaltung, insbesondere Abwägung mit den Rechten der Betreiber

Bei der Ausgestaltung der Informations- und Erklärungspflichten und Auskunftsrechte ist stets zu berücksichtigen, dass diese auch die **rechtlich geschützten Interessen der Betreiber** algorithmischer Systeme sowie derjenigen, die deren Ergebnisse einsetzen, beeinträchtigen können. Dazu gehört allen voran der Schutz von Geschäftsgeheimnissen sowie das Interesse, Manipulationen an den Systemen und beim Gebrauch der Systeme zu verhindern. Private Betreiber können sich zwar im Grundsatz darauf berufen, dass sie über Ergebnisse eines algorithmischen Systems ihre eigene freie Willens- und Vertragsentscheidung definieren. Das entbindet sie jedoch nicht von der Kontrolle, ob ihr Handeln rechtskonform ist, denn das Grundrecht auf allgemeine Handlungsfreiheit findet in den Diskriminierungsverboten (insbesondere das AGG), den Grundrechten der betroffenen Personen oder Dritter und allgemein den – auch vertragsspezifischen – Vorgaben der Rechtsordnung eine Grenze. Zudem sind Transparenzrechte stets mit den datenschutzrechtlichen Vorgaben zum Schutz der im System gespeicherten personenbezogenen Daten Dritter in Ausgleich zu bringen.

Die DEK sieht es daher als sachgerecht an, dass der Gesetzgeber Transparenzpflichten durch Regelungen flankiert, die auf Initiative der Betreiber oder auch möglicherweise betroffener Dritter eine **Abwägung der kollidierenden Rechte und Interessen** mit dem Transparenzinteressen der betroffenen Personen oder sonstiger anspruchsberechtigter Privater ermöglicht. **Starre Vorrangregeln**, etwa eine generelle Präferenz für den Schutz von Geschäftsgeheimnissen im Verhältnis zu den Transparenzinteressen, sind hingegen nach Auffassung der DEK (trotz des damit womöglich verbundenen Gewinns an Rechtssicherheit) **nicht sachadäquat**. In jedem Fall, in dem sich Betreiber oder Dritte auf kollidierende Interessen berufen, ist sorgfältig zu prüfen, ob diesen Interessen nicht durch konkrete Schutzmaßnahmen Rechnung getragen werden kann, bevor eine Transparenzpflicht ganz verneint wird. Im Falle auskunftsberechtigter Privatpersonen sind die Anforderungen an die Schutzmaßnahmen und deren Nachweis so zu gestalten, dass sie auch für verletzliche Verbraucher keine prohibitive Schwelle zur Erlangung von Informationen darstellen. Drittinteressen sind etwa durch Anonymisierung zu schützen.

4.1.3 Risikofolgenabschätzung

Die Folgenabschätzung i.S.d. Art. 35 Abs. 1 DSGVO umfasst ausschließlich Informationen zu den Folgen für den Schutz *personenbezogener Daten*, jedoch keine umfassende Risikoanalyse eines algorithmischen Systems. Bei algorithmischen Systemen ab einem gewissen Schädigungspotenzial (ab Stufe 2) ist es jedoch sachgerecht und zumutbar, dem Anbieter/Anwender eine angemessene Risikofolgenabschätzung zur Einschätzung des mit einem System verbundenen Risikos und ihre Veröffentlichung gesetzlich abzuverlangen. Je kritischer das System, desto umfangreicher muss die Risikofolgenabschätzung ausfallen. Sie sollte auch eine Abschätzung der **Risiken für Selbstbestimmung, Privatheit, körperliche Unversehrtheit, persönliche Integrität sowie für Vermögen, Eigentum, und Gleichbehandlung** umfassen und auch Qualitätsmaße und Fairnessmaße zu den Daten und zur Modellgüte enthalten, etwa zu Bias oder (statistischen) Fehlerquoten (insgesamt oder für bestimmte Teilgruppen), die ein System bei der Vorhersage/Kategorienbildung aufweist.

Use Case: Personalisierte Preise I – Transparenzanforderungen

Der zunehmende Einsatz von Preissetzungsalgorithmen im Online-Handel stellt nicht nur das Verbraucherschutzrecht, sondern auch das Wettbewerbsrecht vor Herausforderungen: Preissetzungsalgorithmen können den Markt überblicken, um in Echtzeit Preise an Nachfrage und die Angebote der Wettbewerber anzupassen.

Im Online-Handel können Anbieter dadurch personalisierte Preise (für einzelne Nutzer oder Gruppen) direkt oder über individuelle Rabatte ausspielen. Algorithmische Systeme lassen sich beispielsweise gezielt dazu nutzen, die maximale Zahlungsbereitschaft der Konsumenten abzuschöpfen, oder Nutzer dazu zu bewegen einen Kaufvorgang nicht abzubrechen. Grundlage dieser Personalisierung sind Scoringverfahren, etwa auf der Basis von Echtzeitanalysen des Surfverhaltens der Nutzer oder anderweitig gesammelter Daten. Die zugrunde liegenden algorithmischen Systeme sind i.d.R. „Blackboxes", so dass für Außenstehende die Datengrundlage und Entscheidungslogik der Preisbildung nicht nachvollziehbar sind. Somit besteht ein Risiko preislicher Diskriminierung, etwa von geschützten Bevölkerungsgruppen im Sinne des AGG.

Das Schädigungspotenzial durch höhere personalisierte Preise für einzelne betroffene Verbraucher kann stark variieren. Nichtsdestotrotz können selbst geringe Preisaufschläge für einzelne Güter und Dienstleistungen in der Summe für die einzelnen Betroffenen und für die betreffenden Bevölkerungsgruppen zu beträchtlichen Wohlfahrtsverlusten führen. Insbesondere kann es durch lernende Systeme, etwa über Signaling, auch zu quasi-abgestimmten hohen Marktpreisen kommen. Wenn Wettbewerber Preise oder Konditionen auf dem Umweg über Algorithmen abstimmen, schadet das dem Wettbewerb, der Innovationskraft der Volkswirtschaft und schlussendlich dem Verbraucher; das gilt sowohl für die beabsichtigte Nutzung von Algorithmen zur Preisbeeinflussung als auch dann, wenn Parallelverhalten und hohe Preise (Tacid Collusion) ohne eine solche konkrete Absicht durch lernende Algorithmen zustande kommen und keine direkte Preisabsprache durch Menschen stattfand.

Es reicht nicht aus, wenn diese hohe Kritikalität insgesamt nur Transparenzanforderungen und Kennzeichnungspflichten für Pricing-Systeme auslöst. Auch eine umfassende Folgenabschätzung kann dazu beitragen, Diskriminierungsrisiken eines algorithmischen Pricing-Systems zu erkennen: Ist die verwendete Datengrundlage bekannt, nach denen personalisierte Preise berechnet werden, sollten unabhängige Experten prüfen können, ob diese mit geschützten Bevölkerungsgruppen korrelieren (sog. Proxys), d. h. ob z. B. Frauen oder bestimmte religiöse Gruppen höhere Preise bezahlen müssen. Wenn Verbraucher zudem über Kennzeichnungspflichten darauf aufmerksam gemacht werden, dass Preise bzw. Rabatte personalisiert ausgespielt werden, könnten die Betroffenen über Auskunftsrechte, die für „ihren" Preis verwendeten Daten auf Richtigkeit oder mögliche diskriminierende Faktoren hin überprüfen.

Auch ist die Transparenz über preisrelevante Faktoren wichtig, um die steuernden Effekte indivualisierter Preissetzung auf das Verhalten der einzelnen Verbraucher zu beobachten, da diese freiheitsrelevante Ausmaße annehmen können.

4.1.4 Pflicht zur Dokumentation und zur Protokollierung

Je komplexer, dynamischer und verteilter einzelne IT-Systeme einen Input in einen Output verwandeln, desto wichtiger ist es aus regulatorischer Sicht, die konkreten Ursachen für eine bestimmte Entscheidung nachvollziehbar zu machen. Nur dann lassen sich Fehler aufdecken und Rechtsverletzungen effektiv ahnden. Ein Ansatzpunkt, um die Wirkweise softwarebasierter Verfahren besser zu verstehen, ist es, einzelne Programmschritte digital mitzuschneiden und für Prüfzwecke zu verwenden. Dies kann im Bereich der Verarbeitung personenbezogener Daten auch gemäß dem Datenschutzrecht geboten sein, um das Gebot der Rechenschaftspflicht umzusetzen.

Zum einen sollte eine solche Anforderung von Dokumentation und Protokollierung in Bezug auf die verwendeten Datensätze und Modelle , die Granularität, die Aufbewahrungszeiten und die Verwendungszwecke im Datenschutzrecht konkretisiert werden, damit die Verantwortlichen und Auftragsverarbeiter Rechtsklarheit erhalten. Zum anderen sollte für Systeme, die ein erhebliches Schädigungspotenzial haben (Stufe 4), eine Pflicht etabliert werden, die Programmabläufe zu dokumentieren und zu protokollieren. Die verwendeten Datensätze und Modelle sind so zu beschreiben, dass diese für Aufsichtsinstitutionen im Falle einer Kontrolle nachvollziehbar sind (etwa hinsichtlich der Herkunft und Aufbereitung von Datensätzen oder der Optimierungsziele der Modelle).

4.2 Sonstige Vorgaben für algorithmische Systeme

4.2.1 Allgemeine qualitative Vorgaben an algorithmische Systeme

Der Normgeber sollte Betreibern ein Mindestmaß an **technischen und mathematischen prozeduralen Qualitätsgarantien** abverlangen, welche die Rechtmäßigkeit der algorithmisch ermittelten Ergebnisse durch Verfahrensvorgaben absichern. Dazu können insbesondere Vorgaben für das mathematische Modell und spezifische Verarbeitungsmethoden oder Vorgaben für Korrektur- und Kontrollmechanismen oder für die Datenqualität sowie die Sicherheit des Systems gehören. Um die widerstreitenden Grundrechtspositionen des Softwarebetreibers sowie der Entscheidungsadressaten auszutarieren, sollten die Anforderungen an die Validität der mathematischen Modelle sowie die Sachnähe der zugrunde gelegten Informationen **mit dem Schädigungspotenzial algorithmischer Systeme steigen**.

Bei algorithmenbasierten und algorithmengetriebenen Entscheidungen bedarf es auch eines **kompetenz-sensitiven Designs**. Dieses kann den bewussten Einsatz zwingend zu absolvierender **Trainingsmodule** beinhalten. Auch hat es sich etwa bei Entscheidungsassistenten insbesondere bewährt, systemische **Rollenwechsel** einzuführen, d. h. dem Anwender etwa immer wieder auch einmal die Erstentscheidung ohne Kenntnis des algorithmischen Entscheidungsvorschlags zuzuweisen. Eine für den einzelnen Anwender möglicherweise unangenehmere Variante ist diejenige des **Aufmerksamkeitstests**, d. h. programmiert falsche Entscheidungsvorschläge der Maschine einzustreuen, deren Eigenschaft als Aufmerksamkeitstest noch rechtzeitig aufgedeckt wird, bevor es zu Auswirkungen für andere Menschen kommen kann.

Ferner ist zu gewährleisten, dass Verbesserungsprozesse fair und im Sinne aller Betroffener durchgeführt werden; insbesondere ist sicherzustellen, dass geeignete **Feedbackschleifen** auch den Interessen der betroffenen Personen, nicht nur der Betreiber, Rechnung tragen. Hinsichtlich der Datenqualität sind auch Vorgaben angezeigt, inwieweit für bestimmte Anwendungsbereiche die Nutzung von Schätz- oder sog. Proxy-Daten (Teil C, 2.2.2 f.) zulässig oder verboten sein sollte.

Zusätzlich zu den Anforderungen, die der eigentliche Zweck der Verarbeitung an das algorithmische System stellt, sollten bei der Gestaltung die Anforderungen der **Sicherheit** erfüllt werden. Dazu sollten die individuellen Anforderungen sämtlicher Beteiligter Berücksichtigung finden, um bei der Konzeptionierung, Implementierung und im Betrieb die geeigneten Gestaltungsentscheidungen zu treffen. Die Einschätzung der Risiken obliegt zwar in der Regel vorrangig dem Betreiber des Systems, doch kann er dies nur leisten, wenn er auf eine ausreichende Dokumentation, z. B. eine Risikofolgenabschätzung des Herstellers, zurückgreifen kann. Nötig ist auch die Klarheit darüber, wer für welchen Bereich verantwortlich ist. Die DEK empfiehlt dabei in als kritisch identifizierten Bereichen rechtliche Vorgaben bzgl. der

- Mindeststandards an erforderlicher Sicherheit und den zu treffenden Maßnahmen;
- Spezifika, wie und unter welchen Voraussetzungen Testverfahren (etwa zur Identifikation von Bias bzw. diskriminierenden Verzerrungen) algorithmischer Systeme bei Herstellern oder Betreibern auszugestalten und durchzuführen sind;
- Rechtsfolgen bei Sicherheitslücken oder anderen Fehlern;
- Dokumentationspflichten hinsichtlich der Funktionsweise und der Tests, die Anwender erhalten, um das Risiko abschätzen zu können; und
- Verpflichtung, Systemaktualisierungen in einem definierten Zeitraum durchzuführen und darüber zu berichten.

4.2.2 Besondere Schutzmaßnahmen beim Einsatz algorithmischer Systeme im Kontext menschlicher Entscheidungen

Der Mensch darf nicht zum Objekt der Technik werden. Dieser für die Regulierung algorithmischer Systeme zentrale Grundsatz entfaltet seine Wirkung insbesondere dort, wo algorithmische Systeme zum Einsatz kommen, um menschliche Entscheidungen zu unterstützen oder um Entscheidungsprozesse zu automatisieren, d. h. menschliche Entscheidungen durch technische Prozesse zu ersetzen.

Im geltenden Recht kodifiziert Art. 22 DSGVO diesen Grundsatz bereits für bestimmte algorithmische Systeme, die in den Anwendungsbereich der DSGVO fallen: Niemand darf einer ausschließlich auf einer automatisierten Verarbeitung – einschließlich Profiling – beruhenden Entscheidung unterworfen werden, die ihm gegenüber rechtliche oder andere erhebliche Folgen hat – es sei denn, es ist für den Abschluss oder die Erfüllung eines Vertrags notwendig, der Betroffene hat explizit eingewilligt oder es gibt eine gesetzliche Erlaubnis. Soweit eine solche vollständig automatisierte Entscheidung zulässig ist, muss der Verantwortliche Schutzmaßnahmen treffen, um die Rechte und Interessen der betroffenen Personen zu wahren.[6] Zudem gelten verschärfte Informationspflichten und Auskunftsrechte.[7]

6 Vgl. Art 22 Abs. 3 DSGVO.
7 Vgl. Art. 13 Abs. 2 lit. f DSGVO, Art. 14 Abs. 2 lit. g DSGVO und Art. 15 Abs. 1 lit. h DSGVO.

Gegenwärtig besteht aus Sicht der DEK in Bezug auf diese Normen an verschiedenen Punkten noch **Klarstellungsbedarf**. Die mit Art. 22 DSGVO – „einschließlich Profiling" – verbundenen Informationspflichten und Auskunftsrechte sollten sich auch auf die automatisierte **Profilbildung als solche** beziehen. So sehen sich beispielsweise einzelne Wirtschaftsauskunfteien nicht von diesen Normen erfasst, da sie lediglich eine Profilbildung vornähmen, die „Entscheidungen" aber erst durch die Unternehmen, die beispielsweise einen Kreditscore abfragen, getroffen würden. Diese Argumentation trägt der Intention der DSGVO aus der Sicht der DEK jedoch nicht ausreichend Rechnung. Denn die langfristigen Auswirkungen einer solchen Profilbildung auf die Betroffenen können zum einen erheblich sein, zum anderen hebt die DSGVO die Profilbildung besonders hervor. Soweit die Datenschutzbehörden und Gerichte das geltende Recht im Wege einer am Schutzzweck der DSGVO orientierten Auslegung entsprechend weit anwenden können, ist dies zu begrüßen. Parallel dazu ist jedoch aufgrund der hohen Grundrechtssensibilität dieser Frage der demokratisch legitimierte Gesetzgeber aufgerufen, die rechtlichen Rahmenbedingungen zeitnah weiter zu konkretisieren, um möglichst schnell Rechtssicherheit zu schaffen. Die DEK empfiehlt der Bundesregierung, sich hierfür im Rahmen der Evaluation der DSGVO einzusetzen.

Klarstellungs- und Konkretisierungsbedarf besteht auch hinsichtlich der Frage, wann eine Entscheidung gemäß **Art. 22 DSGVO** „ausschließlich" auf einer automatisierten Verarbeitung personenbezogener Daten „beruht" und wie weit der Begriff der „Beeinträchtigung in ähnlicher Weise" sowie die Schutzrechte des Art. 22 Abs. 3 DSGVO reichen. Die DEK empfiehlt der Bundesregierung, sich bei der Evaluation der DSGVO dafür einzusetzen, dass der Anwendungsbereich des Art. 22 DSGVO eine Konkretisierung erfährt. Das Schädigungspotential algorithmendeterminierter Entscheidungssysteme, die ursprünglich das Leitbild des Art. 22 DSGVO gebildet hatten, unterscheidet sich insbesondere nicht kategorial von demjenigen vieler algorithmengetriebener Entscheidungssysteme. Dafür ist auch die Neigung menschlicher Akteure, Empfehlungen algorithmischer Systeme schlicht zu übernehmen und bestehendes Ermessen nicht auszuüben, mitverantwortlich.

Im Lichte des im Einzelnen stark differierenden Schädigungspotentials algorithmenbasierter Systeme erscheint es der DEK nicht angemessen, das Verbotsprinzip des Art. 22 DSGVO generell auszuweiten. Insbesondere eignet sich der Grundsatz menschlicher Letztentscheidung aus Art. 22 Abs. 3 DSGVO nicht für alle algorithmischen Systeme gleichermaßen. So wäre für algorithmische Systeme, bei denen keine „Entscheidung" des Systems im Sinne der bisherigen Fassung des Art. 22 Abs. 1 DSGVO vorliegt, auch ein Recht auf menschliche Letztentscheidung regelmäßig wenig praktikabel und zudem oft auch nicht wünschenswert. Stattdessen empfiehlt die DEK ein risikoadaptiertes Regulierungsregime, das dem Einzelnen angemessene Schutzgarantien (insbesondere gegen Profiling) und Verteidigungsmöglichkeiten gegen Fehler und Bedrohungen seiner Rechte vermittelt.

Der Rechtsgedanke, dass der Mensch nicht zum Objekt technischer Systeme werden darf, sollte auch in dem horizontalen EU-Rechtsakt einer EUVAS (→ oben 3.3) zur risikoadaptierten Regulierung algorithmischer Systeme, den die DEK anregt, sowie in den begleitenden sektoralen Rechtsakten einen **zentralen normativen Ankerpunkt** bilden. In diese Rechtsakte sind daher Regelungen aufzunehmen, die auch außerhalb des Anwendungsbereichs des Art. 22 DSGVO Vorgaben für algorithmenbasierte Entscheidungssysteme treffen. Soweit die neue Regulierungsschicht algorithmische Systeme miterfasst, die auch in den – gegebenenfalls im Lichte der hiesigen Empfehlungen modifizierten – Anwendungsbereich des Art. 22 DSGVO fallen, ist auf eine präzise **Synchronisierung der Regelungssysteme** zu achten.

4.2.3 Recht auf angemessene algorithmische Schlussfolgerungen?

Die Prozesse datenbasierter Generierung sog. **algorithmischer Schlussfolgerungen** über die vermeintlichen Interessen, Neigungen und Charaktereigenschaften individueller Personen, insbesondere von Verbrauchern, verdienen aus Sicht der DEK höchste gesellschaftliche und politische Aufmerksamkeit. In der digitalen Wirtschaft sind derartige Schlussfolgerungen ubiquitär. Für viele digitale Geschäftsmodelle, die auf die feingranulare Personalisierung bestimmter Angebote oder Dienste ausgerichtet sind, sind sie geradezu kennzeichnend. Viele Verbraucherinnen und Verbraucher schätzen den Komfort solcher Angebote und Dienste, doch ergeben sich aus ihnen auch Gefahren, wenn Schlussfolgerungen auf einer falschen Datenbasis erfolgen oder infolge der Unzulänglichkeit anderer Systemkomponenten sonst Ergebnisse erzielt werden, die inhaltlich unangemessen sind.

Um den Gefahren zu begegnen, die durch bestimmte algorithmische Schlussfolgerungen erwachsen können, wollen manche dem Betroffenen ein „Recht auf angemessene Schlussfolgerungen“ normativ verbürgen.[8] Dieser Vorschlag sieht ein Gesamtpaket an Maßnahmen vor, das den jeweils Betroffenen ein wirksames Kontrollinstrument über die Schlussfolgerungen an die Hand geben soll, die Betreiber algorithmischer Systeme über sie erstellt haben. Neben einem materiellen Recht, angemessenen Schlussfolgerungen unterworfen zu werden, sieht es eine Pflicht des Betreibers vor, den Betroffenen ohne Auskunftsverlangen darüber unterrichten müssen, dass und wieso die gezogenen Schlussfolgerungen „angemessen“ waren.

Die Datenethikkommission begrüßt die Diskussion, die der Vorschlag eines solchen „Rechts auf angemessene Schlussfolgerungen“ angestoßen hat. Sie gibt jedoch zu bedenken, dass ein derartiges Recht verfassungsrechtlich geschützte Interessen der Betreiber algorithmischer Systeme tangieren kann. Auf diese Schutzpositionen ist nach Auffassung der Datenethikkommission bei einer etwaigen regulatorischen Ausgestaltung des Vorschlags Rücksicht zu nehmen, etwa durch eine Beschränkung des Anwendungsbereichs auf Systeme, die aufgrund ihrer Teilhabe- und Grundrechtsrelevanz eine hohe Kritikalität aufweisen.

4.2.4 Gesetzlicher Diskriminierungsschutz

Ein wesentliches Ziel der Regulierung algorithmenbasierter, -getriebener und -determinierter Entscheidungssysteme besteht darin, die Diskriminierung eines Menschen aufgrund eines in Art. 3 Abs. 3 GG bzw. Art. 21 Abs. 1 GRCh genannten Merkmals – und darüber hinaus jede sachlich nicht gerechtfertigte Ungleichbehandlung – zu verhindern sowie die persönliche Integrität von Betroffenen zu schützen. Während staatliche Stellen bei jeder Form hoheitlichen Handelns unmittelbar einer **Grundrechtsbindung** und damit auch einem umfassenden Diskriminierungsverbot unterliegen, bedarf es bei privaten Akteuren dafür einer einfachgesetzlichen Grundlage. Den regelungstechnischen Anknüpfungspunkt dafür markiert grundsätzlich das **Allgemeine Gleichbehandlungsgesetz (AGG)** – flankiert durch Generalklauseln des Privatrechts, etwa zu sittenwidrigen Verträgen.

Damit eine Ungleichbehandlung unter Privaten dem AGG unterfällt, muss es sich zum einen um eine Ungleichbehandlung aufgrund eines **sensiblen Merkmals** handeln (Rasse, ethnische Herkunft, Geschlecht, Religion, Behinderung, Alter, sexuelle Identität); zum anderen muss der **situative Anwendungsbereich** eröffnet sein (Beschäftigungskontext oder Zugang zu Gütern und Dienstleistungen, einschließlich Wohnraum, die der Öffentlichkeit zur Verfügung stehen).

8 Omer Tene / Jules Polonetsky: Northwestern Journal of Technology and Intellectual Property, 2013 (11:5), S. 240, 270 f.; Sandra Wachter / Brent Mittelstadt: Columbia Business Law Review, 2019 (2), S. 1, 1 ff. Der Vorschlag besteht aus einer materiellen Komponente und einer Verfahrenskomponente.

Zwar erfassen die Normen des AGG im Grundsatz bereits nach geltendem Recht Ungleichbehandlungen durch algorithmische Systeme. Allerdings sind nicht alle diskriminierungsanfälligen Sachmaterien in den Anwendungsbereich des AGG einbezogen und erfasst das AGG nicht sämtliche sensiblen Konstellationen, in denen algorithmisch ermittelte Ergebnisse eine Diskriminierung auslösen oder begünstigen (z. B. im Falle der Vergabe eines Immobilienkredits aufgrund einer individuellen Risikoprüfung). Es ist daher erwägenswert, den **situativen Anwendungsbereich des AGG** etwa auf alle automatisierten Entscheidungsverfahren auszudehnen oder einzelne Sachbereiche besonders persönlichkeitssensibler algorithmischer Schlussfolgerungen ergänzend aufzunehmen.[9] Dies betrifft vor allem solche Sachbereiche, welche die Lebensgestaltung nachhaltig beeinträchtigen können, wie z. B. Verbraucherverträge, die auf der Grundlage eines Scorings zustande kommen oder auf besonders risikoträchtigen Verfahren basieren, Methoden der Gesichtserkennung oder Preisdiskriminierung in bestimmten Lebensbereichen wie der Gesundheitsversorgung. Der gleichfalls grundrechtlich geschützten allgemeinen Handlungsfreiheit des vertraglichen Gegenübers ist dabei angemessen Rechnung zu tragen.

Zu diskutieren ist auch, ob im Zusammenhang mit algorithmischen Systemen der Gesetzgeber die Beschränkung auf bestimmte Diskriminierungsmerkmale aufgeben sollte. In diskriminierenden Effekten algorithmischer Systeme spiegelt sich nur teilweise eine in der Gesellschaft bestehende Verzerrung hinsichtlich **klassischer Diskriminierungsmerkmale** wider, etwa soweit die Verzerrung in den Trainingsdaten oder im verwendeten Modell liegt. Dies wäre etwa der Fall, wenn ein zur Bewerberauswahl verwendetes System anhand der Daten erfolgreicher Führungskräfte der Vergangenheit trainiert wurde, die ganz überwiegend männlich waren. Allerdings geht das Diskriminierungspotenzial algorithmischer Systeme deutlich darüber hinaus, z. B. wenn eine Benachteiligung systematisch an nicht gesetzlich verbotene Gruppenmerkmale (z. B. Wohnadresse in einem bestimmten Bezirk) oder gar an im Wege der Mustererkennung ermittelte, aber eher zufällige Korrelationen anknüpft. Teilweise lassen sich diese Konstellationen bereits über die Figur der **mittelbaren Diskriminierung** in den Griff bekommen. Insoweit bedarf es dann gegebenenfalls ergänzend geeigneter Beweiserleichterungen. Teilweise stellen sich hier aber auch ganz neue Gerechtigkeitsfragen. Diese betreffen nicht nur die Verteilung von Chancen zu Lasten traditionell marginalisierter Gemeinschaften, sondern auch den Ausschluss von Gruppen, die anhand mehr oder weniger zufälliger Merkmale zusammengewürfelt sind: Die Eigenheiten maschinellen Lernens schaffen **neue Diskriminierungsmerkmale**, die aber dadurch, dass trainierte Algorithmen auch in anderen Einsatzbereichen verwendet werden, enorme Breitenwirkung entfalten können.

9 Mario Martini: Juristenzeitung (JZ), 2017, S. 1017, 1021.

Daher ist es angezeigt, eine Erweiterung des Diskriminierungsschutzes auf jede systematische und sachlich ungerechtfertigte Benachteiligung aufgrund eines Gruppenmerkmals zu erwägen. Die DEK empfiehlt der Bundesregierung, auch diesbezüglich eine entsprechende **Erweiterung des AGG oder alternativ die Verankerung in künftiger, spezieller Algorithmen-Gesetzgebung zu prüfen**. Eine besondere regulatorische Schwierigkeit geht dabei davon aus, dass eine – prinzipiell nicht abgeschlossene – Fülle von Gruppenmerkmalen existiert, die eine derartige algorithmische Diskriminierung nach sich ziehen können und damit einziges Abgrenzungskriterium zwischen diskriminierungsrechtlich relevanten und irrelevanten Benachteiligungen der systematische Charakter wäre. Eine entsprechende Regelung des materiellen Diskriminierungsschutzes müsste daher jedenfalls einerseits durch entsprechende Auskunfts- und Begründungspflichten und andererseits durch verschiedene Mechanismen interner und externer Kontrolle flankiert sein, für welche die neue Regelung den materiellen Prüfungsmaßstab bilden würde. Die Folgen einer derartigen Regulierung auf alle Beteiligten wären in jedem Fall sorgfältig abzuschätzen und abzuwägen.

Ganz unabhängig von der Frage einer tatbestandlichen Erweiterung ist zu erwägen, ob die **Beweisregelungen** den Charakteristika algorithmischer Systeme bereits hinreichend gerecht werden. Zwar erfordert die Feststellung einer mittelbaren Diskriminierung weder den Nachweis einer diskriminierenden Absicht noch einen eindeutigen Kausalitätsnachweis. Vielmehr reicht es aus Sicht der Geschädigten aus, eine Korrelation zwischen den Entscheidungen und sensiblen Kriterien aufzuzeigen. Beim Einsatz algorithmischer Systeme ist aber dieser Nachweis für die Betroffenen in der Regel schwer zu erbringen.

Daher empfiehlt die DEK dem Gesetzgeber, die Anforderungen an diesen Nachweis einer Diskriminierung durch den Betreiber algorithmischer Systeme klarstellend gesetzgeberisch zu regeln und ggf. für Betroffene noch weiter abzusenken. Darum ist das AGG stets zusammen mit **Auskunftsrechten und Begründungspflichten** (→ siehe) zu denken, ohne die dem Geschädigten eine Wahrnehmung seiner Rechte oft nicht möglich sein wird. Den dadurch betroffenen Schutzinteressen Dritter sowie der Verwender der Systeme muss dabei hinreichend Rechnung getragen werden.

4.2.5 Präventives behördliches Zulassungsverfahren für besonders riskante algorithmische Systeme

Zusätzlich zu bereits bestehender Regulierung ist es für algorithmische Systeme mit deutlichem oder regelmäßigem (Stufe 3) oder sogar erheblichem Schädigungspotenzial (Stufe 4) sinnvoll, Zulassungsverfahren oder Vorabprüfungen von algorithmischen Systemen durch Aufsichtsinstitutionen zu etablieren, um Schäden für einzelne Betroffene, Bevölkerungsgruppen oder die Gesellschaft als Ganzes abzuwenden.

Zusammenfassung der wichtigsten Handlungsempfehlungen

Instrumente

45

Die DEK empfiehlt bei algorithmischen Systemen erhöhter Systemkritikalität (ab Stufe 2) eine **Kennzeichnungspflicht**: Eine solche Pflicht trägt Betreibern auf, deutlich zu machen, wann und in welchem Umfang algorithmische Systeme zum Einsatz kommen (Information über das „Ob“). Eine Kennzeichnungspflicht sollte unabhängig von der Systemkritikalität stets im Falle einer ethisch relevanten Verwechselungsgefahr zwischen Mensch und algorithmischem System bestehen.

46

Das Recht einer betroffenen Person auf aussagekräftige **Informationen** über die „involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen“ eines algorithmischen Systems (vgl. DSGVO) sollte nicht nur für vollständig automatisierte Systeme, sondern bereits für **Profilbildungen als solche** und unabhängig von einer nachgelagerten Entscheidungssituation bestehen. Es sollte – abgestuft nach der Systemkritikalität – künftig auch bereits für algorithmenbasierte Entscheidungen greifen. Dazu sollte teilweise eine gesetzliche Klarstellung und teilweise eine Erweiterung der Regelung auf europäischer Ebene erfolgen.

47

In bestimmten Bereichen kann es sachgerecht sein, dem Betreiber algorithmischer Systeme zusätzlich zur allgemeinen Erläuterung der Logik (Vorgehensweise) und Tragweite des Systems eine **individuelle Erklärung** der getroffenen Entscheidung abzuverlangen. Wesentlich ist dabei, dass betroffene Personen verständlich, relevant und konkret informiert werden. Die DEK begrüßt daher die technischen Bemühungen, die Erklärbarkeit algorithmischer (insbesondere selbstlernender) Systeme zu stärken („Explainable AI“), und empfiehlt der Bundesregierung, die weitere Forschung und Entwicklung in diesem Bereich zu fördern.

48

In bestimmten Sektoren, in denen nicht nur individuelle, sondern in besonderem Maße auch gesellschaftliche Interessen berührt sind, sollten auch **nicht unmittelbar betroffene Personen** ein Recht auf Zugang zu bestimmten Informationen über die algorithmischen Systeme erhalten. Entsprechende Rechte werden in erster Linie für journalistische und Forschungszwecke infrage kommen und sind zudem mit Blick auf die betroffenen Interessen der Betreiber durch hinreichende Schutzmaßnahmen zu flankieren. Unter Umständen, insbesondere beim staatlichen Einsatz von algorithmischen Systemen mit einem erheblichen Schädigungspotenzial (Stufe 4), kommen nach Ansicht der DEK darüber hinaus auch voraussetzungslose Informationszugangsansprüche in Frage.

49

Bei algorithmischen Systemen ab einem gewissen Schädigungspotenzial (ab Stufe 2) ist es sachgerecht und zumutbar, dem Betreiber gesetzlich die Erstellung und Veröffentlichung einer angemessenen **Risikofolgenabschätzung** abzuverlangen, die auch bei der Verarbeitung nicht-personenbezogener Daten greift und Risiken außerhalb des Datenschutzes berücksichtigt. Sie sollte insbesondere auch eine Abschätzung der Risiken für Selbstbestimmung, Privatheit, körperliche Unversehrtheit, persönliche Integrität sowie Vermögen, Eigentum und Diskriminierung umfassen. Außerdem sollte sie neben den zugrundeliegenden Daten und der Logik des Modells auch Qualitätsmaße und Fairnessmaße zu den Daten und zur Modellgüte berücksichtigen, etwa zu Bias oder (statistischen) Fehlerquoten (insgesamt oder für bestimmte Teilgruppen), die ein System bei der Vorhersage/Kategorienbildung aufweist.

50

Die Anforderungen an **Dokumentation und Protokollierung** in Bezug auf die verwendeten Datensätze und Modelle, die Granularität, die Aufbewahrungszeiten und die Verwendungszwecke sollten konkretisiert werden, damit die Verantwortlichen und Auftragsverarbeiter Rechtsklarheit erhalten. Zum anderen sollte für sensible Anwendungen künftig eine Pflicht etabliert werden, die Programmabläufe einer Software, die nachhaltige Schäden verursachen können, zu dokumentieren und zu protokollieren. Die verwendeten Datensätze und Modelle sind so zu beschreiben, dass diese für Aufsichtsinstitutionen im Falle einer Kontrolle nachvollziehbar sind (etwa hinsichtlich der Herkunft und Aufbereitung von Datensätzen oder der Optimierungsziele der Modelle).

51

Der Normgeber sollte Betreibern ein Mindestmaß an **technischen und mathematisch-prozeduralen Qualitätsgarantien** abverlangen, welche die Korrektheit und Rechtmäßigkeit der algorithmisch ermittelten Ergebnisse durch Verfahrensvorgaben absichern. Dazu können insbesondere Vorgaben für Korrektur- und Kontrollmechanismen oder für die Datenqualität sowie die Sicherheit des Systems gehören. So wäre es beispielsweise sachgerecht, qualitative Anforderungen an das Verhältnis zwischen der Datengrundlage und dem Ergebnis des algorithmischen Datenverarbeitungsprozesses vorzugeben.

52

Beim Einsatz algorithmischer Systeme im Kontext menschlicher Entscheidungen sieht die DEK zunächst Klarstellungs- und Konkretisierungsbedarf betreffend die Anwendungsvoraussetzungen und Rechtsfolgen von Art. 22 DSGVO. Darüber hinaus empfiehlt die DEK, **Schutzmechanismen auch für algorithmenbasierte und -getriebene Entscheidungssysteme** vorzusehen, da sich der Einfluss dieser Systeme in der Praxis nahezu ebenso stark auswirken kann wie bei algorithmendeterminierten Anwendungen. Diesbezüglich empfiehlt sich anstelle des von Art. 22 DSGVO bislang verfolgten Verbotsprinzips ein flexibleres, risikoadaptiertes Regulierungsregime, das dem Einzelnen angemessene Schutzgarantien (insbesondere im Falle von Profiling) und Verteidigungsmöglichkeiten gegen Fehler und Bedrohungen seiner Rechte vermittelt.

53

Es ist erwägenswert, den **Anwendungsbereich des Antidiskriminierungsrechts** in situativer Hinsicht auf Diskriminierungen auszudehnen, die auf einer automatisierten Datenauswertung oder einem automatisierten Entscheidungsverfahren beruhen. Der Gesetzgeber sollte darüber hinaus Maßnahmen eines wirksamen Schutzes gegen **Diskriminierungen aufgrund von Gruppenmerkmalen** etablieren, die an sich nicht zu den gesetzlich geschützten Diskriminierungsmerkmalen zählen, und bei denen Diskriminierungen derzeit vielfach auch nicht als mittelbare Diskriminierung aufgrund eines geschützten Merkmals qualifiziert werden können.

54

Zusätzlich zu bereits bestehender Regulierung ist es für algorithmische Systeme mit deutlichem oder regelmäßigem (Stufe 3) oder sogar erheblichem Schädigungspotenzial (Stufe 4) sinnvoll, **Zulassungsverfahren oder Vorabprüfungen** von algorithmischen Systemen durch Aufsichtsinstitutionen zu etablieren, um Schäden für einzelne Betroffene, Bevölkerungsgruppen oder die Gesellschaft als Ganzes abzuwenden.

5. Institutionen

Die Verantwortlichkeit für den ethisch vertretbaren und rechtmäßigen Einsatz algorithmischer Systeme muss nach Auffassung der DEK auf mehrere Schultern verteilt werden. Die gegenwärtig bereits bestehenden Institutionen und Aufsichtsstrukturen sind nicht ausreichend darauf vorbereitet, um die abgestufte Kontrolle algorithmischer Systeme hinreichend effektiv wahrnehmen zu können. Daher fordert die DEK die Bundesregierung dazu auf, die bestehenden Institutionen und Strukturen im Rahmen ihrer Zuständigkeit zu stärken, neu auszurichten und, wo erforderlich, auch neue Institutionen und Strukturen zu schaffen.

5.1 Behördliche Kompetenzen und fachliche Expertise

5.1.1 Verteilung der Aufsichtsaufgaben im sektoralen Kontrollverbund

Die DEK empfiehlt der Bundesregierung, die behördlichen Aufsichtsaufgaben und Kontrollbefugnisse im Grundsatz jeweils den Behörden zuzuweisen, die bereits **sektorspezifische Sachkompetenzen** haben. Gleiches sollte aus Sicht der DEK bei jenen Materien geschehen, die in die Verwaltungskompetenz der Länder fallen.

Konkret hält es die DEK für sinnvoll, die Aufsicht über den Einsatz algorithmischer Systeme durch Private in den Bereichen der digitalen Wirtschaft, in denen bereits Behörden mit sektorspezifischen Zuständigkeiten existieren, an die **bestehenden Behörden** anzubinden. Zu denken ist hier an Behörden wie die Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), die Bundesnetzagentur (BNetzA), das Bundesamt für Sicherheit in der Informationstechnik (BSI) oder das Kraftfahrtbundesamt (KBA). Eine besondere Stellung kommt ferner dem Bundeskartellamt (BKartA) und den Datenschutzaufsichtsbehörden zu, die je horizontale, d. h. über die verschiedenen Wirtschaftsbereiche hinweg reichende Zuständigkeiten innehaben.

Um die Tätigkeit der mit Algorithmenaufsicht befassten Behörden zu koordinieren, hält die DEK einen **„Kontrollverbund für kritische algorithmische Systeme“** auf nationaler und EU-Ebene für sachgerecht. Für diese Zwecke sind insbesondere Regelungen zur Verteilung von Zuständigkeiten im Verbund, zum Informationsaustausch, zur Organisation verbundförmig durchgeführter Verwaltungsverfahren und zum Rechtsschutz sachgerecht.

Um Aufsichtslücken zu vermeiden, fordert die DEK Bund und Länder dazu auf, Bereiche zu identifizieren, in denen für eine Kontrolle kritischer algorithmischer Systeme **bisher keine hinreichend sektorspezifisch-sachkundige Behörde** existiert, für die eine Zuweisung der Kontrollaufgabe nahe liegt. Regelmäßig wird es in diesen Fällen nach Auffassung der DEK zweckmäßig sein, bei einem entsprechenden Kontrollbedarf eine der horizontal zuständigen Behörden mit der Materie zu betrauen. Im Falle algorithmischer Systeme, die sensible personenbezogene Daten verarbeiten, können etwa die Datenschutzbehörden sachadäquate Kompetenzträger sein. Im Einzelfall hält es die DEK jedoch für möglich, dass es erforderlich ist, ganz neue behördliche Kontrollstrukturen aufzubauen. Im Lichte der sich stetig wandelnden technischen Entwicklungen sollten Bund und Länder diese Überprüfung regelmäßig vornehmen.

Um ihrer Aufgabe der Aufsicht über algorithmische Systeme wirkungsvoll nachzukommen, stehen Behörden vor einer strukturellen Herausforderung: Der Gegenstand ihrer Aufsicht weist eine hohe technische Komplexität auf und unterliegt dynamischen Veränderungsprozessen. Die DEK hält daher die **praktische Befähigung der Behörden** für besonders wichtig. Sie empfiehlt der Bundesregierung nachdrücklich, die auf Bundesebene zuständigen Behörden mit den erforderlichen finanziellen, personellen und technischen Ressourcen auszustatten. Der Entwurf eines Besoldungsstrukturenmodernisierungsgesetzes, das ab 2020 die Gehälter und Zulagen von IT-Fachkräften im öffentlichen Dienst erhöhen und neu regeln soll, ist als erster Schritt ausdrücklich zu begrüßen. Im Lichte der erheblichen Herausforderung, gut geschulte Fachkräften für die Verwaltung zu gewinnen, werden jedoch weitere Maßnahmen zeitnah erforderlich sein.

Darüber hinaus empfiehlt die DEK der Bundesregierung, eine behördliche Einheit in Form eines **Kompetenzzentrums Algorithmische Systeme** zu etablieren, welche die sektoralen Behörden bei der Aufgabe unterstützt, algorithmische Systeme zu überwachen. Die Aufgabe einer solchen Einrichtung sollte es nicht nur sein, das für die Aufsicht über kritische algorithmische Systeme erforderliche sachlich-methodische Wissen zu gewinnen, auszuwerten, weiterzuentwickeln und weiterzugeben. Das Kompetenzzentrum sollte (in Abstimmung und auf Anforderung der sektorspezifischen Behörden) vor allem die sektorspezifischen Aufsichtsbehörden beim Aufbau der Expertise unterstützen, die erforderlich ist, um die Aufgaben zu erledigen und algorithmische Systeme mit Blick auf ihre Kritikalität zu evaluieren. Dies erstreckt sich insbesondere auf die Aufgabe des Kompetenzzentrums**, Kriterien, Verfahren und Werkzeuge** für die Kontrolle algorithmischer Systeme fortzuentwickeln. Dazu gehören auch **Maßstäbe, um die Kritikalität zu beurteilen**, und die Konformität kritischer algorithmischer Systeme zu prüfen. Ein solches Kompetenzzentrum nimmt darüber hinaus wichtige **vermittelnde Beratungsfunktionen** war: Es berät im Rahmen ihrer Möglichkeiten nicht nur Stellen des Bundes, der Länder und Kommunen, sondern auch der Hersteller, Betreiber, Anwender und betroffener Personen im Umgang mit und bei der Entwicklung von algorithmischen Systemen. Darüber hinaus nimmt sie an internationalen und europäischen Initiativen zum Aufbau hinreichender Kontrollexpertise einschließlich Normierungsverfahren teil. Eigene Aufsichtsbefugnisse hat das Kompetenzzentrum demgegenüber nicht. Diese verbleiben bei den sektoralen Aufsichtsbehörden. Die Serviceeinheit sollte entweder als eigenständige Bundesbehörde neu errichtet werden oder an eine bestehende Querschnittsbehörde, wie etwa das BSI, angebunden werden.

Perspektivisch erscheint es nach Auffassung der DEK sinnvoll, auch auf der **Ebene der Europäischen Union** eine entsprechende Stelle, etwa in Form einer Agentur, anzusiedeln. Hierauf sollte die Bundesregierung hinwirken.

Soweit staatliche Stellen sich bei der Erledigung ihrer Aufgaben und zusätzlich zum Aufbau eigenen Sachverstandes auch der **Expertise Privater** bedienen wollen oder in die Aufgabenerledigung Private einbeziehen wollen, stehen dem nach Auffassung der DEK keine prinzipiellen Hindernisse entgegen, solange die allgemeinen verfassungs- und verwaltungsrechtlichen Vorgaben für derartige Kooperationen beachtet werden. Im Gegenteil können entsprechende Kooperationen, etwa auch in Form der Beleihung, genutzt werden, um dem gegenwärtigen Mangel an Fachkräften und Fachkenntnissen in der Verwaltung entgegenzuwirken.

5.1.2 Aufgabenangemessene Ausgestaltung der Kontrollbefugnisse

Den jeweils zuständigen Behörden sollte der Normgeber die zur Aufsicht über algorithmische Systeme notwendigen **Eingriffsbefugnisse**, u. a. Auskunfts-, Einsichts- und Zugangsrechte, hinreichend klar **durch Gesetz zuweisen**. Blaupausen für derartige behördliche Befugnisse zur inhaltlichen Kontrolle bestehen in verschiedenen Rechtsgebieten.[10]

Die zuständigen Aufsichtsbehörden müssen jederzeit die Möglichkeit haben, algorithmische Systeme in sensiblen Anwendungsfeldern oder solche mit hohem Schädigungspotenzial zu **überprüfen**. Die dabei zur Anwendung kommenden Überprüfungs- und Testverfahren müssen insbesondere Systeme umfassen, bei denen eine Interaktion mit dem Nutzer erfolgt. Dies kann beispielsweise über standardisierte Schnittstellen erfolgen. Mit diesem Zugang lassen sich sog. Input-Output-Tests durchführen, die z. B. prüfen, ob ein algorithmisches System systematisch Gruppen benachteiligt. Dies ist insbesondere bei lernenden Systemen sinnvoll, die im Laufe der Zeit ihre internen Regeln anpassen. Dabei muss sichergestellt sein, dass die Prüfung von lernenden Systemen nicht zu einer Änderung des Regelsystems führt, indem das System während der Prüfung aus den Prüfungsdaten lernt.

10 Beispielsweise regelt Art. 58 DSGVO die Untersuchungsbefugnisse der Datenschutzaufsicht, § 32e GWB regelt die Sektoruntersuchungen durch das Bundeskartellamt. Die Kontrolle des Hochfrequenzhandels durch Finanzaufsichtsbehörden basiert auf § 6 Abs. 4 WpHG, § 3 Abs. 4 Satz 4 Nr. 5 BörsG n.F. i.V.m. § 7 Abs. 3 BörsG.

Bei der gesetzlichen Kompetenzzuweisung ist sicherzustellen, dass die Aufsichtsbehörden im Falle eines festgestellten Rechtsverstoßes die Befugnis haben, die Betreiber der algorithmischen Systeme zu verpflichten, die Systeme rechtskonform zu gestalten (zum Beispiel durch Anpassung der Datenbasis) und ggf. **Sanktionen** auszusprechen. Die Aufsichtsbehörden sollen, sofern dies im Einzelfall verhältnismäßig ist, auch behördliche **Verbote** des Einsatzes rechtswidriger algorithmischer Systeme (oder ihrer Komponenten) aussprechen können.

5.1.3 Kritikalitätsangemessene Kontrolltiefe

Wer das Verhalten eines algorithmischen Systems wirksam überprüfen will, muss **alle Elemente des algorithmischen Systems** im Blick haben. Eine behördliche Überprüfung kann sich – und muss sich ggf. – auf die Trainingsdaten und verwendeten Lernverfahren, das finale Regelmodell sowie die verwendeten Inputdaten und Outputdaten, die den Entscheidungen zugrunde liegen, erstrecken. Um Biases oder (statistische) Fehlerquoten (insgesamt oder für bestimmte Teilgruppen) zu identifizieren, die ein System aufweist, können zudem Qualitätsmaße zur Datengrundlage und Modellgüte (Trainingsmodell, finales Entscheidungsmodell) Berücksichtigung finden. In methodischer Hinsicht kann eine Prüfung durch Analyse großer Datenmengen, die Prüfung der Gewichtung von Faktoren in komplexen multidimensionalen Modellen sowie eine Input-Throughput-Output-Analyse erfolgen.

Aufgrund der Komplexität der Materie und involvierten Datenmengen, kann der Einsatz von Kontrollalgorithmen die Effizienz und Effektivität der Überprüfung erheblich steigern. Sie können systematisch nach auffälligen Mustern in der Datenbasis und den Ergebnissen eines algorithmischen Systems suchen, die beispielsweise Aufschluss über eine Diskriminierung geben können.

Welches Maß an Kontrolltiefe im konkreten Fall erforderlich ist, sollte sich nach dem Einsatzbereich und der Kritikalität des Systems bestimmen. Bei Systemen, die nur ein gewisses Schädigungspotential aufweisen (Stufe 2), kann es ausreichen, wenn der Gesetzgeber die behördliche Kontrolle auf eine Ergebniskontrolle im Falle eines dokumentierten Fehlversagens des Systems beschränkt. In Bereichen, die ein hohes Schädigungspotential aufweisen, kann es hingegen erforderlich sein, den Systembetreibern vorzuschreiben, eine standardisierte Schnittstelle vorzuhalten.

Ob eine behördliche Kontrolle Betriebs- und **Geschäftsgeheimnisse** der Systembetreiber oder **Persönlichkeitsrechte** Dritter berührt, spielt nach Ansicht der DEK bei der behördlichen Kontrolle auf keiner Stufe der Kritikalitätspyramide eine Rolle. Da Aufsichtsbehörden verpflichtet sind, die im Wege der Kontrolle gewonnenen Informationen als Teil des Amtsgeheimnisses vertraulich zu behandeln, stehen diese Gesichtspunkte einer weitreichenden Befugnis zur vollständigen und detaillierten Überprüfung auf gesetzlicher Grundlage nicht entgegen.

Testergebnisse sachgerecht zu interpretieren, ist aus technischer Sicht alles andere als trivial. Insbesondere ist nicht immer eindeutig, ob sie wirklich einen Fehler eines algorithmischen Systems ans Tageslicht befördern. Das schränkt ihre Beweisfunktion ein. Es bedarf deswegen auch einer Verständigung über die Qualität und den Erkenntniswert der unterschiedlichen Testverfahren und Audits – insbesondere darüber, welcher Beweiswert ihnen in gerichtlichen Verfahren zukommt, um Betroffenenrechte durchzusetzen. Die DEK empfiehlt daher der Bundesregierung Initiativen zu unterstützen, die – ggf. nach Anwendungsbereichen differenzierte – **technisch-statistische Standards für Testverfahren und Audits entwickelt**. Dem Kompetenzzentrum Algorithmische Systeme (→ siehe dazu bereits oben 5.1.1) sollte bei diesen Bemühungen eine führende Rolle zukommen.

Use Case: Personalisierte Preise II – Ex-post Kontrolle durch Aufsichtsinstitutionen

Aufsichtsinstitutionen könnten überprüfen, ob sich algorithmische Pricing-Systeme im Online-Handel rechtskonform verhalten oder etwa geschützte Bevölkerungsgruppen (im Sinne des Allgemeinen Gleichbehandlungsgesetzes (AGG)) diskriminieren. So könnten Aufsichtsbehörden nach auffälligen Mustern in der Datenbasis und den ausgegebenen Preisen suchen, die Aufschluss über eine mögliche Diskriminierung geben können.

Dafür muss die Aufsicht nicht die (möglicherweise hochkomplexen) Regeln des zugrunde liegenden Algorithmus über eine Analyse des Codes nachvollziehen.

Eine effektive Kontrolle kann mit Hilfe statistischer Tests erfolgen. Diese analysieren, wie sich ausgegebene Preise – ceteris paribus – in Abhängigkeit von Input-Daten ändern, die mit bestimmten Bevölkerungsgruppen assoziiert werden. Gibt das System beispielsweise für Verbraucher höhere Preise aus, wenn bei deren Input-Daten nur das Geschlecht von „männlich“ auf „weiblich“ geändert wird oder korrelieren die ausgegebenen Preise mit durch das Gleichstellungsrecht geschützten Eigenschaften einzelner Bevölkerungsgruppen (etwa über sog. Proxys), lässt sich das mathematisch-statistisch ermitteln. [11]

11 Vgl. Gesellschaft für Informatik: Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. Gutachten der Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V. im Auftrag des Sachverständigenrats für Verbraucherfragen, Berlin, URL: www.svr-verbraucherfragen.de/wp-content/uploads/GI_Studie_Algorithmenregulierung.pdf [Zugriff: 07.03.2019] S. 63 ff.

5.2 Unternehmerische Selbstregulierung und Ko-Regulierung

Algorithmische Systeme flächendeckend gesetzgeberisch regulatorisch zu erfassen, ist weder möglich noch notwendig. Vielmehr können grundsätzlich auch verschiedene Modelle der Selbst- und Ko-Regulierung für bestimmte Konstellationen ergänzend adäquate Antworten liefern. Ko-Regulierung zeichnet sich dadurch aus, Regulierungswege zwischen staatlicher Regulierung und privater Selbstregulierung zu beschreiten. Prägend ist das Zusammenwirken einer staatlich-hoheitlichen Komponente und einer privat-institutionellen Komponente.

5.2.1 Selbstregulierung und -zertifizierung

Selbstregulierung in Form einer internen Überprüfung durch den Hersteller oder Betreiber des algorithmischen Systems empfiehlt sich aus Sicht der Datenethikkommission bereits für die unterste Stufe der Kritikalitätspyramide. Dies kann durch eine Selbstzertifizierung der Hersteller oder Betreiber auf der Grundlage spezifischer Standards für algorithmische Systeme unterstützt werden. Der Vorteil eines solchen Systems liegt dabei insbesondere darin, dass die Selbstzertifizierungseinrichtungen aufgrund ihrer **inhaltlichen Nähe zu spezifischen Thematiken** über das notwendige Know-how verfügen. So können Experten auch aus betroffenen Unternehmen selbst bei der Entwicklung die rechtlichen Maßstäbe und die Kontrolle ihrer Einhaltung berücksichtigen und diesen unternehmerischen Sachverstand gegebenenfalls auch institutionell in die Regulierungsmechanismen einbinden. Freilich gewährleistet eine rein interne und freiwillige Selbstkontrolle keine unabhängige Überwachung und stellt im Falle von Verstößen keine effektive Sanktionierung sicher.

Ergänzen ließe sich die Selbstregulierungsarchitektur durch ein Modell der regulierten Selbstkontrolle, das externe Standards für das Qualitäts- und Risikomanagement der Selbstkontrolle vorgibt, die sodann auch extern überwachbar sind. Ein vergleichbares System sieht die DSGVO vor. So eröffnet Art. 40 DSGVO die Möglichkeit, Generalklauseln der DSGVO zu konkretisieren und auf bestimmte, für die Adressaten der Verhaltensregeln bedeutsame Lebenssachverhalte anwendbar zu machen sowie branchenintern Mindeststandards zu setzen. Um die damit beabsichtigte Effektivität der Regulierung gewährleisten zu können, muss eine wirksame Überwachung sicherstellen, dass die genehmigten Verhaltensregeln aus Art. 40 DSGVO tatsächlich eingehalten werden. Verpflichtend müssen nicht nur die materiellen Verhaltensregeln, sondern auch die prozeduralen Vorschriften zur Überwachung, Steuerung und Sanktionierung für den Fall der Nichteinhaltung festgelegt werden.

Sofern ein Anbieter sich der freiwilligen Selbstkontrolle anschließt und das dort vereinbarte Verfahren nachweislich einhält, kann der Normgeber Privilegien bei Aufsichtsmaßnahmen gewähren. Bedingung eines solchen Vorgehens ist, dass Anbieter in Wahrnehmung ihrer unternehmerischen Verantwortung im Zusammenwirken mit einer privaten Selbstkontrolleinrichtung Verfahrensstandards entwickeln, welche die Aufsicht anerkennt. Die Einbindung zivilgesellschaftlicher Organisationen in die Erarbeitung ist dabei erforderlich, um die Bürger- und Verbraucherinteressen angemessen zu repräsentieren und berücksichtigen zu können.

5.2.2 Erarbeitung eines Verhaltenscodex

Als Teil des Konzepts regulierter Selbstregulierung ist ein **sog. Algorithmic Accountability Codex** erwägenswert, der dem Comply-or-explain-Ansatz folgt, wie er aus anderen Teilen der Rechtsordnung bekannt ist. Er könnte die Regulierungsadressaten dazu verpflichten, sich dazu zu erklären, ob und inwieweit sie den Empfehlungen des Kodex folgen oder nicht.[12] An falsche Erklärungen knüpfen sich dann Sanktionen. Auf diese Weise könnte ein zu erarbeitender Kodex Bindungswirkung entfalten, indem er Unternehmen und Behörden für die Folgen des Einsatzes algorithmischer Systeme in die Verantwortung nimmt. Dieser Kodex kann sich zum Beispiel im Kontext von sog. Corporate Digital Responsibility-Leitlinien (→ hierzu oben Teil D, 2) herausbilden oder umgekehrt auch solche Leitlinien beeinflussen. Dabei wird sich zeigen, welche Granularität für Kodizes und Leitlinien sich als praktikabel erweist bzw. für welche bereichsspezifischen ethischen Herausforderungen ein eigener Kodex sinnvoll sein kann.

Maßgeblich für eine Steuerungsfunktion eines Kodex sind die Qualität der definierten Anforderungen und die Rahmenbedingungen, also die Kontrollmöglichkeiten durch unabhängige Externe und die Sanktionsfähigkeit bei Verstößen. Einen solchen Kodex zu erarbeiten, sollte die Aufgabe eines unabhängigen paritätisch besetzten Gremiums sein. Das Gremium müsste gleichermaßen Hersteller, Betreiber, Wissenschaft und die Zivilgesellschaft einbinden. Ob die „Regierungskommission Deutscher Corporate Governance Kodex“ (*www.dcgk.de*) hierfür als Vorbild dienen kann, bleibt zu prüfen.

Darüber hinaus oder alternativ kommen bindende Erklärungen der Hersteller und Betreiber algorithmischer Systeme untereinander in Betracht.

12 Mario Martini: Juristenzeitung (JZ), 2017, S. 1017, 1022 f.

5.2.3 Gütesiegel für algorithmische Systeme

Um eine wirksame Algorithmenregulierung zu unterstützen, ist es sinnvoll, Gütesiegel für algorithmische Systeme zu etablieren. Dabei kann es sich um freiwillige oder verpflichtende Schutzzeichen handeln. Sie machen dem Nutzer transparent, inwieweit ein algorithmisches System bestimmte Anforderungen erfüllt. Zu klären ist dabei, wer die Anforderungen eines Gütesiegels bestimmt und wer dafür im Detail zuständig ist, die mit dem Gütesiegel verbundenen Anforderungen zu erfüllen und inwieweit Verstöße sanktionsbewehrt sind. Ebenso wie im Falle eines Algorithmic Accountability Codex sollte die Aufgabe, die Anforderungen eines Gütesiegels zu definieren, in den Händen einer unabhängigen, paritätisch besetzten Kommission liegen, die sich aus der Reihe der Betreiber algorithmischer Systeme, Wissenschaft und Zivilgesellschaft zusammensetzt.

5.2.4 Ansprechpartner für algorithmische Systeme in Unternehmen und Behörden

Unternehmen und Behörden, die mit kritischen algorithmischen Systemen (ab Stufe 2) arbeiten, sollten (jedenfalls ab einer bestimmten Unternehmens- bzw. Behördengröße) einen Ansprechpartner benennen müssen, der für die Kommunikation mit Behörden zur Verfügung steht und zu einer Mitwirkung verpflichtet ist. In jedem Fall muss der Ansprechpartner **spezifischen Sachverstand** haben. Er überwacht die Verwendung von algorithmischen Systemen intern und berät die Unternehmens- und Behördenleitung. Er ist in seiner Funktion unabhängig. In Anlehnung an den Datenschutzbeauftragten könnte er als Bindeglied zwischen Aufsicht, Betreiber eines algorithmischen Systems und betroffenen Personengruppen fungieren. Dies trägt zusätzlich dazu bei, in Unternehmen und Behörden für ein stärkeres Problembewusstsein und für einen erhöhten Kontrolldruck von innen heraus zu sorgen.

5.2.5 Einbindung zivilgesellschaftlicher Akteure

Um sicherzustellen, dass bei der Überprüfung von algorithmischen Systemen die Interessen der Zivilgesellschaft und der betroffenen Unternehmen angemessen Berücksichtigung finden, sollten **Beiräte** bei den sektorspezifisch zuständigen Behörden eingerichtet werden und zivilgesellschaftliche Akteure sollten auch etwa im Zusammenhang eines Kodex beteiligt werden. In diesen Beiräten sollten Vertreter zivilgesellschaftlicher Organisationen und Benannte der Unternehmen in einem ausgewogenen Verhältnis vertreten sein, um sicherzustellen, dass sowohl den Interessen betroffener Individuen und Gruppen als auch denen betroffener Unternehmen bei der Prüfung angemessen Rechnung getragen wird.

5.3 Technische Standardisierung

Normungsorganisationen wie ISO/IEC, IEEE, IETF, ITU, ETSI, W3C, CEN oder DIN, die technische Standards für Informations- und Kommunikationstechnologien setzen, können aus Sicht der DEK einen wichtigen Beitrag dazu leisten, die Anforderungen an algorithmische Systeme bereichsspezifisch zu konkretisieren. Technische Standards, die ethische und rechtliche Anforderungen berücksichtigen, können die Rechtssicherheit derjenigen Unternehmen, die die algorithmischen Systeme entwickeln und einsetzen. Zudem können sie in Einzelbereichen auch die Anforderungen an die Rechtmäßigkeit von algorithmischen Systemen handhabbar in konkrete Handlungsanweisungen übersetzen.

Die DEK sieht technische Standards grundsätzlich als ein sinnvolles Instrument zwischen „klassischer“ staatlicher Regulierung und rein privater Selbstregulierung an. Sie empfiehlt daher der Bundesregierung, in geeigneter Weise auf die Entwicklung und Verabschiedung technischer Standards hinzuwirken, die vor Risiken schützen, welche von algorithmischen Systemen ausgehen.

Die Bundesregierung sollte aus Sicht der DEK allerdings auch die **Grenzen technischer Normung** nicht aus den Augen verlieren (→ oben Teil D, 6). Technische Normen können weder die Definition klarer gesetzlicher Anforderungen an algorithmische Systeme noch die behördliche Aufsicht über den Einsatz derartiger Systeme ersetzen. An dem Grundsatz, dass gesetzliche Vorgaben umso detaillierter ausfallen müssen, je intensiver Grundrechte von B Bürgern betroffen sind, gilt es schon aus verfassungsrechtlichen Gründen festzuhalten. Konkret bedeutet das, dass zunächst der Gesetzgeber den gesetzlichen Rahmen abstecken muss – nicht Gremien zur technischen Standardsetzung. Darin manifestiert sich nicht zuletzt ein Integritätsschutz der Entscheidungsfindung, da durch die aktive Beteiligung der Vertreter von Branchen bzw. betroffenen Unternehmen neben großem technischen Sachverstand natürlich auch die Interessen dieser Unternehmen bzw. Branchen in die Formulierung der technischen Norm häufig ungefiltert einfließen.

5.4 Institutioneller Rechtsschutz (insbesondere Verbandsklagerechte)

Die in Deutschland bewährten Klagerechte von Wettbewerbern und von Wettbewerbs- und Verbraucherverbänden sind ein zentraler Baustein einer **zivilgesellschaftlichen Kontrolle** des Einatzes algorithmischer Systeme. Besonders legitimierte zivilgesellschaftliche Akteure können durch private Klagerechte die Einhaltung von Rechtsvorschriften im Bereich des Vertragsrechts und des Lauterkeitsrechts sicherstellen, ohne hierbei auf das Tätigwerden von Behörden oder die Mandatierung durch einzelne Betroffene angewiesen zu sein. Dieser zivilrechtliche Ansatz ist besonders marktnah und reaktionsschnell sowie dadurch im internationalen Vergleich erfolgreich. Verbände sind grundsätzlich politisch und administrativ unabhängig und können so eigenverantwortlich dafür eintreten, dass die Wettbewerbsordnung und das Verbraucherrecht im gemeinsamen Interesse von Verbrauchern und Unternehmen effizient vor unlauteren und verbraucherschädigenden Geschäftspraktiken geschützt wird.

Wer sich nicht an Regulierungsvorgaben hält, der erlangt im Wettbewerb unter Umständen einen – allerdings unlauteren – Vorteil. Um einen „Vorsprung durch Rechtsbruch“ zu verhindern, sollten Wettbewerbs- und Verbraucherverbände die Möglichkeit haben, solche Rechtsverletzungen abzustellen.

Zusammenfassung der wichtigsten Handlungsempfehlungen

Institutionen

55

Die DEK empfiehlt der Bundesregierung, die bestehenden Aufsichtsinstitutionen und -strukturen im Rahmen ihrer Zuständigkeit zu stärken, neu auszurichten und, wo erforderlich, auch neue Institutionen und Strukturen zu schaffen. Dabei sollten die behördlichen Aufsichtsaufgaben und Kontrollbefugnisse primär jeweils denjenigen **sektoralen Aufsichtsbehörden** zugewiesen werden, die bereits sektorspezifische Sachkompetenzen ausgebildet haben. Von großer Bedeutung ist es dabei, dass die zuständigen Behörden mit den erforderlichen finanziellen, personellen und technischen **Ressourcen** ausgestattet werden.

56

Darüber hinaus empfiehlt die DEK der Bundesregierung die Schaffung eines **bundesweiten Kompetenzzentrums Algorithmische Systeme**, welches die sektoralen Aufsichtsbehörden durch technischen und regulatorischen Sachverstand in ihrer Aufgabe unterstützt, algorithmische Systeme im Hinblick auf die Einhaltung von Recht und Gesetz zu kontrollieren.

57

Aus Sicht der DEK sollten Initiativen unterstützt werden, die – ggf. differenziert nach kritischen Anwendungsbereichen – technisch-statistische **Standards für die Qualität von Testverfahren und Audits** festlegen. Für die Überprüfbarkeit algorithmischer Systeme können derartige Testverfahren künftig eine zentrale Rolle spielen, wenn sie hinreichend aussagekräftig, verlässlich und sicher ausgestaltet sind.

58

Innovative Formen der **Ko- und Selbstregulierung** verdienen aus Sicht der DEK neben und in Ergänzung zu staatlichen Formen der Regulierung besondere Aufmerksamkeit. Die DEK empfiehlt der Bundesregierung die Prüfung verschiedener Modelle der Ko- und Selbstregulierung, die für bestimmte Konstellationen adäquate Antworten liefern können.

59

Die DEK hält es für erwägenswert, den Betreibern – nach dem Regulierungsmodell „Comply or Explain“ – die gesetzliche Pflicht aufzuerlegen, sich zu den Regeln eines **Algorithmic Accountability Codex** zu bekennen. Die Erarbeitung eines solchen bindenden Codex für die Betreiber von algorithmischen Systemen könnte dabei durch eine unabhängige, paritätisch besetzte Kommission erfolgen, die nicht unter staatlichem Einfluss stehen dürfte. Vertreter der Zivilgesellschaft sollten bei der Erarbeitung eines solches Codex in angemessener Weise beteiligt werden.

60

Auch ein spezifisches **Gütesiegel** als freiwilliges oder verpflichtendes Schutzzeichen kann Verbrauchern Orientierung über vertrauenswürdige algorithmische Systeme geben und gleichzeitig marktwirtschaftliche Anreize für Entwickler und Betreiber setzen, vertrauenswürdige Systeme zu entwickeln und zu verwenden.

61

Ähnlich wie schon heute Unternehmen ab einer bestimmten Größe einen Datenschutzbeauftragten benennen müssen, sollten nach Auffassung der DEK künftig auch solche Unternehmen und Behörden, die kritische algorithmische Systeme betreiben, einen **Ansprechpartner** benennen müssen. Er soll für die Kommunikation mit Behörden zur Verfügung stehen und zu einer Mitwirkung verpflichtet sein.

62

Um sicherzustellen, dass bei der behördlichen Überprüfung algorithmischer Systeme auch die Interessen der Zivilgesellschaft und betroffener Unternehmen angemessen berücksichtigt werden, sollten geeignete **Beiräte bei den sektoralen Aufsichtsbehörden** gebildet werden.

63

Die DEK stuft technische Standards **akkreditierter Normungsorganisationen** als ein grundsätzlich sinnvolles Instrument zwischen staatlicher Regulierung und rein privater Selbstregulierung an. Sie empfiehlt daher der Bundesregierung, in geeigneter Weise auf die Entwicklung und Verabschiedung technischer Standards hinzuwirken.

64

Die in Deutschland bewährten **Klagerechte von Wettbewerbern** und von **Wettbewerbs- und Verbraucherverbänden** sind ein zentraler Baustein für eine zivilgesellschaftliche Kontrolle des Einsatzes von algorithmischen Systemen. Besonders legitimierte zivilgesellschaftliche Akteure können durch solche privaten Klagerechte die Einhaltung von Rechtsvorschriften im Bereich des Vertragsrechts, des Lauterkeitsrechts oder des Antidiskriminierungsrechts sicherstellen, ohne hierbei auf das Tätigwerden von Behörden oder die Mandatierung durch einzelne Betroffene angewiesen zu sein.

6. Besonderes Augenmerk: Algorithmische Systeme bei Medienintermediären

6.1 Die Relevanz für den demokratischen Prozess am Beispiel sozialer Netzwerke

Längst sind soziale Netzwerke, Suchmaschinen und ähnliche Dienste aus dem Alltag vieler Menschen nicht mehr wegzudenken: Sie ermöglichen Nutzern, sich in Echtzeit über das Neueste aus den Nachrichten und dem Freundeskreis zu informieren, sind Plattformen etwa für Selbstdarstellung und Kommunikation, dienen der Unterhaltung und der wirtschaftlichen Betätigung, einschließlich der Werbung.

In der Summe entwickeln sie eine immer größere Bedeutung für die private und öffentliche Meinungsbildung. Um der Masse an Informationen Herr zu werden, nutzen die Betreiber derartiger Dienste algorithmische Systeme. Diese sollen unter anderem die Interessen, Neigungen und Überzeugungen der Nutzer erkennen, die für sie potenziell relevanten Beiträge identifizieren, ihnen ähnliche Beiträge präsentieren, um Interaktionen mit dem Netzwerk hervorzurufen, sowie illegale oder anstößige Beiträge ausfiltern. Das wirtschaftliche Ziel besteht in erster Linie darin, hohe Werbeeinnahmen zu generieren.

Abhängig von ihrer Reichweite und ihren Inhalten können Medienintermediäre einen tiefgreifenden Einfluss auf den demokratischen Prozess haben. So nutzen immer mehr Menschen soziale Netzwerke auch, um sich über Politik und Weltgeschehen zu informieren. Dabei eröffnen soziale Netzwerke den Nutzern neue Möglichkeiten der Partizipation an der Informationsgesellschaft. In diesem Sinne sind sie **Medium und Faktor für Informationen und Meinungsaustausch**.

Zugleich stellt die Konzentration der öffentlichen Debatte auf einigen wenigen privaten Plattformen aber auch eine Herausforderung für die Demokratie dar. Denn als Wirtschaftsakteure haben die privaten Betreiber sozialer Netzwerke ein Interesse daran, den Zugang zu ihrem Netzwerk und das Verhalten darauf in erster Linie nach ökonomischen Gesichtspunkten auszurichten, statt gesellschaftliche Interessen an einem vielfältigen, am Gemeinwohl orientierten Meinungsbildungsprozess in den Vordergrund zu rücken. Der Einsatz algorithmischer Systeme, die **überwiegend an ökonomischen Kriterien orientiert** sind, kann dabei negative Folgen für die Meinungsvielfalt in sozialen Netzwerken haben.

Zudem kann es durch die Nutzung von Services zur Manipulation der Meinungsbildung kommen. Dies kann einerseits unbeabsichtigt durch bestimmte Charakteristika zugrundliegender Software, wie beispielsweise Recommender Systeme, geschehen. Andererseits können diese Systeme auch bewusst von diversen Akteuren manipulativ eingesetzt werden. Bislang haben die Betreiber sozialer Netzwerke solchen demokratiegefährdenden Aktivitäten nicht hinreichend vorgebeugt. Zugleich fehlt es insbesondere mit Blick auf ihre **hohe Kritikalität** an einem staatlichen Ordnungsrahmen und an einer gesellschaftlichen Kontrolle.

Die DEK sieht perspektivisch bei Medienintermediären mit Torwächterfunktion ein hohes Gefährdungspotential für die Demokratie und dementsprechenden **Regulierungsbedarf**. Die DEK hält es für unerlässlich, dass der Gesetzgeber einen angemessenen Ordnungsrahmen für den Einsatz algorithmischer Systeme durch Medienintermediäre schafft. Zwar obliegt es nach Auffassung der DEK zunächst den Betreibern solcher Plattformen und Dienste selbst, Grundregeln für ein faires Miteinander im Meinungsbildungsprozess zu definieren und durchzusetzen. Dieses „digitale Hausrecht" hat jedoch Grenzen, insbesondere dort, wo die Integrität des demokratischen Prozesses berührt ist. Abhängig von der Marktmacht und Torwächterfunktion solcher Plattformen und Dienste bestehen im Wege der mittelbaren Drittwirkung[13] grundrechtliche Verpflichtungen an die Betreiber. Diese sollte der Gesetzgeber nach Auffassung der DEK konkretisierende Regelungen – insbesondere auch mit Blick auf den Einsatz algorithmischer Systeme durch und in Plattformen und Diensten mit Marktmacht und Torwächterfunktion – stärker als bisher einfachgesetzlich ausformen und präzisieren. Dies ist auch relevant für die von der DEK empfohlene EUVAS (→ oben 3.3).

Regulierungsbedarf besteht auch vor dem Hintergrund der Regulierungsgerechtigkeit im Vergleich zu Rundfunkanbietern. Die DEK empfiehlt der Bundesregierung zu prüfen, wie Gefahren durch besonders meinungsmächtige Anbieter begegnet werden kann. Hierzu kann ein Spektrum von Maßnahmen in Frage kommen, das sich prinzipiell von Steigerung der Transparenz bis hin zu einer Ex-ante-Kontrolle in der Form eines Lizensierungsverfahrens für demokratierelevante algorithmische Systeme erstreckt.

6.2 Vielfalt bei Medienintermediären am Beispiel sozialer Netzwerke

Die Funktionspluralität sozialer Netzwerke sowie die überwiegend hohe Kritikalität der von ihnen genutzten algorithmischen Systeme stellen den von der DEK empfohlenen Ansatz einer risikoadaptierten Regulierung für algorithmische Systeme allerdings vor besondere Herausforderungen. Für besonders zielführend hält die DEK vor diesem Hintergrund positive gesetzliche Vorgaben für soziale Netzwerke, die etwa die **Transparenz und Vielfalt des dortigen Diskurses** verbessern und die **Rechte der Nutzer** stärken.

Jedenfalls dort, wo soziale Netzwerke eine beherrschende Marktmacht haben, fordert die DEK weitergehende Maßnahmen zur **Vielfaltssicherung**, weil ausschließlich abwehrende Maßnahmen nicht ausreichen. In derartigen Netzwerken operierende algorithmische Systeme, die Auswirkungen auf die für die Demokratie konstitutive Freiheit und Vielfalt der Meinungsbildung haben, weisen bereits aufgrund ihrer Reichweite eine besonders hohe Kritikalität auf. Den Gesetzgeber trifft daher nach Auffassung der DEK eine ethische und eine verfassungsrechtliche Pflicht, zum Schutz der Demokratie eine **positive Medienordnung** für Medienintermediäre zu etablieren. Dies kann durch eine Transformation der Medienrechtsordnung geschehen.

Der Gesetzgeber muss geeignete Maßnahmen treffen, um sicherzustellen, dass im Gesamtangebot die plurale Vielfalt der Meinungen abgebildet sowie die **Ausgewogenheit, Neutralität und Tendenzfreiheit in der Informationsgesellschaft** gewährleistet ist.[14] Das gilt für meinungsmächtige Medienintermediäre mit Torwächterfunktion erst recht. Laut Bundesverfassungsgericht bedarf es zur Sicherung pluraler Vielfalt materieller, organisatorischer und Verfahrensregelungen, die an der Aufgabe der Herstellung der Kommunikationsfreiheit orientiert sind und deshalb geeignet sind zu bewirken, was Art. 5 Abs. 1 GG gewährleisten will.

13 BVerfGE 128, 226, 249 (FRAPORT); 148, 267 ff. Rn. 32 ff. (Stadionverbot).
14 Vgl. BVerfGE 136, 9, 28 m.w.N.

Vor diesem Hintergrund stehen die Gesetzgeber in den Bundesländern, bei denen die Zuständigkeit für das Medienrecht liegt, in der Pflicht, die genannten Vorgaben umzusetzen. Dasselbe gilt für den Gesetzgeber einer EU-Verordnung für Algorithmische Systeme (EUVAS) (s.o.). Schon aktuell unterfallen Medienintermediäre als sog. Video-Sharing-Plattformen (VSP) der AVMD-Richtlinie[15], weil sie nutzergenerierte Inhalte für die Allgemeinheit bereitstellen. Auch der Entwurf des Mediendienste-Staatsvertrages nimmt Medienintermediäre in seinen Anwendungsbereich auf. Die DEK begrüßt insoweit erneut die im Entwurf für einen **Medienstaatsvertrag (MStV-E)** vorgesehenen Vorgaben für die Transparenz sozialer Netzwerke als ersten Schritt in diese Richtung.

Bei der Ausgestaltung der Vorgaben haben die **Landesgesetzgeber** weitreichende Gestaltungsspielräume. Allerdings müssen sie die Entscheidung über das Regulierungsmodell selber treffen und dürfen sie nicht einer Vereinbarung der Privaten überlassen. Aus Sicht der DEK sollten Pluralitätspflichten für Medienintermediäre jedenfalls die Verpflichtung zum Einsatz solcher algorithmischer Systeme umfassen, die zumindest als zusätzliches Angebot auch einen Zugriff auf eine tendenzfreie, ausgewogene und die plurale Meinungsvielfalt abbildende Zusammenstellung von Beiträgen und Informationen verschaffen.[16]

Auf Basis dieser Überlegungen empfiehlt die DEK der Bundesregierung ferner, zu prüfen, ob es weitere Bereiche gibt, in denen unabhängig von der hier diskutierten demokratierelevanten Situation eine entsprechende Pflicht zur Statuierung von Neutralitätsgeboten und Vielfaltsvorgaben geboten erscheint. In Betracht kommt hier etwa der **Schutz Minderjähriger** vor Beeinflussung durch und über soziale Netzwerke.

6.3 Kennzeichnungspflicht für Social Bots

Der demokratische Prozess beruht im Kern auf der freien Meinungs- und Willensbildung menschlicher Akteure. Auf diversen Plattformen werden jedoch Bots, d. h. Softwareprogramme, eingesetzt, welche den **Anschein erwecken, menschliche Nutzer** zu sein. Nach Auffassung der DEK ist es hochproblematisch, wenn solche Bots dazu genutzt werden, individuelle Nutzer bzw. den öffentlichen Diskurs zu manipulieren oder gar bei anstehenden politischen Entscheidungen das Abstimmungsergebnis in eine bestimmte Richtung zu lenken. Die Vortäuschung der Menschlichkeit suggeriert zum einen fälschlicherweise, dass die verbreiteten Äußerungen das Ergebnis autonomer Reflexion und eigenständiger politischer Meinungsbildung seien. Zum anderen kann durch Automatisierung die Anzahl und Frequenz von Meinungsäußerungen massiv erhöht werden, wodurch auch die Beurteilung faktischer Mehrheitsverhältnisse von Meinungen erschwert bzw. unmöglich wird. Nach Ansicht der DEK ist hier ein regulatorisches Eingreifen erforderlich.

Vor diesem Hintergrund empfiehlt die DEK als **transparenzsteigernde Maßnahme** eine Kennzeichnungspflicht für Social Bots in sozialen Netzwerken. Schon nach allgemeinen Erwägungen empfiehlt die DEK eine solche Kennzeichnungspflicht überall dort, wo eine Verwechselungsgefahr von Social Bots mit menschlichen Gesprächspartner besteht (→ oben). Aufgrund des besonderen Gefährdungspotentials für den demokratischen Prozess hält die DEK darüber hinaus jedenfalls eine Kennzeichnungspflicht für solche Social Bots, die Einfluss auf politische Meinungsbildungsprozesse nehmen, auch unabhängig von einer konkreten Verwechslungsgefahr, für unbedingt geboten.

15 Richtlinie 2010/13/EU vom 10.3.2019 zur Koordinierung bestimmter Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über die Bereitstellung audiovisueller Mediendienste (Richtliche über audiovisuelle Mediendienste).
16 Dazu Rolf Schwartmann / Maximilian Hermann / Robin Muhlenbeck: MultiMedia und Recht (MMR), 2019 (8), S. 498, 498 ff.

6.4 Maßnahmen gegen „Fake News“

Eine Kennzeichnungspflicht für Social Bots kann der automatisierten Verbreitung sog. Fake News entgegenwirken. Darüber hinaus ist die DEK aber der Auffassung, dass das Konzept von Fake News sich **nicht als Anknüpfungspunkte für eine medienrechtliche Regulierung** eignet. Die Vorstellung eines gesetzlichen Fake News-Tatbestands, der eine objektive, trennscharfe Linie zwischen zugespitzter oder satirischer Meinungsäußerung und absichtlich-falscher Darstellung von Nachrichten zieht, scheitert an der Komplexität menschlicher Kommunikation. Zudem kann eine – typischerweise mit dem Begriff der Fake News assoziierte – Desinformation und Manipulation der öffentlichen Meinungsbildung auch durch eine selektive Darstellung wahrer Tatsachen erfolgen.

Darüber hinaus empfiehlt die DEK insbesondere dem Gesetzgeber den Betreibern sozialer Netzwerke, den Nutzern ein einfach handhabbares **Recht auf Gegendarstellung** einzuräumen, bei der die Richtigstellung einer nachgewiesen falschen Behauptung (z. B. ein erfundenes Zitat) in die „Timeline“, den „Newsfeed“ o.ä. aller Nutzer eingespeist werden muss, von denen das Netzwerk anhand vorhandener Daten rekonstruieren kann, dass sie die falsche Tatsachenbehauptung angeboten bekommen hatten.

Die DEK betont, dass der Staat keine Anreizstruktur zu einer Kollateralzensur („collateral censorship“) durch soziale Netzwerke schaffen darf. Zum Schutz vor sog. overblocking ist es daher aus Sicht der DEK erforderlich, parallel zu den den Betreibern auferlegten Pflichten den Betroffenen zeitnahe und effiziente verfahrensrechtliche Schutzmechanismen einzuräumen. Hierzu gehört nach Auffassung der DEK insbesondere ein **Recht auf ein wirksames Verfahren, um gelöschte Beiträge wieder einzustellen,** solange diese keinen Gesetzen widersprechen; eine Berufung der Netzwerke auf ihre eigenen Regeln darf als Grund für eine dauerhafte Löschung/Blockade allein nicht ausreichen. Derartige Rechte müssen nach Auffassung der DEK Nutzerinnen und Nutzer gegenüber allen sozialen Netzwerken gelten.

6.5 Transparenzpflichten für News-Aggregatoren

Soweit soziale Netzwerke algorithmische Systeme verwenden, die auch journalistisch-redaktionelle Angebote Dritter aggregieren, selektieren und allgemein zugänglich präsentieren, sollten sie Nutzern und interessierten Dritten in dem Maße Einblick in ihr technisches Verfahren der Nachrichtenauswahl und -priorisierung gewähren müssen, der nachvollziehbar macht, wie eine Empfehlung im Einzelfall zu Stande kommt. Dabei genießt das demokratische Informationsinteresse grundsätzlich Vorrang vor den Geschäftsgeheimnissen der Medienintermediäre. Solche Offenlegungspflichten sollten sich im Interesse eines fairen Meinungsbildungsprozesses und –austausches auch auf die wirtschaftlichen Verflechtungen erstrecken. Auch aus diesem Grund begrüßt die DEK die aktuellen Reformüberlegungen zum Medienstaatsvertrag (MStV-E), der für Medienintermediäre ab einer gewissen Reichweite entsprechende Transparenzpflichten vorsieht.

Zusammenfassung der wichtigsten Handlungsempfehlungen

Besonderes Augenmerk: Algorithmische Systeme bei Medienintermediären

65

Vor dem Hintergrund der besonderen Gefahren von Medienintermediären mit **Torwächterfunktion für die Demokratie** empfiehlt die DEK, auch mit Blick auf eine Einwirkung auf den EU-Gesetzgeber (→ siehe oben Empfehlung Nr. 43) zu prüfen, wie den mit einer solchen Torwächterfunktion verbundenen Gefahren begegnet werden kann. Dabei sollte ein ganzes Spektrum gefahrenabwehrender Maßnahmen erwogen werden, das bis hin zu einer Ex-ante-Kontrolle (z. B. in Form eines Lizenzierungsverfahrens) reichen kann.

66

Den nationalen Gesetzgeber trifft die verfassungsrechtliche Pflicht, die Demokratie vor den Gefahren für die freie demokratische und plurale Meinungsbildung, die von Anbietern mit Torwächterfunktion ausgehen, durch **Etablierung einer positiven Medienordnung** zu schützen. Die DEK empfiehlt, die Anbieter in diesem engen Bereich zum Einsatz solcher algorithmischer Systeme zu verpflichten, die den Nutzern zumindest als zusätzliches Angebot auch einen Zugriff auf eine tendenzfreie, ausgewogene und die plurale Meinungsvielfalt abbildende Zusammenstellung von Beiträgen und Informationen verschaffen.

67

Für alle Medienintermediäre und auch bei Anbietern ohne Torwächterfunktion oder bei geringerem Schädigungspotenzial für die demokratische Meinungsbildung sollte die Bundesregierung Maßnahmen prüfen, die den charakteristischen Gefahren des Mediensektors Rechnung tragen. Dies könnte Mechanismen zur **Transparenzsteigerung** (z. B. Einblick in technische Verfahren der Nachrichtenauswahl und -priorisierung, **Kennzeichnungspflichten für Social Bots**) und ein Recht auf Gegendarstellung in Timelines umfassen.

7. Der Einsatz algorithmischer Systeme durch staatliche Stellen

7.1 Chancen und Risiken beim Einsatz algorithmischer Systeme durch staatliche Stellen

Die Bürger erwarten zu Recht, dass der Staat **die beste verfügbare Technik nutzt**, um seine Aufgaben zu erledigen. Hierzu gehören, je nach Aufgabenbereich, auch algorithmische Systeme. Bereits heute existieren Systeme, die staatliche Stellen von repetitiven Tätigkeiten entlasten können – und die dadurch Verfahren beschleunigen und Kapazitäten für komplexe Fälle freisetzen –, die in bestimmten Konstellationen Einheitlichkeit und Qualität staatlichen Handelns verbessern oder die – etwa in Form von Chatbots oder Sprachassistenten – Bürgern den Zugang zum Recht erleichtern.

Zugleich müssen staatliche Stellen beim Einsatz von algorithmischen Systemen besonders hohe Standards wahren. Denn zum einen sind sie als Träger hoheitlicher Gewalt unmittelbar an die Grundrechte gebunden. Zum anderen kommt staatlichem Handeln generell eine **Modell- und Vorbildcharakter** für die gesamte Gesellschaft zu. Die institutionellen und Wissenskapazitäten, die der Staat zur Gewährleistung hinreichender Kontrolle der von Privaten eingesetzten algorithmischen Systeme aufbauen muss, müssen dementsprechend auch genutzt werden, um das Handeln der eigenen staatlichen Stellen anzuleiten und zu beaufsichtigen. Insbesondere dem von der DEK geforderten Kompetenzzentrums Algorithmische Systeme dürfte in diesem Zusammenhang eine Schlüsselrolle zukommen.

Der Einsatz algorithmischer Systeme durch staatliche Stellen muss **grundsätzlich als besonders sensibel** i.S.d. Kritikalitätspyramide (mindestens Stufe 3) gelten. Daher gehört aus Sicht der DEK eine umfassende Risikofolgenabschätzung zu den zwingenden Voraussetzungen jedes ethisch verantwortbaren Einsatzes algorithmischer Systeme. Darüber hinaus sind – je nach Kritikalität der staatlich genutzten Systeme – gegebenenfalls weitere der oben erörterten Instrumente zum Schutz der Bürger auch für hoheitlich genutzte algorithmische Systeme in Stellung zu bringen. Weitergehende datenschutzrechtliche Anforderungen bleiben davon ebenso wie sonstige verfassungs- und verwaltungsrechtliche Vorgaben für die Ausgestaltung der Systeme unberührt. Hinzu kommt, dass nach Auffassung der DEK in bestimmten Bereichen, in denen der Einsatz algorithmischer Systeme mit übergeordneten verfassungsrechtlichen Gütern kollidiert, die Nutzung algorithmischer Systeme unabhängig von im Einzelfall getroffenen Schutzmaßnahmen ausgeschlossen oder nur unter sehr restriktiven Bedingungen zulässig ist. Dies betrifft insbesondere den Einsatz algorithmischer Systeme für Zwecke der Rechtsetzung und der Rechtsprechung.

7.2 Algorithmische Systeme in der Rechtsetzung

Grenzen sind dem Einsatz algorithmischer Systeme im staatlichen Kontext bei der Rechtsetzung gezogen. Die DEK hält den demokratischen Prozess im Sinne einer möglichst freien Meinungs- und Willensbildung menschlicher Akteure für prinzipiell unantastbar. Automationsunterstützung in der Rechtsetzung ist daher **allenfalls für untergeordnete Hilfsaufgaben** (z. B. Aufdeckung begrifflicher Inkonsistenzen) bzw. **sehr weit von der demokratischen Willensbildung entfernte Rechtsakte** (z. B. Kataloge technischer Vorgaben in nachgelagerten Verordnungen) akzeptabel. In beiden Fällen sind höchste Anforderungen an die Qualität und Sicherheit der eingesetzten Systeme zu stellen.

Die DEK spricht sich in diesem Zusammenhang insbesondere auch gegen den Anspruch an neu erlassene Rechtsakte aus, diese müssten bereits im Hinblick auf eine mögliche künftige maschinelle Anwendung konzipiert werden; **die Technik hat auch insoweit dem Recht zu folgen, und nicht umgekehrt das Recht der Technik**. Allenfalls dann, wenn nach herkömmlichen Kriterien zur Bewertung von Gesetzgebung (Konformität mit Grundrechten und anderem höherrangigem Recht, Folgenabschätzung usw.) zwei gleichwertige Versionen denkbar sind, darf das Argument der leichteren Algorithmisierbarkeit einer Version den Ausschlag geben.

7.3 Algorithmische Systeme in der Rechtsprechung

Auch in der Rechtsprechung ist die Nutzung algorithmischer Systeme nach Auffassung der DEK **nur in Randbereichen** zulässig. Recht wird im „im Namen des Volkes" und das heißt jedenfalls im streitigen Verfahren sowie in verwaltungsgerichtlichen und in Strafverfahren stets durch menschliche Richter gesprochen. Der Befriedungseffekt eines Gerichtsverfahrens wird nicht nur durch das Urteil selbst (Ergebnisgerechtigkeit), sondern auch durch die menschliche Anhörung und Abwägung widerstreitender Interessen und insbesondere die strukturelle Abarbeitung der Tatbestands- und Rechtsfolgenseite (Verfahrensgerechtigkeit) – im Unterschied zu einer intransparenten Black-Box-Entscheidung – erreicht.

Aufgrund des oftmals hohen Vertrauens in die vermeintliche „Unfehlbarkeit" technischer Systeme („Automation Bias") sowie der geringen Bereitschaft, abweichende Entscheidungen zu treffen, insbesondere wenn dies mit zusätzlicher Argumentations- und Beweislast sowie dem Risiko eines „Fehlurteils" verbunden ist (sog. Default-Effekte), sind **auch rechtlich unverbindliche Entscheidungsvorschläge** für Urteile durch algorithmische Systeme aus Betroffenensicht in der Regel **hoch problematisch**.

Hilfreich können algorithmische Systeme dagegen – unter der Voraussetzung strenger Qualitätskontrolle und hoher Sicherheitsmaßstäbe – bei nicht unmittelbar die richterliche Entscheidung betreffenden **Vorbereitungsarbeiten** (z. B. Akten-Management, Dokumentkontrolle) sein.

Denkbar ist schließlich auch der Einsatz von Systemen, die **richterliche Entscheidungen retrospektiv analysieren**, ausschließlich der freiwilligen Nutzung durch Richter offenstehen und durch hohe Sicherheitsmaßnahmen vor einem Zugriff durch Dritte geschützt sind. Solche Systeme könnten z. B. herausarbeiten, ob und welche Entscheidungen durch externe Faktoren beeinflusst wurden, um Richtern künftig zur eigenen Verwendung Wege zur Vermeidung derartiger Verzerrungen zu unterbreiten und somit zu einer besseren und einheitlicheren Rechtsprechungspraxis beizutragen. Auch die Forschung kann ein legitimes Interesse am Zugang zu derartigen Systemen haben, wobei es hier hinreichender Sicherheitsgarantien im Einzelfall bedarf. Der Einsatz von Systemen zum Zweck der Kontrolle des Wegs zur richterlichen Entscheidungsfindung oder zum Abgleich der Spruchtätigkeit von Richtern mit externen Zielvorgaben (z. B. der durchschnittlichen Bearbeitungszeit für einen Fall) ist hingegen mit Blick auf die sachliche richterliche Unabhängigkeit unzulässig.

Im **vorgerichtlichen Bereich** (Beispiel: Geltendmachung von Fluggastrechten) oder auch im Mahnverfahren odgl. ist nach Ansicht der DEK eine vollautomatisierte Behandlung rechtlicher Ansprüche zulässig, sofern dadurch Verfahrensrechte einzelner Beteiligter gewahrt werden. Letzteres ist allerdings nicht gegeben, wenn algorithmische Systeme Korrelationen herstellen, die nicht den festgeschriebenen rechtlichen Vorgaben und Verfahrensschritten folgen. Beim jetzigen Stand der Technik kommen daher in der Regel ausschließlich auf klassischen deterministischen Algorithmen basierende Systeme in Betracht, die z. B. Entscheidungen über das Einhalten formaler (nicht wertungsoffener) Kriterien treffen. Aus systemischer Sicht drohende Kompetenzverluste werden hier durch das Freiwerden von Ressourcen für komplexe Einzelfälle ausgeglichen.

7.4 Algorithmische Systeme in der Verwaltung

In der Verwaltung ist tendenziell am ehesten Raum für den Einsatz algorithmischer Systeme. Eine verstärkte **Automatisierung behördlicher Routinefälle**, die sich unter präzise definierte Tatbestands- und Rechtsfolgevoraussetzungen subsumieren lassen, kann dabei im Sinne des Effizienzgebots (§ 10 S. 2 VwVfG) geboten sein, um Verwaltungsverfahren möglichst zweckmäßig und zügig durchzuführen. Insbesondere hier gilt, dass die Entlastung der Verwaltungsmitarbeiter von Routineaufgaben Ressourcen freisetzt. Diese können wiederum für die Bearbeitung nicht-automatisierbarer Verfahren eingesetzt werden.

Potentiale hierfür bestehen insbesondere in der **Leistungsverwaltung**. Dort können und sollten nach Auffassung der DEK algorithmische Systeme dazu genutzt werden, ein proaktives Verfahrensmanagement auszubauen, durch das bei Vorliegen aller erforderlichen Daten auf Seiten der Behörden Leistungen verstärkt antragslos gewährt. Hiervon könnten besonders bildungsferne und hilfebedürftige Menschen profitieren (vgl. die antragslose Familienbeihilfe in Österreich anlässlich der Geburt eines Kindes).

In der **Eingriffsverwaltung** ist der Einsatz algorithmischer Systeme hingegen wegen der besonderen Grundrechtsbetroffenheit vorsichtig zu handhaben. Dies gilt wie bei der gerichtlichen Nutzung nicht nur für algorithmendeterminierte Verwaltungsentscheidungen, sondern bereits dort, wo durch die Nutzung der Systeme der behördliche Entscheidungskorridor verengt wird. Allgemein sind bei der Beurteilung der Zulässigkeit des Einsatzes der Systeme die Tiefe des dadurch erfolgenden Eingriffs und die Reversibilität von Entscheidungen zu berücksichtigen. Grundsätzlich sind für die Gestaltung der Systeme die Technologien zu verwenden, die einer Kontrolle am ehesten zugänglich sind. Regelmäßig wird die Verwaltung daher in sensiblen Feldern allein auf klassischen deterministischen Algorithmen basierende Systeme verwenden dürfen. Aus demselben Grund sollte die Nutzung proprietärer Software vermieden werden.

Bei **Ermessensentscheidungen** der Exekutive und Entscheidungen mit einem Beurteilungsspielraum, die rechtliche Außenwirkung entfalten, hält es die DEK derzeit für geboten, dass Menschen die letzte Entscheidung treffen, sofern die Entscheidung nicht lediglich begünstigende Auswirkungen hat. Denkbar ist es allerdings, durch Bildung von Fallgruppen und weitere Konkretisierung das Ermessen so weit zu reduzieren, dass aus Sicht des algorithmischen Systems letztlich eine gebundene Entscheidung vorliegt. Aus Sicht der DEK bildet § 35a VwVfG die Vielzahl der hier möglichen Fallgestaltungen nicht hinreichend ab, sondern ist zu schematisch. Unter Beachtung der verfassungsrechtlich gebotenen sowie der aus Art. 22 DSGVO ableitbaren Sicherungsmechanismen sollte der Gesetzgeber daher den **Anwendungsbereich des § 35a VwVfG vorsichtig erweitern** bzw. im Fachrecht differenzierte Vorgaben für den teilweisen und vollständig automationsgestützten Erlass von Verwaltungsakten machen. Die Fortentwicklung der Regelungen zur Teil- und Voll-Automatisierung von Verwaltungsverfahren sollte mit den von der DEK empfohlenen horizontalen und sektoralen Regelungen für algorithmische Systeme (→ oben) erfolgen.

7.5 Algorithmische Systeme im Sicherheitsrecht

Besonders kritisch wird in der Öffentlichkeit der Einsatz algorithmischer Systeme durch die Sicherheitsbehörden diskutiert. Da administrative Maßnahmen in diesem Bereich besonders tief in Grundrechte eingreifen können, ist der Einsatz algorithmischer Systeme tendenziell **restriktiv** zu handhaben.

Werden algorithmische Systeme zur Vorhersage von Straftaten oder Gefährdungslagen genutzt (sog. **Predictive Policing**), ist zu berücksichtigen, dass auch solche Systeme, die unmittelbar keine personenbezogenen Daten nutzen, grundrechtsrelevante Effekte haben können. Dies ist insbesondere dann der Fall, wenn durch ggf. allzu detaillierte Ortsangaben ein Personenbezug (wieder-)hergestellt werden kann. Ferner kann es durch sog. lagebezogene Risikoprognosen zur übermäßigen Kontrolle bestimmter als sog. Hot Spots identifizierter Nachbarschaften und dadurch zu ethnischen oder sozialen Profilierungen dort ansässiger Bevölkerungsgruppen kommen. Ebenfalls können derartige Maßnahmen Kriminalitätsverlagerungs- und Verdrängungseffekte auslösen. Die DEK empfiehlt daher, die Sicherheitsbehörden für derartige Effekte zu sensibilisieren und in die Vorhersagesysteme Randomisierungen einzubauen, um entsprechende Effekte und sonstige systembedingte Verzerrungen zu reduzieren; zudem muss sichergestellt werden, dass den Sicherheitsbehörden eine menschliche Prüfung weiterer Fälle als der vom System ausgewählten Risikofälle stets möglich bleibt (vgl. § 88 AO). Die Sicherheitsbehörden dürfen zudem nicht allein auf der Basis lagebezogener Prognosen weitergehende ermessensbasierte Eingriffsmaßnahmen anordnen.

Soweit **personenbezogene Risikoprognosen** im Sicherheitsbereich rechtlich zulässig sind, dürfen solche Prognosen nicht vollautomatisiert erstellt werden, sofern sich an die Erstellung der Prognose negative Rechtsfolgen für den Betroffenen knüpfen. Aufgrund des Risikos eines „Automation Bias“ schon bei algorithmenbasierten Entscheidungen ist zudem die Unterstützung menschlicher Entscheidungsträger durch algorithmische Systeme bei derartigen Profilierungen allenfalls in sehr engen Grenzen zulässig.

7.6 Transparenzanforderungen beim Einsatz algorithmischer Systeme durch staatliche Akteure

Staatliche Entscheidungen, die unter Nutzung algorithmischer Systeme zu Stande kommen, müssen **transparent und begründbar** bleiben. Dies ist aufgrund der Grundrechtsbindung und der Notwendigkeit einer demokratischen Rückbindung aller hoheitlichen Gewalt im staatlichen Bereich tendenziell noch wichtiger als im privaten Sektor. Für staatliche Stellen gelten daher nicht nur die allgemeinen Transparenzanforderungen (→ oben). Vielmehr müssen sich staatliche Stellen darüber hinaus in besonderem Maße um Offenheit bemühen.

Die DEK weist darauf hin, dass hoheitliche algorithmische Systeme vielfach bereits in den Anwendungsbereich der bestehenden Informationsfreiheits- bzw. Transparenzgesetze fallen. Darüber hinaus begrüßt die DEK das im Rahmen der 36. Konferenz der Informationsfreiheitsbeauftragten in Deutschland verabschiedete Positionspapier zur „Transparenz der Verwaltung beim Einsatz von Algorithmen“, wonach öffentliche Stellen über aussagekräftige, umfassende und allgemein verständliche Informationen bezüglich der eigenen Datenverarbeitungen verfügen müssen und diese, soweit rechtlich möglich, veröffentlichen sollten, einschließlich Informationen (i) zu den Datenkategorien der Ein- und Ausgabedaten des Verfahrens, (ii) zur darin enthaltenen Logik, insbesondere zu den verwendeten Berechnungsformeln einschließlich der Gewichtung der Eingabedaten, zum zugrundeliegenden Fachwissen und zur individuellen Konfiguration durch die Anwendenden sowie (iii) zur Tragweite der darauf basierenden Entscheidungen sowie zu den möglichen Auswirkungen der Verfahren.[17]

17 Positionspapier im Rahmen der 36. Konferenz der Informationsfreiheitsbeauftragten in Deutschland – „Transparenz der Verwaltung beim Einsatz von Algorithmen für gelebten Grundrechtsschutz unabdingbar“, Ulm, 16. Oktober 2018 (abrufbar unter: *https://www.datenschutzzentrum.de/uploads/informationsfreiheit/2018_Positionspapier-Transparenz-von-Algorithmen.pdf*).

Bei der gesetzgeberischen Konkretisierung entsprechender Transparenzpflichten bzw. Informationszugangspflichten gibt die DEK zudem zu bedenken, dass unzureichende Vorgaben zur Transparenz zu Misstrauen in die Systeme führen kann, was sich in erhöhten Anfechtungsraten niederschlagen kann, die den durch den Einsatz algorithmischer Systeme intendierten Effizienzgewinn konterkarieren können. Aus diesem Grund hält es die DEK schließlich allenfalls in wenigen Fällen für vertretbar, den Informationszugang zu hoheitlichen algorithmischen Systemen unter Verweis auf ein Manipulationsrisiko oder auf den Schutz von Geschäftsgeheimnissen pauschal auszuschließen. Im Regelfall ist daher eine Abwägung vorzunehmen.

Informationen über die generelle Funktionsweise des Systems offenzulegen, reicht beim Einsatz algorithmischer Systeme durch Hoheitsträger nicht in jedem Fall aus. Regelmäßig müssen hoheitliche Entscheidungen den Betroffenen gegenüber auch begründet werden, d. h. es sind die **„wesentlichen tatsächlichen und rechtlichen Gründe“**, die die Entscheidung im Einzelfall motiviert haben, anzugeben (vgl. § 39 Abs.1 S.2 VwVfG). Wo eine solche einzelfallbezogene Erläuterung verfassungs- bzw. einfachrechtlich geboten, aber aufgrund der technischen Komplexität des Systems nicht bzw. nicht in einer Art und Weise möglich ist, die im Zuge eines behördlichen Beanstandungsverfahren oder vor Gericht eine effektive Überprüfung der Tragfähigkeit der Begründung erlaubt, muss der Einsatz algorithmischer Systeme ausscheiden. Im Übrigen ist der Staat aus Sicht der DEK dazu aufgefordert, in der Verwaltung und an den Gerichten hinreichende **Expertise** aufzubauen, die eine entsprechende Kontrolle der systeminternen Entscheidungsprozesse gewährleisten kann.

Die DEK weist darauf hin, dass die Transparenz staatlichen Handelns auch dadurch beeinträchtigt werden kann, dass sich der Staat bei der Erfüllung seiner Tätigkeiten proprietärer Software (sog. Closed Source-Software) von privaten Anbietern bedient. Allgemein erschwert proprietäre Software Änderungen und Anpassungen durch den Benutzer, was ein Abhängigkeitsverhältnis entstehen lässt. Zudem führt die Nutzung proprietärer Software zu einem Mangel an Transparenz und kann damit die gesellschaftliche Akzeptanz der Systeme gefährden. Insbesondere in grundrechtssensiblen Bereichen wie dem Sicherheitsrecht sollte daher nach Möglichkeit auf proprietäre Software verzichtet werden. Stattdessen sollten staatliche Stellen auf **Open-Source-Lösungen** zurückgreifen oder – idealerweise in Form interdisziplinär besetzter Entwicklerteams – eigene Systeme entwickeln. Wo dies nicht praktikabel ist, empfiehlt die DEK der Bundesregierung, Anpassungen im Vergaberecht zu prüfen, die die gerade beschriebenen negativen Effekte proprietärer Software minimieren. Wo nicht zu befürchten ist, dass durch Transparenz die Effektivität des Systems leidet, also Ausnutzungseffekte ausgeschlossen werden können, sollte die Software-Entwicklung in einem offenen und konsultativen Prozess unter Einbeziehung zivilgesellschaftlicher Akteure erfolgen.

7.7 Das Risiko eines automatisierten Totalvollzugs

Die DEK verwehrt sich zwar dagegen, dass es aus ethischer Sicht ein generelles Recht auf Freiheit zur Nichtbefolgung von Normen gebe. Allerdings bestehen gegen einen automatisierten Totalvollzug des Rechts eine Reihe ethischer Bedenken. So können sich Bürger durch eine perfektionierte Vollzugspraxis unter einen Generalverdacht gestellt sehen, unter dem der allgemeine Normbefolgungswille leitet. Ferner besteht beim automatisierten Vollzug die Gefahr, dass die Komplexität der Lebenswirklichkeit nicht hinreichend abgebildet und insbesondere unvorhergesehenen Ausnahmesituationen (Beispiel: Geschwindigkeitsüberschreitung bei Privattransport eines Schwerverletzten ins Krankenhaus) nicht genügend Rechnung getragen wird. Schließlich sind viele Gesetze ursprünglich nicht für einen Totalvollzug erlassen worden. Zu fordern ist daher regelmäßig ein technisches Design, bei dem der Mensch im Einzelfall den technischen Vollzug außer Kraft setzen kann. Zudem stellt sich jede Maßnahme der Rechtsdurchsetzung als eigener staatlicher Eingriff dar und hat sich als solcher am **Verhältnismäßigkeitsprinzip** zu orientieren.

Zusammenfassung der wichtigsten Handlungsempfehlungen

Der Einsatz von algorithmischen Systemen durch staatliche Stellen

68

Der Staat ist im Interesse seiner Bürger zur Nutzung der besten verfügbaren Technik – einschließlich algorithmischer Systeme – verpflichtet, muss dabei jedoch im Lichte seiner Grundrechtsbindung sowie der Vorbildfunktion allen staatlichen Handelns besondere Sorgfalt walten lassen. Der Einsatz algorithmischer Systeme durch Hoheitsträger ist daher **im Allgemeinen als besonders sensibel im Sinne des Kritikalitätsmodells** einzustufen und erfordert mindestens eine umfassende Risikofolgenabschätzung.

69

Aufgaben in der **Rechtsetzung** und der **Rechtsprechung** dürfen algorithmischen Systemen allenfalls in Randbereichen übertragen werden. Insbesondere dürfen algorithmische Systeme nicht genutzt werden, um die freie Willensbildung im demokratischen Prozess und die sachliche Unabhängigkeit der Gerichte zu unterminieren. Große Potenziale für den Einsatz algorithmischer Systeme bestehen hingegen in der **Verwaltung**, vor allem in der Leistungsverwaltung. Um dem Rechnung zu tragen, sollte der Gesetzgeber verstärkt teil- und vollautomatisierte Verwaltungsverfahren zulassen. Dazu bedarf es auch einer vorsichtigen Fortentwicklung des zu engen § 35a VwVfG sowie der entsprechenden einfachrechtlichen Normen. Bei alledem gilt es, hinreichende Schutzmaßnahmen für die Bürger vorzusehen.

70

Staatliche Entscheidungen, die unter Nutzung algorithmischer Systeme zustande kommen, müssen **transparent und begründbar** bleiben. Dazu bedarf es ggf. Klarstellungen bzw. Erweiterungen der bestehenden Informationsfreiheits- und Transparenzgesetze. Ferner entbindet der Einsatz algorithmischer Systeme nicht vom Grundsatz, dass hoheitliche Entscheidungen regelmäßig im Einzelfall begründet werden müssen; im Gegenteil kann dieser Grundsatz dem Einsatz allzu komplexer algorithmischer Systeme Grenzen setzen. Schließlich trägt die Nutzung von Open-Source-Lösungen wesentlich zur Transparenz staatlichen Handelns bei und sollte daher verstärkt angestrebt werden.

71

Zwar ist aus ethischer Sicht ein generelles Recht auf Freiheit zur Nichtbefolgung von Normen nicht anzuerkennen. Gleichzeitig wirft ein automatisierter Totalvollzug des Rechts eine Reihe ethischer Bedenken auf. Daher ist regelmäßig ein technisches Design zu fordern, bei dem der Mensch im Einzelfall den **technischen Vollzug** außer Kraft setzen kann. Ferner muss stets die Verhältnismäßigkeit zwischen der potenziellen Normübertretung und der automatisierten (ggf. präventiven) Vollzugsmaßnahme gewahrt sein.

8. Haftung für algorithmische Systeme

8.1 Bedeutung

Strafrechtliche Verantwortlichkeit, Verwaltungssanktionen oder Haftung auf Schadensersatz sind unverzichtbarer Bestandteil eines ethisch vertretbaren Ordnungsrahmens, auch und gerade für algorithmische Systeme und andere digitale Technologien. Die DEK unterstreicht dabei aus ethischer Sicht insbesondere die Rolle des Schadensersatzrechts, welches sowohl der Kompensation als auch der Prävention von Schäden dient und damit ganz maßgeblich zu einem **grundrechtskonformen Rechtsgüterschutz** beiträgt.

Aus ethischer Sicht sind an ein Haftungssystem, welches neuen digitalen Technologien gerecht werden soll, u. a. die folgenden Anforderungen zu stellen:

a) Ausreichende **Kompensation** für Geschädigte, insbesondere bei besonders grundrechtsrelevanten Rechtsgütern und soweit Kompensation in einer vergleichbaren Situation bei Einsatz von Menschen oder herkömmlicher Technologie geschuldet wäre;

b) Setzen der richtigen **Verhaltensanreize**, indem Schäden von denjenigen Akteuren getragen werden, welche die Schäden durch vermeidbares und unerwünschtes Verhalten verursacht haben oder aus deren Sphäre das betreffende Risiko stammt;

c) **Fairness**, indem diejenigen Akteure für Schäden aufkommen, welche das System etwa in den Verkehr gebracht haben oder welche die Kontrolle über das System ausüben und aus seinem Einsatz den Nutzen ziehen;

d) **Effizienz**, indem Kosten von denjenigen Akteuren getragen (internalisiert) werden, die diese Kosten mit dem geringsten Aufwand vermeiden oder versichern können.

8.2 Schäden durch den Einsatz algorithmischer Systeme

8.2.1 Haftung der „Elektronischen Person“?

Die DEK **rät ausdrücklich davon ab**, Robotern bzw. sog. autonomen Systemen Rechtspersönlichkeit zu verleihen (oft unter dem Stichwort **„E-Person“** diskutiert) mit dem Ziel, die Systeme selbst haften zu lassen (z. B. ein autonom fahrendes Fahrzeug ohne Halter, das sich als Mobilitätsdienstleistung „selbst betreibt“). Eine solche Maßnahme wäre nicht geeignet, Verantwortlichkeit und Haftung für Schäden denjenigen Akteuren zuzuweisen, welche den Einsatz des Systems zu verantworten haben und letztlich von diesem Einsatz ökonomisch profitieren. Vielmehr könnte die Maßnahme im Gegenteil dazu genutzt werden, sich der Verantwortlichkeit zu entziehen. Durch die Rechtspersönlichkeit von Maschinen als eines neuen Typs juristischer Person ließe sich kein wünschenswertes Ergebnis erzielen, das nicht zwangloser auf andere Weise zu erzielen wäre (z. B. mithilfe des Gesellschaftsrechts). Autonom agierende Maschinen gar analog zu natürlichen Personen zu behandeln, wäre aus der Sicht der DEK eine gefährliche Verirrung.

8.2.2 Gehilfenhaftung für „Autonome“ Systeme

Die DEK hält es allerdings für geboten, eine Zurechnung entsprechend den Regelungen über die Haftung für **Gehilfen** (vgl. insbes. § 278 BGB) bei sog. autonomen Systemen vorzunehmen. Ein Akteur, der sich zur Erweiterung seines Aktionsradius eines solchen Systems bedient (z. B. ein Krankenhaus bedient sich eines chirurgischen Roboters), sollte sich im Falle einer Fehlfunktion nicht exkulpieren können, da auch ein Akteur, der sich eines menschlichen Erfüllungsgehilfen (z. B. eines menschlichen Chirurgen) bedient, für das – als Verhalten des Akteurs selbst gedacht – schuldhafte Fehlverhalten dieses Gehilfen haftet. Dies erlangt besondere Bedeutung bei der **Haftung für algorithmische Systeme**, wo anderenfalls leicht Haftungslücken entstehen, wenn kein Sorgfaltspflichtverstoß der Hinterperson bei Einsatz und Überwachung des algorithmischen Systems nachgewiesen werden kann.

Beispiele 18
Ein chirurgischer Roboter in einem Krankenhaus verursacht einen zu langen Operationsschnitt mit Komplikationen. Oder: Durch ein algorithmisches System wird die Kreditwürdigkeit eines Bankkunden falsch abgeleitet und dieser kann das einmalig günstige Angebot einer Immobilie nicht annehmen.

Dabei mag es vereinzelt schwierig sein, ein für Autonome Systeme adäquates Pendant zum „Sorgfaltsmaßstab" zu ermitteln, v.a. sobald die Fähigkeiten einer Maschine diejenigen eines Menschen übersteigen. In der Mehrzahl der Fälle aber werden Fehlfunktionen von Normalfunktionen zu unterscheiden sein und daher kann dies nicht generell gegen die Haftung des Betreibers angeführt werden. Der Maßstab muss dann durch am Markt verfügbare vergleichbare Systeme bestimmt werden, wobei die Frage, der Einsatz welcher Technologie dem Betreiber zugemutet werden konnte, nach allgemeinen Grundsätzen zu entscheiden ist (z. B. unterscheidet sich insofern die Frage, welche Qualität von chirurgischem Roboter einzusetzen war, nicht von der Frage, welche Qualität von Röntgengerät einzusetzen war).

8.2.3 Gefährdungshaftung

Dass die Regeln der klassischen Verschuldenshaftung nicht immer ausreichen, um die rechtlichen Probleme, die bei gefährlichen Produkten auftreten, zu lösen, ist grundsätzlich bekannt. Die Rechtsordnung hat auf diese Herausforderung bislang eine Reihe von Antworten gefunden. Dazu gehören insbesondere:

- **Modifikation der Verschuldenshaftung** (z. B. durch Anpassungen des Sorgfaltsmaßstabs und diverse Beweiserleichterungen bis hin zur Beweislastumkehr);
- verschiedene Tatbestände der **Gefährdungshaftung** (d. h. für Anlagen und Tätigkeiten, die typischerweise Schäden verursachen, aufgrund ihres gesamtgesellschaftlichen Nutzens aber nicht verboten werden sollen); und
- die **Produkthaftung** nach dem ProdHaftG; diese stellt sich dabei als spezielle Form der verschuldensunabhängigen Haftung dar, die sich von der Gefährdungshaftung dadurch unterscheidet, dass sie u. a. einen Fehler des Produkts voraussetzt und sie dadurch der Verschuldenshaftung etwas angenähert wird.

Es muss sichergestellt sein, dass diese Antworten auch bei gefährlichen digitalen Anwendungen zu einer rechtssicheren Kompensation von Schadensereignissen führen.

Gegenwärtig bestehen bei digitalen Anwendungen **Rechtsunsicherheiten und Haftungslücken**, die vor allem aus der Unvorhersehbarkeit von Schadensereignissen, u. a. beim Inverkehrbringen – und damit gegebenenfalls einem Versagen der klassischen Verschuldenshaftung – resultieren sowie daraus, dass durch das Zusammenwirken verschiedener Akteure und verschiedener Anwendungen in aller Regel kaum nachvollzogen werden kann, wo ein Fehler aufgetreten ist und/oder wo die Fehlerursache gesetzt wurde. Auch der offene und dynamische Charakter digitaler Ökosysteme und die enge funktionale Verknüpfung von Produkten, digitalen Inhalten und digitalen Dienstleistungen stellen eine Herausforderung dar. Diese Rechtsunsicherheiten sind aus Sicht von Unternehmen wie Verbrauchern **Hindernisse für Innovationen und für die Akzeptanz neuer Technologien**. Wenn Schadensereignisse regelmäßig nicht zugeordnet und kompensiert werden können, kann die durch Haftungsbestimmungen intendierte Marktwirkung nicht erzeugt werden. Um einen angemessenen Interessenausgleich herzustellen, muss der Gesetzgeber Transparenz und Verantwortung schaffen. Nur wenn die Verantwortlichkeiten geklärt sind, ist auch eine praxisgerechte Versicherbarkeit von Schäden möglich.

Die DEK kann den komplexen rechtstechnischen Fragen nach der richtigen Verortung einer haftungsrechtlichen Lösung nicht vorgreifen, zumal teilweise erst Chancen auf eine Lösung auf europäischer Ebene ausgelotet werden sollten. Aus ethischer Sicht ist entscheidend, dass **Rechtsklarheit und Rechtssicherheit, insbesondere in Bezug auf die oben beschriebenen Haftungsgrundsätze,** geschaffen wird. Dass neben einer sachgerechten Anpassung der Produkthaftungsrichtlinie (→ dazu sogleich) auch punktuelle Modifikationen der Verschuldenshaftung und/oder neue Tatbestände der Gefährdungshaftung erforderlich sein können, erscheint nach derzeitigem Stand der Diskussion jedoch sehr wahrscheinlich.

Im Gesetzgebungsprozess wird dabei zunächst zu klären sein, **für welche Produkte, digitalen Inhalte und digitalen Dienstleistungen** welches Haftungsregime sachgerecht und wie dieses konkret auszugestalten ist, wobei es wesentlich wiederum auf die Kritikalität (→ oben) des betreffenden Systems ankommen wird, aber auch auf weitere, speziell im Haftungskontext relevante Kriterien. So kann eine Gefährdungshaftung (nach dem Modell etwa der Kfz-Halterhaftung) bei Geräten, deren Betriebsrisiko ähnlich unkontrollierbar ist und zu Schäden an Leib und Leben führen kann, durchaus angemessen sein. Dabei muss immer die Frage nach der Versicherbarkeit bzw. einer allfälligen Pflichtversicherung eine Rolle spielen. Stets wäre zugleich mitzuentscheiden, **für welche Schäden** gehaftet werden soll (z. B. Personen- und Sachschäden, Datenverlust, reine Vermögensschäden, immaterielle Schäden).

Schließlich wird jeweils zu entscheiden sein, wer unter Berücksichtigung der oben beschriebenen Haftungsgrundsätze der richtige **Adressat der Haftung** ist. Dabei zeichnen sich vor allem drei mögliche Haftungsadressaten ab, von denen gegebenenfalls auch jeweils zwei als Gesamtschuldner haften könnten:

- der individuelle **Halter** des Systems (d. h. der Eigentümer oder derjenige, der in einer ähnlichen Position das System für seine eigenen Zwecke einsetzt);
- der **Hersteller** des Systems;
- der **Betreiber** des Systems (d. h. je nachdem, wer das höhere Maß an Kontrolle über den Systembetrieb ausübt, der individuelle Halter als Frontend-Betreiber oder aber ein Backend-Betreiber, der mit dem Hersteller identisch sein kann, aber nicht sein muss).[18]

Eine Bestimmung des Adressaten und der Art der Haftung ist dabei stets abhängig von der konkreten Art des vernetzten oder Autonomen Systems und der Identifikation der konkreten Haftungssphären.

8.2.4 Produktsicherheit und Produkthaftung

Insgesamt ist derzeit ein Paradigmenwechsel vom einmaligen Inverkehrbringen eines Produkts hin zum Inverkehrbringen eines Produkts und zusätzlicher, fortwährender Leistungserbringung für das Produkt zu verzeichnen. Daher kommt der laufenden **Produktbeobachtung** und **Produktpflege** eine gesteigerte Bedeutung zu. IT-Sicherheits- und Datenschutzstandards müssen nicht nur erfüllt sein, wenn ein Produkt das Werktor verlässt, sondern dürfen auch bei späteren Software-Updates nicht verloren gehen. Umgekehrt sollte den Hersteller bei später auftretenden Sicherheitslücken – entsprechend der Regelungen in den Richtlinien zu digitalen Inhalten und digitalen Dienstleistungen sowie zum Warenhandel – im Rahmen der berechtigten Verbrauchererwartungen zur Nutzungsdauer eine Pflicht zu **Sicherheitsupdates** treffen.

18 Zum Haftungskonzept einer derart differenzierten Betreiberhaftung in digitalen Ökosystemen siehe den Bericht „Liability for Artificial Intelligence and other emerging digital technologies“ der von der Europäischen Kommission eingesetzten Expert Group on Liability and New Technologies (New Technologies Formation), vsl. Oktober 2019 (im Erscheinen), Nr. [11], S. 41 ff.

Beispiel 19
Für eine intelligente Hausanlage werden keine Sicherheitsupdates angeboten, weshalb es nach einem Cyberangriff zu einem Wohnungseinbruch kommt.

Die aus den 1980er Jahren stammende Produkthaftungsrichtlinie konnte die Charakteristiken vernetzter, hybrider und autonomer Produkte noch nicht einbeziehen. Die DEK empfiehlt der Bundesregierung, bei der **Evaluierung und Überarbeitung der Produkthaftungsrichtlinie** auf europäischer Ebene auf rechtssichere und rechtsklare Regelungen insbesondere für folgende Aspekte zu dringen:

a) das Unterfallen digitaler Inhalte und digitaler Dienstleistungen, wie etwa auch algorithmische Systeme, unter den Produktbegriff;

b) die Haftung für Produktfehler, die erst nach dem Inverkehrbringen infolge sich selbst verändernder Software, erfolgter oder pflichtwidrig unterlassener Updates, oder produktspezifischer Daten auftreten;

c) die Haftung für Verletzungen der Produktbeobachtungspflicht;

d) die Einbeziehung typischerweise von digitaler Produktsicherheit betroffener Rechtsgüter, insbesondere die Verletzung des informationellen Selbstbestimmungsrechts, im Rahmen von Schadensersatzregelungen;

e) die Anpassung der Einwendung des Entwicklungsrisikos.

8.3 Bedarf nach einer Neubewertung des Haftungsrechts

Digitale Ökosysteme werfen eine Vielzahl weiterer Probleme im Zusammenhang mit Haftung und Verantwortlichkeit auf. So besteht teilweise eine Haftungslücke im geltenden Deliktsrecht bei **Schäden an Daten und digitalen Gütern**, soweit weder ein anerkanntes absolut geschütztes Rechtsgut verletzt ist (z. B. Eigentum am Speichermedium) noch ein existierendes Schutzgesetz verletzt wurde (z. B. Normen des Strafrechts) noch die Voraussetzungen vorsätzlicher sittenwidriger Schädigung vorliegen. Neue digitale Technologien gehen auch vielfach mit der **opportunistischen Nutzung fremder Infrastrukturen** einher (z. B. indem von privaten IoT-Geräten generierte Sensordaten von Dritten systematisch gesammelt und verwertet werden, oder auch durch unmittelbare Nutzung von Rechenkapazität oder Sendefunktionen), was schwierige Haftungsfragen aufwerfen kann. In stärker vertragsrechtlich geprägten Zusammenhängen können immense Schäden – insbesondere zulasten von Verbrauchern – dadurch verursacht werden, dass die **Nutzbarkeit von hochwertigen Gütern** (Immobilien, Maschinen, Kraftfahrzeuge usw.) immer mehr von der langfristigen Erbringung digitaler Dienste abhängig ist (Software-Updates, Nutzerkonten u. a.) und die Erbringung dieser Dienste nicht gesichert ist bzw. sogar gezielt unterbrochen werden kann, um Einzelne unter Druck zu setzen (**Electronic Repossession**).

Auch sind digitale Ökosysteme teilweise durch das Zusammenwirken zahlreicher Komponenten und Betreiber gekennzeichnet, wobei es für den Geschädigten vielfach unverhältnismäßig schwierig ist, nachzuweisen, **welcher von mehreren potenziellen Schädigern** (z. B. Hardware-Lieferant, Lieferanten mehrerer Software-Komponenten, Lieferant von Daten-Feeds oder Netzwerkbetreiber) einen Schaden verursacht hat. Andererseits schaffen digitale Technologien nicht nur neue Intransparenzen in Bezug auf die Schadensverursachung, sondern können umgekehrt auch dazu beitragen, Kausalverläufe in nie da gewesener Weise zu dokumentieren. Es stellt sich daher die Frage, welchen Akteur welche Verpflichtung trifft, durch **Logging von Daten** bereits ex ante zur Aufklärung der Schadensverursachung beizutragen und wem die tatsächlich durch Logging aufgezeichneten Daten im Schadensfall offenzulegen sind.

Die DEK empfiehlt daher der Bundesregierung insgesamt, zu prüfen, inwieweit das geltende Haftungsrecht den **Herausforderungen digitaler Ökosysteme** gewachsen ist oder einer Überarbeitung bedarf. Dabei ist vorrangig eine Lösung auf europäischer Ebene anzustreben. Die DEK warnt in diesem Zusammenhang vor einer Tendenz, den Blick einseitig auf bestimmte technologische Merkmale, insbesondere das Merkmal maschinellen Lernens, zu richten. Während maschinelles Lernen bestimmte zusätzliche Gefahren schafft und bestimmte zusätzliche Zurechnungsprobleme mit sich bringt, sind die meisten Herausforderungen für das Haftungsrecht durch andere Faktoren (z. B. Unkörperlichkeit, Zusammenwirken vieler Komponenten, Vernetzung, Dezentralisierung) bedingt.

Zusammenfassung der wichtigsten Handlungsempfehlungen

Haftung für algorithmische Systeme

72

Neben strafrechtlicher Verantwortlichkeit und Verwaltungssanktionen ist auch die Haftung auf Schadensersatz unverzichtbarer Bestandteil eines ethisch vertretbaren Ordnungsrahmens. Es ist bereits jetzt erkennbar, dass algorithmische Systeme – u. a. aufgrund der Komplexität und Dynamik der Systeme sowie aufgrund ihrer wachsenden „Autonomie" – das bestehende Haftungsrecht vor Herausforderungen stellen. Die DEK empfiehlt daher eine umfassende Prüfung und, soweit erforderlich, **Anpassung des geltenden Haftungsrechts**. Der Blick sollte sich dabei nicht allein auf bestimmte technologische Merkmale – wie etwa auf das Merkmal Maschinellen Lernens oder Künstlicher Intelligenz – verengen.

73

Der Gedanke, algorithmischen Systemen hoher Autonomie künftig Rechtspersönlichkeit zuzuerkennen und sie selbst für Schäden haften zu lassen **(„elektronische Person")**, sollte **nicht weiterverfolgt** werden. Soweit dieser Gedanke auf eine Analogie zwischen Mensch und Maschine gestützt wird, ist er schon ethisch nicht vertretbar, und soweit es schlicht um die Anerkennung einer neuen Gesellschaftsform im Sinne des Gesellschaftsrechts geht, löst er keine Probleme.

74

Dagegen ist es geboten, für den Einsatz sog. autonomer Systeme – abhängig von der Natur der dem System übertragenen Aufgaben – auch eine Zurechnung schädigender Vorgänge entsprechend den Regelungen über die Haftung für **Gehilfen** (vgl. insbes. § 278 BGB) vorzunehmen. Beispielsweise sollte eine Bank, die sich für die Prüfung der Kreditwürdigkeit eines autonomen Systems bedient, gegenüber ihrem Kunden mindestens in gleichem Maße haften, wie wenn sie sich eines menschlichen Mitarbeiters bedient hätte.

75

Daneben erscheint es nach derzeitigem Stand der Diskussion sehr wahrscheinlich, dass zusätzlich zu einer sachgerechten Anpassung der aus den 1980er Jahren stammenden **Produkthaftungsrichtlinie** und Verknüpfung mit neuen Standards der Produktsicherheit auch punktuelle Modifikationen der **Verschuldenshaftung** und/oder neue Tatbestände der **Gefährdungshaftung** erforderlich sein werden. Dabei wird jeweils zu klären sein, für welche Produkte, digitalen Inhalte und digitalen Dienstleistungen welches Haftungsregime sachgerecht und wie dieses konkret auszugestalten ist, wobei es wiederum wesentlich u. a. auf die Kritikalität des betreffenden algorithmischen Systems ankommen wird. Dabei sollten auch innovative Haftungskonzepte, wie sie derzeit auf europäischer Ebene entwickelt werden, in Betracht gezogen werden.

Teil G

Für einen europäischen Weg

Die Fülle an Fragen, die sich der Datenethikkommission gestellt haben und deren Diskussion jeweils wieder neue Fragen aufwarf, lässt deutlich werden, dass dieses Gutachten lediglich einen weiteren Grundstein für einen **Zukunftsdiskurs über Ethik, Recht und Technologie** legen kann, der immer weiter und auf breiter Basis geführt werden muss. Dieser Diskurs muss von vorneherein ein interdisziplinärer sein und ein breites Spektrum an Wissenschaften ebenso wie eine Vielfalt von Vertretern aus Wirtschaft, Zivilgesellschaft und Politik umfassen. Angesichts des hohen ökonomischen Drucks und der Geschwindigkeit des technischen Wandels müssen die Ergebnisse dieses Diskurses auf unterschiedlichen Ebenen kontinuierlich von allen beteiligten Akteuren umgesetzt werden, damit eine wertefundierte Gestaltung der technologisch geprägten Zukunft gewährleistet werden kann.

In Anbetracht des Umstandes, dass der Transfer von Daten und die Anwendung algorithmischer Systeme vor nationalen Grenzen keinen Halt machen, kann eine vorausschauende Erörterung ethischer und rechtlicher Fragen zu Daten und algorithmischen Systemen nicht allein auf nationaler Ebene erfolgen. Wir brauchen einen **globalen Blick** auf die Probleme und müssen – vice versa – auch bestrebt sein, unsere Erkenntnisse und Herangehensweisen stärker als bisher in die außereuropäische Debatte einzubringen. Die Erfahrung mit der Umsetzung der DSGVO zeigt, dass die ökonomische Macht des europäischen Wirtschaftsraumes und seine Bedeutung als Absatzmarkt für Betreiber und Anbieter algorithmischer Systeme dazu führen können, dass letztere europäische Rahmenbedingungen bei der Entwicklung und Umsetzung ihrer Produkte und Dienstleistungen aus wirtschaftlichen Interessen heraus berücksichtigen. Zudem dienen europäische Rahmenbedingungen zunehmend auch außereuropäischen Regierungen zur Orientierung bei der Gestaltung ihres eigenen Ordnungsrahmens.

Deswegen sollte der erforderliche Diskussionsprozess insbesondere auch Schwerpunktthema internationaler Foren wie EU, OECD, Europarat, Vereinte Nationen, G7 und G20 sein. Vor diesem Hintergrund empfiehlt die DEK der Bundesregierung, sich aktiv in die entsprechenden internationalen Gremien einzubringen. Insbesondere die **deutsche Ratspräsidentschaft** in der zweiten Jahreshälfte 2020 sollte dazu genutzt werden, die in diesem Gutachten vorgeschlagenen Maßnahmen zur Regulierung des Umgangs mit Daten und algorithmischen Systemen auf europäischer Ebene voranzutreiben. Außerdem plädiert die DEK für eine frühzeitige Mitgestaltung des auf G7-Ebene initiierten Prozesses der Einrichtung eines International Panel on Artificial Intelligence (IPAI) sowie für eine kontinuierliche, aktive Teilnahme der Bundesregierung.

Deutschland und Europa sehen sich im globalen Wettlauf um die Entwicklung von Zukunftstechnologien mit Wertesystemen, Gesellschaftsmodellen und Kulturen konfrontiert, die sich von unseren unterscheiden. Dies hat zu einer Debatte geführt, ob Deutschland und Europa sich an das eine oder andere außereuropäische Modell anpassen müssen, um wettbewerbsfähig zu bleiben. Die DEK unterstützt den bislang eingeschlagenen **„europäischen Weg“** (in der Debatte oft auch als „dritter Weg“ zwischen den Strategien der USA und Chinas bezeichnet), wonach sich europäische Technologien durch konsequente Ausrichtung an europäischen Werten und Grundrechten, wie sie insbesondere auch in der Charta der Grundrechte der Europäischen Union und in der Konvention zum Schutz der Menschenrechte und Grundfreiheiten des Europarats zum Ausdruck kommen, auszeichnen sollten.

Um im Zukunftsdiskurs über das Zusammenspiel von Ethik, Recht und Technologie handlungsfähig zu bleiben, muss die digitale Souveränität Deutschlands und Europas weitestmöglich gewährleistet sein. Digitale Souveränität von Staaten oder Organisationen umfasst das gesamte Feld der Verarbeitung von Daten, das heißt die Kontrolle über die Speicherung, Übertragung und Verwendung ihrer schutzwürdigen Daten inklusive der unabhängigen Entscheidung darüber, wer darauf zugreifen darf. Da grenzüberschreitende Datenflüsse für ein globalisiertes Miteinander von Menschen, Staaten und Unternehmen erforderlich sind und das Internet als Basis für solche Datenflüsse ein globales „Netz der Netze" ist, macht die verteilte weltweite Struktur, die sehr unterschiedliche Rechts- und Gesellschaftssysteme umfasst, eine vollständige Souveränität unmöglich. Damit betrifft digitale Souveränität ganz zentral Fragen der technischen Infrastruktur einschließlich der Hardware, der Netze, der Steuerungskomponenten, wie Router oder Adress-Server, und der Datenzentren. Gerade in Anbetracht der großen Abhängigkeit von ausländischen Produkten sieht die DEK erheblichen Handlungsbedarf auf deutscher und europäischer Ebene durch **Investitionen in die Entwicklung und Sicherung entsprechender Technologien und Infrastrukturen**, um die digitale Souveränität Deutschlands und Europas zu gewährleisten.

Da in Deutschland und sogar europaweit derzeit die wichtigsten der eingesetzten Basiskomponenten für die Internetinfrastruktur fast ausschließlich von Unternehmen aus anderen Kontinenten bezogen werden können, kann sich der Souveränitätsanspruch derzeit nur auf die Fähigkeit beschränken, die verwendeten Basiskomponenten zunächst kritisch zu analysieren und zu bewerten und dann möglichst sicher zu betreiben, um die Gefahr der missbräuchlichen Nutzung durch fremde Staaten und Organisationen zu minimieren. Perspektivisch aber hält es die DEK für bedeutsam, dass Deutschland und Europa ein **höheres Maß an digitaler Souveränität auch in der technischen Infrastruktur** entwickeln. Forschung und Entwicklung von möglichst sicheren Systemen sollten unterstützt werden. Dies umfasst sowohl neu gestaltete Komponenten, um bisherige Systeme zu ersetzen, als auch Ansätze, um auf Basis vorhandener Komponenten trotz bekannter oder vermuteter Unzulänglichkeiten oder Sicherheitsrisiken zu Gesamtlösungen zu kommen, die dem Schutzbedarf angemessen sind.

Digitale Souveränität eines Staates ist allerdings nicht nur im Verhältnis zu anderen Staaten zu sehen, sondern auch im Verhältnis zu nicht-staatlichen Machtzentren. Mit der Entwicklung der Datenwirtschaft gehen **ökonomische Konzentrationstendenzen** einher, die das **Entstehen neuer Machtungleichgewichte** beobachten lassen. Die Forschungs- und Entwicklungsbedingungen im Bereich algorithmischer Systeme und anderer digitaler Technologien werden zunehmend von einigen wenigen großen Digitalunternehmen bestimmt, die häufig auch noch zu wichtigen Finanzquellen für öffentliche Forschung werden und diese mitprägen. Auch die meinungsbildende Funktion von Intermediären und ihr damit verbundener Einfluss auf den gesellschaftspolitischen Diskurs wuchs in den vergangenen Jahrzehnten – ebenso wie die damit verbundene Missbrauchsgefahr – stetig an. Die DEK hält es auf der Grundlage der ethischen und rechtlichen Grundwerte und -freiheiten sowie mit Blick auf die digitale Souveränität Deutschlands und Europas für dringend geboten, die Verschiebung solcher Machtverhältnisse, die für das Funktionieren eines demokratischen Staates und einer sozialen Marktwirtschaft von zentraler Bedeutung sind, umfassend zu beobachten und in den einschlägigen Bereichen wirkungsvoll zu regulieren.

Wer von anderen übermäßig abhängig ist, wird vom „rule maker" zum „rule taker" und setzt seine Bürgerinnen und Bürger letztlich Vorgaben aus, die von Akteuren aus anderen Regionen der Welt formuliert werden. Bemühungen um die **langfristige Sicherung der digitalen Souveränität** sind daher nicht nur ein Gebot politischer Weitsicht, sondern auch Ausdruck ethischer Verantwortung.

Anhang

1. Leitfragen der Bundesregierung an die Datenethikkommission

Koalitionsvertrag:

„Wir werden zeitnah eine Daten-Ethikkommission einsetzen, die Regierung und Parlament innerhalb eines Jahres einen Entwicklungsrahmen für Datenpolitik, den Umgang mit Algorithmen, künstlicher Intelligenz und digitalen Innovationen vorschlägt. Die Klärung datenethischer Fragen kann Geschwindigkeit in die digitale Entwicklung bringen und auch einen Weg definieren, der gesellschaftliche Konflikte im Bereich der Datenpolitik auflöst."

Leitfragen an die Datenethikkommission:

Die Digitalisierung verändert unsere Gesellschaft grundlegend. Neuartige datenbasierte Technologien können zu einem Nutzen für den Alltag des Einzelnen, für Wirtschaft, für Umwelt und Wissenschaft und für die Gesellschaft als Ganzes führen und bergen große Potentiale.

Gleichzeitig werden auch die Risiken der Digitalisierung wahrgenommen. Es stellen sich zahlreiche ethische und rechtliche Fragen, in deren Mittelpunkt die Auswirkungen dieser Entwicklungen und die gewünschte Rolle der neuen Technologien stehen. Wenn der digitale Wandel zum Wohl der gesamten Gesellschaft führen soll, müssen wir uns mit möglichen Folgen der neuen Technologien befassen und ethische Leitplanken definieren.

Eine Herausforderung besteht darin, das Recht für das 21. Jahrhundert so fortzuentwickeln, dass die Menschenwürde („ein Mensch darf nicht zum bloßen Objekt werden") gewahrt bleibt und Grund- und Menschenrechte wie das allgemeine Persönlichkeitsrecht, die Privatsphäre, das Recht auf informationelle Selbstbestimmung, die Diskriminierungsfreiheit, die Wissenschaftsfreiheit, die unternehmerische Freiheit und die Meinungs- und Informationsfreiheit garantiert und zu einem Ausgleich gebracht werden. Dabei bestehen vielfältige Spannungsverhältnisse zwischen Gemeinwohlorientierung, Fortschritt, Innovation und Solidarprinzip.

Diese Kommission soll – unter Berücksichtigung des Diskussions- und Regelungsstandes auf europäischer und internationaler Ebene, nationaler Gestaltungsmöglichkeiten und besonderer Berücksichtigung sensibler Bereiche – ethische Maßstäbe und Leitlinien für den Schutz des Einzelnen, die Wahrung des gesellschaftlichen Zusammenlebens und die Sicherung und Förderung des Wohlstands im Informationszeitalter entwickeln. Die Kommission soll der Bundesregierung auch Empfehlungen oder Regulierungsoptionen vorschlagen, wie die ethischen Leitlinien entwickelt, beachtet, implementiert und beaufsichtigt werden können. Die Vorschläge sollen jeweils auch eine Beschreibung des zugrunde gelegten Begriffsverständnisses und Einschätzungen zu möglichen Folge- und Nebenwirkungen umfassen.

Die Öffentlichkeit soll in geeigneter Weise an der Arbeit der Kommission teilhaben können.

Für ihre Arbeit gibt die Bundesregierung der Datenethikkommission Leitfragen in drei Bereichen zur Hand:

I. Algorithmenbasierte Prognose- und Entscheidungsprozesse („algorithmic decision making" = ADM)

Fortgeschrittene Automatisierungssysteme prägen in immer stärkerem Maße das wirtschaftliche und gesellschaftliche Leben und den Alltag des Einzelnen. Datenerfassung und -analyse ermöglichen die Entwicklung neuartiger Deutungsmodelle, die auch dazu genutzt werden, algorithmenbasierte Entscheidungen zu treffen oder vorzubereiten. Algorithmen ermöglichen es beispielsweise, Verhaltensmuster und Unterschiede im Verhalten verschiedener Gruppen zu erkennen. Ob bei der individuellen Preisgestaltung im Onlinehandel, der Einschätzung der Kreditwürdigkeit oder der Bewerberauswahl bei Einstellungsverfahren: Menschen werden in immer mehr Lebensbereichen von technischen Verfahren bewertet. Die Datenauswertung und die Prognosen über individuelles Verhalten können Chancen bieten (z. B. für die Forschung, die Innovationsfähigkeit der Wirtschaft, die Effizienzsteigerung von Datenverarbeitungsprozessen), bergen aber auch Risiken (z. B. für die individuelle Handlungsfreiheit und Selbstbestimmung, Teilhabe und Chancengleichheit einzelner Menschen wie gesellschaftlicher Gruppen). Gesellschaftliche Ungleichheit und Diskriminierung von Individuen oder Personengruppen kann fortgeschrieben werden, wenn in die Programmierung des Algorithmus oder seine Trainingsdaten tendenziöse Vorfestlegungen („biases") oder Diskriminierungen eingeflossen sind. Diese Risiken bestehen vor allem bei teilhaberelevanten und persönlichkeitssensiblen ADM-Prozessen. Vor diesem Hintergrund stellen sich insbesondere mit Blick auf den Schutz von Verbraucherinnen und Verbrauchern folgende Fragen:

- Welche ethischen Grenzen gibt es für den Einsatz von ADM-Prozessen bzw. sollte es geben?
- Kann es ethisch geboten sein, ADM-Prozesse einzusetzen?
- Gibt es Merkmale, Kriterien oder Datenpunkte, die – beispielsweise aufgrund ihres Alters oder ihrer Herkunft – nicht in ADM-Prozesse einfließen sollten?
- Wie kann ermittelt werden, welche Vorurteile und Verzerrungen in welchen Bereichen ethisch unerwünscht sind? Welche Auswirkungen kann der Einsatz von ADM-Prozessen auf gesellschaftliche Gruppen haben?
- Welche Regulierungsansätze sind denkbar, um Manipulationen, Ungleichbehandlung und Diskriminierung zu verhindern?
- Empfiehlt sich ein abgestufter Regulierungsrahmen abhängig vom Risiko für soziale Teilhabe bzw. dem Diskriminierungspotential?
- Wie kann Verlässlichkeit, Reproduzierbarkeit und Überprüfbarkeit von ADM gewährleistet werden?
- Gibt es Grenzen des Einsatzes von ADM, wenn Einsatz und Kriterien den betroffenen Menschen nicht erklärt werden können?
- Sind Testmethoden möglich, die selbstlernende ADM überprüfbar machen?

II. Künstliche Intelligenz (KI)

Mit der Entwicklung von KI werden in Industrie und Verwaltung immer mehr Systeme mit einem hohen Grad an Automatisierung eingesetzt, die Methoden der KI verwenden und etwa über die Fähigkeit verfügen, durch den Einsatz von Trainingsdaten zu „lernen". Darüber hinaus wird an einer Nachbildung der kognitiven Funktionen im menschlichen Gehirn gearbeitet. Die Entwicklungen im Bereich Künstliche Intelligenz werfen die Frage auf, wie die Würde, die Autonomie und die Selbstbestimmung des Einzelnen gewahrt bleiben und gefördert werden kann. In dem Zusammenhang stellen sich unter anderem folgende Fragen:

- Welche ethischen Grundprinzipien müssen bei der Entwicklung, Programmierung und Nutzung von KI eingehalten werden?

- Wo verlaufen ethische Grenzen für den Einsatz von KI und Robotern, insbesondere in besonderen Lebensbereichen wie Pflege und Betreuung und bei besonders schutzbedürftigen Gruppen (Kinder, ältere Menschen, Menschen mit Behinderungen)? Kann es ethisch geboten sein, KI einzusetzen?

- Kann es bei KI „Ethics by Design" geben? Wenn ja, wie ließe sie sich implementieren und kontrollieren?

- Wie kann sichergestellt werden, dass Maschinen, die auf KI-Basis arbeiten, kontrollierbar sind?

- Wem sind die mit KI generierten Schöpfungen/ Erfindungen zuzuordnen? Wer sollte die Verantwortung für fehlerhaft arbeitende Systeme tragen? Wie kann die Verantwortlichkeit der an der Entwicklung und am Einsatz von KISystemen beteiligten Akteure (Programmierer, Datenwissenschaftler, Auftraggeber, usw.) transparent gemacht?

- Was wird darüber hinaus zukünftig nötig sein, um die für unsere Gesellschaft konstitutiven Freiheiten und Grundrechte nachhaltig zu gewährleisten?

III. Daten

Die Digitalisierung ist gekennzeichnet durch eine Zunahme der Datenmenge (Big Data), durch eine enorme Datenakkumulation bei einzelnen Akteuren, durch die Geschwindigkeit der Datenverarbeitung (Echtzeit), durch Vernetzung (Internet, komplexe Akteursnetzwerke, Internet der Dinge), durch zunehmende Ubiquität und Permanenz von Daten und durch die Weiterentwicklung verschiedener Methoden der Datenanalyse. Dabei steigt mit der Menge der verfügbaren Daten auch die Möglichkeit von immer granulareren Analysen. Durch Daten werden neue Geschäftsmodelle entwickelt und Wertschöpfungsketten sowie Arbeitsprozesse verändert. Daten werden zum Teil als Wirtschaftsgut angesehen, das Wertschöpfung ermöglicht („Datenwirtschaft").

Sowohl auf nationaler als auch auf europäischer Ebene gibt es geltendes Recht (u. a. Datenschutz-Grundverordnung, Open Data) und zahlreiche gesetzgeberische Initiativen, die den Umgang mit Daten betreffen (u. a. e-Privacy-Verordnung, Free Flow of Data). Sie sollen einerseits Grundrechte wie das Recht auf informationelle Selbstbestimmung wahren und andererseits in diesem Rahmen nützliche und innovative Datenverarbeitungen ermöglichen. Diskutiert werden weitere Vorschläge, ob und wie der Zugang zu Daten, die Nutzung von Daten, der Handel mit Daten und Rechte an Daten erstmals oder besser reguliert werden könnten.

Dabei können sich folgende Fragen zum Umgang mit Daten allgemein, zum Datenzugang und zur Datennutzung stellen:

- Welche ethischen Grenzen der Ökonomisierung von Daten gibt es?

- Wer darf den ökonomischen Nutzen aus Daten ziehen?

- Sollte es eine Pflicht zum Angebot von Bezahlmodellen geben?

- Sind einheitliche Regelungen, die für alle Daten gleichermaßen gelten, empfehlenswert? Oder sollten bereichsspezifische Regelungen (z. B. für Gehirndaten) bevorzugt werden? Was sollte der Anknüpfungspunkt für bereichsspezifische Regelungen sein?

- Welche Folgen haben bestehende Zugriffs- und Ausschließlichkeitsrechte an Daten für Wettbewerb und Innovation und welche Folgen hätten zusätzliche Zugriffs- und Ausschließlichkeitsrechte an Daten?

- Bedarf es staatlicher Angebote als Teil der Daseinsvorsorge, damit die Bürgerinnen und Bürger sich verantwortlich, kompetent und souverän im Internet und in den sozialen Netzwerken bewegen können und den Umgang mit Daten beherrschen? Kann die Bereitstellung von Daten, insbesondere offener Daten, ein Teil der staatlichen Daseinsvorsorge werden?

- Wieviel Transparenz ist notwendig und angemessen, um das Recht auf informationelle Selbstbestimmung zu wahren und Bürgerinnen und Bürgern eine selbstbestimmte Teilhabe am Wirtschaftsleben zu ermöglichen?

- Erfordern besondere Lebenslagen spezielle Schutzkonzepte für einzelne Nutzergruppen?

- Sind die bestehenden Institutionen in sensiblen Bereichen ausreichend, um eine ethisch vertretbare Nutzung von Daten sicherzustellen? Wie kann eine ausreichende Vertretung der jeweiligen Stakeholder nachhaltig sichergestellt werden?

- Welche Auswirkungen können umfassende Datensammlungen auf das Funktionieren der Marktwirtschaft (z. B. Wettbewerbsfähigkeit, Informationsasymmetrie zwischen Anbietern und Verbrauchern, Möglichkeit, innovative Produkte zu entwickeln) und der Demokratie (z. B. Erfassung und Auswertung des Verhaltens in sozialen Netzwerken) haben? Wie kann erforderlichenfalls gegen Datenmacht/Datensilos (insbesondere Intermediäre) vorgegangen werden?

- Sollten Daten oder der Zugang zu ihnen in bestimmten Fällen zum Allgemeingut erklärt werden? In welchen Fällen und unter welchen ethischen Kriterien?

- Die Nutzung von nicht-personenbezogenen Daten kann kollektive Wirkungen haben. So können zum Beispiel Einzelne oder bestimmte Bevölkerungsgruppen schlechter gestellt werden, weil die Datenanalyse ergibt, dass in einem bestimmten Stadtviertel die Zahlungsmoral geringer ist. Welche Regelungsinstrumente wären hierfür notwendig? In welchen Sektoren?

- Ist eine gesetzliche Regelung zur Verbesserung des Zugangs zu Daten möglich, erforderlich und sinnvoll?

- Muss es aus ethischen Gründen Datenverarbeitungsverbote geben, etwa bei bestimmten Datenarten (z. B. politische Einstellung; Gehirndaten) oder bestimmten Verwendungsbereichen (z. B. Profiling für politische Zwecke oder zur Verwendung bei Wahlen)?

- Unter welchen Voraussetzungen kann es eine ethische Pflicht zur Datennutzung geben?

- Wird ein möglicher Gemeinwohlnutzen der Datenverarbeitung von der Rechtsordnung in hinreichender Weise anerkannt? Wenn nein, wie kann dies erreicht werden?

- Ist es möglich und sinnvoll, Experimentierklauseln zur Erprobung neuer Anwendungen oder neuer Regulierungsinstrumente zu schaffen?

- Ist es sinnvoll, in Dateninfrastrukturen zu investieren? Wenn ja, in welche?

- Wie können die grundrechtlich geschützten Interessen des Einzelnen, der Unternehmen, der Wissenschaft und Kunst und das Gemeinwohlinteresse an der Datennutzung in Einklang gebracht werden?

2. Mitglieder der Datenethikkommission der Bundesregierung

Co-Sprecherinnen

Prof. Dr. Christiane Wendehorst

- Professorin für Zivilrecht an der Universität Wien
- Mitglied im Vorstand des Instituts für Innovation und Digitalisierung im Recht an der Universität Wien
- Präsidentin des European Law Institute (ELI)

Prof. Dr. Christiane Woopen

- Professorin für Ethik und Theorie der Medizin und Leiterin der Forschungsstelle Ethik an der Uniklinik Köln
- Geschäftsführende Direktorin des Cologne Center for Ethics, Rights, Economics, and Social Sciences (ceres) der Universität zu Köln
- Vorsitzende des Europäischen Ethikrates (EGE)

Mitglieder

Prof. Dr. Johanna Haberer

- Leitung der Professur für Christliche Publizistik an der Friedrich-Alexander-Universität Nürnberg-Erlangen
- Geschäftsführerin des Instituts für Praktische Theologie an der Friedrich-Alexander-Universität Nürnberg-Erlangen

Marit Hansen

- Landesbeauftragte für Datenschutz Schleswig-Holstein
- Leiterin des Unabhängigen Landeszentrums für Datenschutz (ULD)

Prof. Dr. Dirk Heckmann

- Inhaber des Lehrstuhls für Recht und Sicherheit der Digitalisierung an der Technischen Universität München
- Direktor am Bayerischen Forschungsinstitut für Digitale Transformation
- Verfassungsrichter am Bayerischen Verfassungsgerichtshof

Prof. Ulrich Kelber

- Bundesbeauftragter für den Datenschutz und die Informationsfreiheit
- Honorarprofessor an der Hochschule Bonn-Rhein-Sieg

Prof. Dieter Kempf

- Präsident des Bundesverbandes der Deutschen Industrie e. V.
- Honorarprofessor an der Friedrich-Alexander-Universität Erlangen-Nürnberg

Prof. Dr. Mario Martini

- Inhaber des Lehrstuhls für Verwaltungswissenschaft, Staatsrecht, Verwaltungsrecht und Europarecht an der DUV Speyer
- Leiter des Programmbereichs „Transformation des Staates durch Digitalisierung“ und Stellvertretender Direktor des Deutschen Forschungsinstituts für öffentliche Verwaltung

Klaus Müller

- Vorstand des Verbraucherzentrale Bundesverbands (vzbv e. V.)
- Lehrbeauftragter an der Heinrich-Heine-Universität Düsseldorf

Paul Nemitz

- Hauptberater in der EU Kommission, Generaldirektion Justiz und Verbraucherschutz

Prof. Dr. Sabine Sachweh

- Professorin für Angewandte Softwaretechnik an der Fachhochschule Dortmund
- Sprecherin und Vorstandsmitglied des Instituts für die Digitalisierung von Arbeits- und Lebenswelten (IDiAL) der Fachhochschule Dortmund
- Ko-Sprecherin im Fachbeirat „Digitalisierung und Bildung für ältere Menschen“ des Bundesministeriums für Familie, Senioren, Frauen und Jugend

Christin Schäfer

- Gründerin und Geschäftsführerin des Unternehmens acs plus, einer Boutique für Data Science
- Beirätin der Forschungsgruppe Big Data Analytics des IW Köln

Prof. Dr. Rolf Schwartmann

- Professor für Bürgerliches Recht und Wirtschaftsrecht an der Technischen Hochschule Köln
- Leiter der Forschungsstelle für Medienrecht an der Technischen Hochschule Köln
- Vorsitzender der Gesellschaft für Datenschutz und Datensicherheit (GDD) e.V.

Prof. Dr. Judith Simon

- Professorin für Ethik in der Informationstechnologie an der Universität Hamburg

Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster

- Professor für Informatik, Lehrstuhl für Künstliche Intelligenz, Universität des Saarlandes
- CEO/CEA des Deutschen Forschungszentrums für Künstliche Intelligenz
- Leiter des Steuerungskreises für die KI-Normungsroadmap beim Deutschen Institut für Normung (DIN)

Prof. Dr. Thomas Wischmeyer

- Juniorprofessor (Tenure Track) für Öffentliches Recht und Recht der Digitalisierung an der Universität Bielefeld

Stand: 10. Oktober 2019

Impressum

Berlin, Oktober 2019

Gutachten der Datenethikkommission
der Bundesregierung

Herausgeber
Datenethikkommission der Bundesregierung
Bundesministerium des Innern, für Bau und Heimat
Alt-Moabit 140
10557 Berlin
Bundesministerium der Justiz und für Verbraucherschutz
Mohrenstraße 37
10117 Berlin

E-Mail
datenethikkommission_gs@bmi.bund.de
datenethikkommission_gs@bmjv.bund.de

Internet
www.datenethikkommission.de

Gestaltung
Atelier Hauer + Dörfler GmbH, Berlin

Bildnachweis
S. 53 shutterstock.com;
S. 234: BMI (Gruppenfoto), Studio Wilke (Christiane Wendehorst), Reiner Zensen (Christiane Woopen), BPA/Kugler (Ulrich Kelber);
S. 235: Christian Kruppa (Dieter Kempf), vzbv/Gert Baumbach (Klaus Müller), Markus Mielek (Sabine Sachweh), TH Köln/Schmülgen (Rolf Schwartmann), UHH/Nicolai (Judith Simon), Jim Rakete (Wolfgang Wahlster)

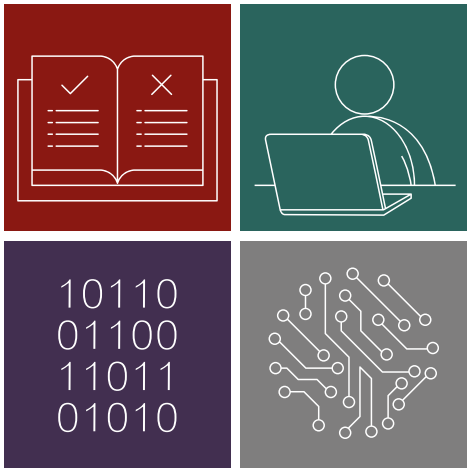
Druck
Brandenburgische Universitätsdruckerei und Verlagsgesellschaft Potsdam mbH (bud)

Ausschließlich zum Zweck der besseren Lesbarkeit wird im vorliegenden Gutachten der Datenethikkommission auf die geschlechtsspezifische Schreibweise verzichtet. Alle personenbezogenen Bezeichnungen sind geschlechtsneutral zu verstehen.

Artificial Intelligence and Public Standards

A Review by the Committee on Standards in Public Life

The Committee on Standards in Public Life



February 2020

Artificial Intelligence and Public Standards

A Review by the Committee on Standards in Public Life

Chair, Lord Evans of Weardale KCB DL

February 2020

The Seven Principles of Public Life

The Seven Principles of Public Life (also known as the Nolan Principles) apply to anyone who works as a public office-holder. This includes all those who are elected or appointed to public office, nationally and locally, and all people appointed to work in the Civil Service, local government, the police, courts and probation services, non-departmental public bodies (NDPBs), and in the health, education, social and care services. All public office-holders are both servants of the public and stewards of public resources. The principles also apply to all those in other sectors delivering public services.

Selflessness

Holders of public office should act solely in terms of the public interest.

Integrity

Holders of public office must avoid placing themselves under any obligation to people or organisations that might try inappropriately to influence them in their work. They should not act or take decisions in order to gain financial or other material benefits for themselves, their family, or their friends. They must declare and resolve any interests and relationships.

Objectivity

Holders of public office must act and take decisions impartially, fairly and on merit, using the best evidence and without discrimination or bias.

Accountability

Holders of public office are accountable to the public for their decisions and actions and must submit themselves to the scrutiny necessary to ensure this.

Openness

Holders of public office should act and take decisions in an open and transparent manner. Information should not be withheld from the public unless there are clear and lawful reasons for so doing.

Honesty

Holders of public office should be truthful.

Leadership

Holders of public office should exhibit these principles in their own behaviour. They should actively promote and robustly support the principles and be willing to challenge poor behaviour wherever it occurs.

Chair's Foreword

Dear Prime Minister,

I am pleased to present the 21st report of the Committee on Standards in Public Life on the subject of artificial intelligence and public standards.

Artificial intelligence – and in particular, machine learning – will transform the way public sector organisations make decisions and deliver public services. The government has committed significant resources to this new technology through the AI Sector Deal, which promises to deliver a more accurate, capable and efficient public sector.

Adherence to high public standards will help fully realise the great benefits of AI in public service delivery. By ensuring that AI is subject to appropriate safeguards and regulations, the public can have confidence that new technologies will be used in a way that upholds the Seven Principles of Public Life. Our conclusion from this review is that these principles will remain a valid guide for public sector practice as AI is deployed across all levels of government.

Our recommendations are directed towards three key audiences.

Our message to government is that the UK's regulatory and governance framework for AI in the public sector remains a work in progress and deficiencies are notable. The work of the Office for AI, the Alan Turing Institute, the Centre for Data Ethics and Innovation (CDEI), and the Information Commissioner's Office (ICO) are all commendable. But on the issues of transparency and data bias in particular, there is an urgent need for practical guidance and enforceable regulation.

Regulators must also prepare for the changes AI will bring to public sector practice. We conclude that the UK does not need a specific AI regulator, but all regulators must adapt to the challenges that AI poses to their specific sectors. Government should establish the CDEI as a centre for regulatory assurance to assist regulators in this area.

Upholding public standards will also require action from public bodies using AI to deliver frontline services. All public bodies must comply with the law surrounding data-driven technology and implement clear, risk-based governance for their use of AI.

Artificial intelligence – particularly in the public sector – is the subject of significant media interest and this report will not be the final word on the matter. Nonetheless, we believe our contribution will help the UK public sector uphold public standards as it adopts AI across a wide range of public service delivery.

Lord Evans of Weardale KCB DL
Chair, Committee on Standards in Public Life

Contents

Executive summary

Artificial intelligence has the potential to revolutionise the delivery of public services, creating an opportunity for more innovative and efficient public service delivery. Machine learning in particular will transform the way decisions are made in areas as diverse as policing, health, welfare, transport, social care, and education.

This review found that the Nolan Principles are strong, relevant, and do not need reformulating for AI. The Committee heard that they are principles of good governance that have stood, and continue to stand, the test of time. All seven principles will remain relevant and valid as AI is increasingly used for public service delivery.

If correctly implemented, AI offers the possibility of improved public standards in some areas. However, AI poses a challenge to three Nolan Principles in particular: openness, accountability, and objectivity. This review examined how public officials and government departments can uphold these principles as AI is increasingly rolled out across our public services.

Our concerns here overlap with key themes from the field of AI ethics. Under the principle of openness, a current lack of information about government use of AI risks undermining transparency. Under the principle of accountability, there are three risks: AI may obscure the chain of organisational accountability; undermine the attribution of responsibility for key decisions made by public officials; and inhibit public officials from providing meaningful explanations for decisions reached by AI. Under the principle of objectivity, the prevalence of data bias risks embedding and amplifying discrimination in everyday public sector practice.

This review found that the government is failing on openness. Public sector organisations are not sufficiently transparent about their use of AI and it is too difficult to find out where machine learning is currently being used in government. It is too early to judge if public sector bodies are successfully upholding accountability. Fears over 'black box' AI, however, may be overstated, and the Committee believes that explainable AI is a realistic goal for the public sector. On objectivity, data bias is an issue of serious concern, and further work is needed on measuring and mitigating the impact of bias.

Governance and regulation

To uphold public standards, government and public sector organisations should set effective governance to mitigate the risks we have identified. In this sense, AI is a new challenge that can be solved with existing tools and established principles. Public standards can be upheld with a traditional risk management approach.

This is not a challenge that public sector organisations can tackle alone. Government needs to identify and embed authoritative ethical principles and issue accessible guidance on AI governance to those using it in the public sector. Government and regulators must also establish a coherent regulatory framework that sets clear legal boundaries on how AI should be used in the public sector.

Attempts to establish this governance and regulatory framework are emerging and developments are fast-moving. In the area of ethical principles and guidance, the Department for Culture, Media and Sport (DCMS), the Centre for Data Ethics and Innovation (CDEI) and the Office for AI have all published ethical principles for data-driven technology and AI. The Office for AI, the Government Digital Service (GDS), and the Alan Turing Institute have jointly issued A Guide to Using Artificial Intelligence in the Public Sector and draft guidelines on AI procurement. The Information Commissioner's Office (ICO) has also published its Auditing Framework for AI.

In the area of regulation, the use of AI is subject to the provisions of the GDPR, the Equality Act, and sections of administrative law. The government has also established the Centre for Data Ethics and Innovation to advise on regulation.

These developments are positive and are to be welcomed. However, at the time of writing, this review has found that the governance and regulatory framework for AI in the public sector is still a work in progress and one with significant deficiencies.

This is mostly because key documents have only recently been published and government AI institutions are very new. Multiple sets of ethical principles are confusing and the application of each is unclear. Public sector guidance is not yet widely used and public officials with no AI expertise may find it difficult to understand and comply with.

We conclude that a new AI regulator is not needed but existing regulators will need to adapt to face the challenges AI brings. They will need assistance from a central body to do so, but the CDEI does not yet have a clearly defined purpose and is not yet on a statutory footing. Two areas in particular – transparency and data bias – are in need of urgent attention in the form of new regulation and guidance.

Our recommendations

Our recommendations to government and regulators are intended to assist in the development of a stronger and more coherent regulatory and governance framework for AI in the public sector.

We recommend that government should establish consistent and authoritative ethical principles and issue easier to use guidance. Procurement processes should be reformed and the Digital Marketplace should offer greater assistance to public bodies seeking technologies that are compliant with public standards.

Though no new AI regulator is needed, the CDEI should advise regulators on how to adapt to new technologies and be set on an independent statutory footing. The application of anti-discrimination law to AI needs to be clarified and new transparency guidelines are needed. AI impact assessments should be mandatory, published, and set by the CDEI, and new guidelines are needed to enforce transparency.

We also provide recommendations to providers of public services, both public and private, to help them develop effective risk-based governance for AI. During project planning, our recommendations focus on legal and legitimate AI, system design, and diversity. During project implementation, our recommendations cover setting responsibility, internal and external oversight, monitoring and evaluation, appeal and redress, and training and education.

The Nolan Principles remain a valid guide for public sector practice in the age of AI. However, this new technology is a fast-moving field, so government and regulators will need to act swiftly to keep up with the pace of innovation. Our recommendations set out what we believe is needed to ensure the Seven Principles of Public Life are upheld as the public sector transitions into a new AI-enabled age.

List of recommendations

Recommendations to government, national bodies and regulators

The Committee makes eight recommendations to government, national bodies and regulators to help create a strong and coherent governance and regulatory framework for AI in the public sector.

Recommendation 1: Ethical principles and guidance

There are currently three different sets of ethical principles intended to guide the use of AI in the public sector – the FAST SUM Principles, the OECD AI Principles, and the Data Ethics Framework. It is unclear how these work together and public bodies may be uncertain over which principles to follow.

a. The public needs to understand the high level ethical principles that govern the use of AI in the public sector. The government should identify, endorse and promote these principles and outline the purpose, scope of application and respective standing of each of the three sets currently in use.

b. The guidance by the Office for AI, the Government Digital Service and the Alan Turing Institute on using AI in the public sector should be made easier to use and understand, and promoted extensively.

Recommendation 2: Articulating a clear legal basis for AI

All public sector organisations should publish a statement on how their use of AI complies with relevant laws and regulations before they are deployed in public service delivery.

Recommendation 3: Data bias and anti-discrimination law

The Equality and Human Rights Commission should develop guidance in partnership with both the Alan Turing Institute and the CDEI on how public bodies should best comply with the Equality Act 2010.

Recommendation 4: Regulatory assurance body

Given the speed of development and implementation of AI, we recommend that there is a regulatory assurance body, which identifies gaps in the regulatory landscape and provides advice to individual regulators and government on the issues associated with AI.

We do not recommend the creation of a specific AI regulator, and recommend that all existing regulators should consider and respond to the regulatory requirements and impact of the growing use of AI in the fields for which they have responsibility.

The Committee endorses the government's intention for CDEI to perform a regulatory assurance role. The government should act swiftly to clarify the overall purpose of CDEI before setting it on an independent statutory footing.

Recommendation 5: Procurement rules and processes

Government should use its purchasing power in the market to set procurement requirements that ensure that private companies developing AI solutions for the public sector appropriately address public standards.

This should be achieved by ensuring provisions for ethical standards are considered early in the procurement process and explicitly written into tenders and contractual arrangements.

Recommendation 6: The Crown Commercial Service's Digital Marketplace

The Crown Commercial Service should introduce practical tools as part of its new AI framework that help public bodies, and those delivering services to the public, find AI products and services that meet their ethical requirements.

Recommendation 7: Impact assessment

Government should consider how an AI impact assessment requirement could be integrated into existing processes to evaluate the potential effects of AI on public standards. Such assessments should be mandatory and should be published.

Recommendation 8: Transparency and disclosure

Government should establish guidelines for public bodies about the declaration and disclosure of their AI systems.

Recommendations to front-line providers, both public and private, of public services

The Committee makes seven recommendations to front-line providers of public services to help establish effective risk-based governance for the use of AI.

Recommendation 9: Evaluating risks to public standards

Providers of public services, both public and private, should assess the potential impact of a proposed AI system on public standards at project design stage, and ensure that the design of the system mitigates any standards risks identified.

Standards review will need to occur every time a substantial change to the design of an AI system is made.

Recommendation 10: Diversity

Providers of public services, both public and private, must consciously tackle issues of bias and discrimination by ensuring they have taken into account a diverse range of behaviours, backgrounds and points of view. They must take into account the full range of diversity of the population and provide a fair and effective service.

Recommendation 11: Upholding responsibility

Providers of public services, both public and private, should ensure that responsibility for AI systems is clearly allocated and documented, and that operators of AI systems are able to exercise their responsibility in a meaningful way.

Recommendation 12: Monitoring and evaluation

Providers of public services, both public and private, should monitor and evaluate their AI systems to ensure they always operate as intended.

Recommendation 13: Establishing oversight

Providers of public services, both public and private, should set oversight mechanisms that allow for their AI systems to be properly scrutinised.

Recommendation 14: Appeal and redress

Providers of public services, both public and private, must always inform citizens of their right and method of appeal against automated and AI-assisted decisions.

Recommendation 15: Training and education

Providers of public services, both public and private, should ensure their employees working with AI systems undergo continuous training and education.

Introduction

The Committee on Standards in Public Life (the Committee) was established in 1994. In its first report, the Committee articulated the Seven Principles of Public Life, commonly referred to as the Nolan Principles: selflessness, integrity, objectivity, accountability, openness, honesty and leadership.

The standards landscape has changed significantly since then and the context within which the Committee operates continues to evolve. The Committee, in its 2013 report Standards Matter, said:

> "The systems and practices of public organisations, the culture and behaviour of public office-holders and the expectations of the public are constantly subject to new influences and constraints, causing them to develop in new and sometimes unexpected ways."[1]

This is particularly true in the case of new technology. Artificial Intelligence (AI) will fundamentally change the way that government and the public sector operates, and new technology could help design better public policy and deliver more efficient and effective public services.

AI could be used in ways that are uncontroversial. For example, AI could be used to create smart traffic lights where the timing of a red or green light is altered to create the most efficient traffic flow possible.

There is already evidence, however, that AI can be used in more controversial ways. In 2017, a machine learning system was used to see whether there were any patterns in publicly available data to indicate whether a school was likely to be inadequate or failing.[2] Key correlations were identified and Ofsted began using machine learning to take a more targeted approach to its inspections. Teachers protested and argued that the tool was unfair, lacked transparency, and had the potential to exacerbate pre-existing biases within the education system.

This controversy speaks to wider concerns shared by the Committee. Any change in how the government makes policy decisions and delivers public services must not undermine public standards and the public's confidence in its institutions. This is particularly important in the context of AI because AI has the potential to change how decisions are made in sensitive policy areas like social care, policing and criminal justice, where the impact on individuals can be significant. It is in these areas that standards will matter most.

The increasing use of AI in public service delivery is also of interest to the Committee because public bodies will not be delivering this change alone. Private companies will often work alongside public bodies to develop and deliver AI solutions. The involvement of commercial organisations in the delivery of public services means that additional care must be given to standards issues.

How the government manages private sector service delivery in a way that exemplifies high ethical standards is not a new issue for the Committee. As the Committee said in its 2018 report The Continuing Importance of Ethical Standards for Public Service Providers, the public is right to expect services to be delivered responsibly and ethically, regardless of how they are being delivered, or who is providing those services.[3]

1 Committee on Standards in Public Life (2013), Standards Matter, Cm 8519, 12. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/348304/Standards_Matter.pdf

2 The Behavioural Insights Team (2017), Using Data Science in Policy, 14. Available at: https://www.bi.team/wp-content/uploads/2017/12/BIT_DATA-SCIENCE_WEB-READY.pdf

3 The Committee on Standards in Public life (2018), The Continuing Importance of Ethical Standards for Public Service Providers. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/705884/20180510_PSP2_Final_PDF.pdf

The Committee undertook this review to ensure that the public sector continues to maintain high standards of conduct as it grapples with the implications and developments made possible by the introduction of AI. This report and its recommendations are designed to help government and public bodies uphold standards as they start to use AI across a wide range of public service provision.

General concerns about privacy in the field of AI are beyond the remit of the Seven Principles of Public Life and are not discussed in this report.

Chapter 1: AI in the UK, is an overview of AI, why it matters for the public sector, and current government policy.

Chapter 2: AI and the Nolan Principles, considers the relevance of the Nolan Principles in this new age of artificial intelligence and examines the risks and opportunities for openness, accountability and objectivity.

Chapter 3: Guidance and ethical principles, examines the recent publication of guidance and ethical principles for using AI in the public sector.

Chapter 4: Regulating AI, assesses the legal and regulatory framework for AI in the public sector.

Chapter 5: The role of public bodies, outlines how public bodies can manage and mitigate the risk AI poses to public standards through good governance.

The challenge AI poses for standards is real, but it does not require a fundamental reworking of public sector practice. Good, proactive and careful governance is key, and it is the role of government and regulators to encourage this through the development and enforcement of clear and effective AI regulation. This will encourage public institutions to establish suitable governance mechanisms to manage the standards risks associated with the technology they use.

The Committee collected a wide range of evidence for this review, meeting individually with experts in the field from government, academia, and the public and private sectors, holding roundtables, and attending external conferences and workshops. The Committee also held focus groups with members of the public and commissioned public polling on attitudes to AI. The Committee is indebted to all those who contributed to this review.

Chapter 1: AI in the UK

> "Artificial Intelligence is one of the most transformative forces of our time, and is bound to alter the fabric of society."[4]
> **European Commission, Independent High-Level Expert Group on AI**

1.1. What is AI?

There is no single uncontested definition of what constitutes AI and the term is used liberally to describe anything from routine data analysis to complex deep neural networks. Whichever definition is used, there is wide agreement that the potential for change and impact on society is immense.

Experts predict it is machine learning that will have the most significant impact and lead to transformative change. Machine learning systems learn from past data by identifying patterns and correlations within it. This allows computers to undertake increasingly complex tasks, like natural language processing and image recognition. These innovations will transform the power of computers to interpret our world. AI systems will be able to analyse and predict human behaviour on an unprecedented scale, in areas as diverse as crime, transport and health.

Complex processes of filtering, cross-referencing and authenticating information, such as the personal data of a benefits claimant to establish their entitlement, could be automated and instantaneous. AI could process and respond coherently to natural language, giving computers the capacity to read legal contracts and converse fluently with human interlocutors. Image recognition could be able to identify distinct people, animals and objects in images and video in real time.

Police forces across the UK are, for example, already using live facial recognition technology to assist in the prevention and detection of crime by identifying wanted criminals.

AI will undoubtedly change the relationship between humans and technology, as well as between citizens and the state. This is because AI allows computers, for the first time, to assist in decision-making processes in a substantive and meaningful way, independent of human judgement.

1.2. The scale of AI

The impact of AI across the public and private sectors is potentially vast. These advances in computing capability will revolutionise areas such as finance, energy, health, education and agriculture. The Office for AI (see opposite) estimates that AI could add £232 billion to the UK's economy by 2030, boosting productivity in some industries by 30%.[5] AI is also an international issue. Over 25 countries have published an AI strategy, and the European Union, United Nations, and OECD have all taken a close interest in AI governance and ethics. The question of how AI can be used effectively and ethically is of global concern and there would be benefit to the UK working with its international partners in a shared approach.

Government has a particular responsibility to exercise care around the use of AI in the public sector. Citizens can choose not to use a particular private company's products or services, but citizens cannot always opt out of public service delivery. Public sector AI will be funded by taxpayers' money, and in some cases AI will be part of the operation of the law. The use of AI in the UK public sector must follow the Seven Principles of Public Life, which outline the ethical values to which the public sector should adhere.

4 Council of Europe (2018), Draft Recommendation of the Committee of Ministers to member States on human rights impacts of algorithmic systems. Available at: https://rm.coe.int/draft-recommendation-on-human-rights-impacts-of-algorithmic-systems/16808ef256

5 Office for Artificial Intelligence (AI) (2019), AI Sector Deal One Year On. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819331/AI_Sector_Deal_One_Year_On__Web_.pdf

1.3. UK policy on AI

In 2017, the government published an independent review led by Professor Dame Wendy Hall and Jerome Pesenti (the Hall-Pesenti review) on how the AI industry could be developed in the UK.[6] The review made a number of important recommendations to improve access to skills and data, maximise AI research and support the uptake of AI.

Following the Hall-Pesenti review, the government published its 2017 Industrial Strategy,[7] which identified AI and data as one of four 'Grand Challenges' to modernise the UK economy.[8] The AI Sector Deal was published in 2018.[9] It made clear that the government sees AI as a "huge global opportunity" and wants to become a global leader in AI and data-driven technology.

The AI Sector Deal led to the creation of three new institutions: a government Office for AI; an industry-led AI Council; and the Centre for Data Ethics and Innovation (CDEI).

The Office for AI

The Office for AI is a joint office sponsored by the Department for Digital, Culture, Media and Sport (DCMS), and the Department for Business, Energy and Industrial Strategy (BEIS).

Their role is to oversee the implementation of the AI Sector Deal, which is part of the AI and Data Grand Challenge. They have recently published the AI Guide on implementing AI in the public sector, as well as draft guidelines for AI procurement. The Office for AI aims to increase adoption of AI across the private and public sectors.

The AI Council

The Hall-Pesenti review recommended that government should work with industry to establish an industry-led AI Council to advise government on AI.

The AI Council is an independent expert committee that advises government on how to promote the growth of AI in the UK. It includes representatives from the public and private sectors.

The Centre for Data Ethics and Innovation (CDEI)

CDEI is an independent public sector body established by DCMS to advise government on artificial intelligence and other data-driven technologies. It is tasked to help develop the right regulation and governance for data-driven technology.

CDEI is currently undertaking thematic reviews on issues of public concern, such as data bias and online targeting. Their remit includes both the public and private sectors.

CDEI is not yet on a statutory footing and its final status is yet to be determined. The government has said it intends to place CDEI on a statutory footing after its initial phase of operation.

6 Professor Dame Wendy Hall and Jerome Pesenti (2017), Growing the Artificial Intelligence Industry in the UK. Available at: https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

7 Department for Business, Energy and Industrial Strategy (2017), Industrial Strategy: Building a Britain fit for the Future, White Paper. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/664563/industrial-strategy-white-paper-web-ready-version.pdf

8 Department for Business, Energy and Industrial Strategy (2019), The Grand Challenges, Policy Paper. Available at: https://www.gov.uk/government/publications/industrial-strategy-the-grand-challenges/industrial-strategy-the-grand-challenges#artificial-intelligence-and-data

9 Department for Business, Energy and Industrial Strategy (2018), Industrial Strategy: Artificial Intelligence Sector Deal. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/702810/180425_BEIS_AI_Sector_Deal__4_.pdf

These new institutions will work alongside other public bodies created before the Sector Deal. Notably, the Alan Turing Institute was made the national institute for artificial intelligence and data science by the government in 2017, in response to a recommendation made in the Hall-Pesenti review. Currently based at the British Library, the Institute convenes academics and industry to research AI and its impact on society. The Institute has been working with the Office for AI to produce guidance on how to use AI ethically and safely.

The Government Digital Service (GDS) also has a significant role in shaping AI policy. GDS is part of the Cabinet Office and is responsible for digital transformation across government. GDS currently sets and enforces standards for digital technology, including around procurement.

In 2019, GDS published joint guidance with the Office for AI and the Alan Turing Institute on how to use AI in the public sector.[10]

AI used in the UK public sector is also subject to the Data Ethics Framework, published by the Department for Digital, Culture, Media and Sport (DCMS) in 2018.[11] The framework guides appropriate data use in government and the wider public sector, and is aimed at anyone working directly or indirectly with data, including data scientists, policymakers and operational staff. It is not binding but builds on the core values of the Civil Service Code – integrity, honesty, objectivity and impartiality – to encourage ethical data use, build better services and inform policy.

DCMS has also published guidance around each principle, and a Data Ethics Workbook to help practitioners align their work with the framework's principles.

The Data Ethics Framework principles

1. Start with clear user need and public benefit
2. Be aware of relevant legislation and codes of practice
3. Use data that is proportionate to the user need
4. Understand the limitations of the data
5. Ensure robust practices and work within your skillset
6. Make your work transparent and be accountable
7. Embed data use responsibly.[12]

10 Office for AI and Government Digital Service (GDS) (2019), A guide to using AI in the public sector. Available at: https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

11 Department for Digital, Culture, Media and Sport (2018), Data Ethics Framework. Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework

12 Same source

1.4. Where is AI being used in the UK public sector?

Despite generating much interest and commentary, our evidence shows that the adoption of AI in the UK public sector remains limited. Most examples the Committee saw of AI in the public sector were still under development or at a proof-of-concept stage.

The Committee heard, however, that many public bodies are beginning to look at how they can use AI to deliver better public services. Our evidence showed that healthcare and policing currently have the most developed AI programmes, with technology being used, for example, to identify eye disease and to predict reoffending rates, though levels of system maturity differ across NHS trusts and police forces.

The Committee found that the Judiciary, the Department for Transport (DfT) and the Home Office are examining how AI can increase efficiency in service delivery. The Home Office also told us that they are currently looking at the governance structures that need to be in place when AI is used in the public sector.

The Committee was told that local government is currently innovating with AI systems in education, welfare and social care. Hampshire County Council, for example, is trialling the use of smart devices, such as Amazon Echo, in the homes of adults receiving social care, to bridge the gap between visits from professional carers.[13] The Guardian reported that one-third of councils use algorithmic systems to make welfare decisions.[14]

It is the view of the Committee, however, that obstacles to widespread and successful adoption remain significant. Public policy experts frequently told this review that access to the right quantity of clean, good-quality data is limited, and that trial systems are not yet ready to be put into operation. It is our impression that many public bodies are still focusing on early-stage digitalisation of services, rather than more ambitious AI projects.

Multiple contributors to the review also commented that the lack of a clear standards framework – including in law and regulation – meant that organisations did not have the confidence to use AI. While standards and regulation are often seen as barriers to innovation, the Committee believes that implementing clear ethical standards around AI may accelerate rather than delay adoption, by building trust in new technologies among public officials and service users.

13 SA Mathieson (2019), 'I feel in control of my life: Alexa's new role in public service', The Guardian. Available at: https://www.theguardian.com/society/2019/feb/07/control-life-alexa-role-public-service-chatbots-councils

14 Sarah Marsh (2019), 'One in three councils using algorithms to make welfare decisions', The Guardian. Available at: https://www.theguardian.com/society/2019/oct/15/councils-using-algorithms-make-welfare-decisions-benefits

Chapter 2: AI and the Nolan Principles

2.1. The Seven Principles of Public Life

The Seven Principles of Public Life apply to anyone who works as a public office-holder. They also apply to those in the private sector delivering public services. These well-established principles set the standards across the whole of public service.

The Nolan Principles have been widely accepted as the basis of good practice throughout the public sector. They are mentioned explicitly in the UK, Scottish, Welsh and Northern Irish Ministerial Codes, are included in the corporate documentation of a large number of public sector organisations, and form the basis of the codes of conduct required of all local authorities. Some organisations, like the Civil Service, have adapted the principles to their own particular context.[15]

The way public bodies view ethical standards has, however, changed over the years. The Committee noted in its 2013 report Standards Matter that in many organisations, the debate on ethical standards has shifted from an emphasis on personal standards to an approach which places greater weight on managing risks to standards in an organisation as a whole.[16]

The increased adoption of AI will bring new challenges to the practices of public organisations and the behaviour of public office-holders, as well as affecting the expectations of the public. As part of this review, the Committee examined whether artificial intelligence would require a fundamental rethinking of public standards.

There was a general consensus among contributors to this review that the Nolan Principles are strong, relevant, and do not need reformulating for AI systems. The Committee heard that they are principles of good governance that have stood, and continue to stand, the test of time.

This was partly because they are already well known and embedded within the cultures of organisations across the public service, and also because they are highly relevant in terms of the ethical challenges AI will have to meet.

The Committee is aware that while principles are important, they are not sufficient on their own as a complete guide for behaviour in public life. To ensure that ethical principles generate changes in behaviour, they need to be elaborated in codes of conduct and guidance and implemented through policy and governance. The application of the principles in AI may not be self-evident, and in some cases it will be unclear exactly how public officials should uphold these ethical principles in practice.

The Committee heard that more needs to be done to achieve this behavioural change. Codes and principles should be embedded into current practices through better governance, leadership and education. The Committee also heard that internal systems for upholding standards in public bodies should be supported by independent scrutiny.

The following chapters of this report outline what government and public bodies can do to translate the Seven Principles into practice for the use of AI.

2.2. Where is AI likely to affect public standards?

All of the Seven Principles of Public Life must be upheld when using AI in the public sector. Three principles are particularly relevant: openness, accountability and objectivity.

15 Civil Service Code of Conduct (2015). Available at: https://www.gov.uk/government/publications/civil-service-code/the-civil-service-code

16 Committee on Standards in Public Life (2013), Standards Matter, Cm 8519. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/348304/Standards_Matter.pdf

2.3. Openness

Holders of public office should act and take decisions in an open and transparent manner. Information should not be withheld from the public unless there are clear and lawful reasons for so doing.

2.3.1. Why openness matters

The public can only scrutinise and understand the decisions of government and public bodies if they have access to information about the evidence, assumptions and principles on which policy decisions have been made. Citizens should have access to information about government policies that affect their lives.

> "When decision systems are introduced into public contexts such as criminal justice, it is important they are subject to the scrutiny expected in a democratic society. Algorithmic systems have been criticised on this front, as when developed in secretive circumstances or outsourced to private entities, they can be construed as rulemaking not subject to appropriate procedural safeguards or societal oversight."[17]
> **Law Society Report, Algorithms in the Criminal Justice System**

Access to this information also facilitates fair and informed public debate. Democratic society cannot make meaningful decisions if the government and the wider public sector are not open about their processes, capabilities and functions.

The use of artificial intelligence in the public sector may also change the relationship between citizens and the state. Surveillance technologies like facial recognition could increase the power of the state and other actors to monitor citizens' lives. Machine learning systems are also likely to shift the impetus of public service delivery from reaction and redress to prediction and prevention. These are not just questions of policy. These issues raise fundamental questions about democracy and human rights. The Committee considers that government openness about its use of AI is essential.

> "States should engage in inclusive, inter-disciplinary, informed and public debates to define what areas of public services profoundly affecting access to or exercise of human rights may not be appropriately determined, decided or optimised through algorithmic systems."[18]
> **The Council of Europe's draft Guidelines for States on actions to be taken vis-à-vis the human rights impacts of algorithmic systems**

While members of the public expressed a clear preference for openness in the focus groups held for this review, they also understood the need to judge the particular context in which AI is being used. Too much information was seen as being as unhelpful as too little. Openness does not mean that every detail around every use of AI in the public sector must be made public, and it may not be necessary or desirable to publish the source code for an AI system. Nonetheless, it is the view of the Committee that fundamental information about the purpose of the technology, how it is being used, and how it affects the lives of citizens must be disclosed to the public.

17 The Law Society (2019), Algorithms in the Criminal Justice System. Available at: https://www.lawsociety.org.uk/support-services/research-trends/documents/algorithm-use-in-the-criminal-justice-system-report/

18 Council of Europe (2018), Draft Recommendation of the Committee of Ministers to member States on human rights impacts of algorithmic systems'. Available at: https://rm.coe.int/draft-recommendation-on-human-rights-impacts-of-algorithmic-systems/16808ef256

2.3.2. How open is government about its use of AI?

Evidence submitted to this review suggests that at present the government and public bodies are not sufficiently transparent about their use of AI. Many contributors, including a number of academics, civil society groups and public officials said that it was too difficult to find out where the government is currently using AI. Even those working closely with the UK government on the development of AI policy, including staff at the Alan Turing Institute and the Centre for Data Ethics and Innovation, expressed frustration at their inability to find out which government departments were using these systems and how.

> "We are not aware of any body with systematic knowledge of where automated decision-making tools are being used in the public sector."
> **Centre for Data Ethics and Innovation**

The government does not publish any centralised audit identifying and making publicly available the extent of AI use across central government or the wider public sector. Public bodies rarely take a proactive approach to publishing information about their AI systems. Most of what we know is the result of work by journalists and academics who often have to rely on Freedom of Information Requests (FOIs), parliamentary questions or poorly formatted procurement data.[19]

Our evidence suggests that this lack of transparency is particularly pressing in policing and criminal justice. Many contributors said that they had trouble accessing important information about the use of new technologies in this area. This is particularly concerning given that surveillance technologies like automated facial recognition have the potential to undermine human rights.

> "There is a serious lack of transparency and concomitant lack of accountability about how the police and other law enforcement agencies are already using these technologies."
> **Professor Karen Yeung, Interdisciplinary Professorial Fellow in Law, Ethics and Informatics, University of Birmingham Law School and School of Computer Science**

Transparency is further complicated by the use of private sector commercial organisations in the development and provision of AI systems for use in the public sector, particularly as private companies may cite the need for commercial confidentiality to avoid certain forms of disclosure. This is concerning given that skills and resource constraints mean that public bodies are more likely to contract private companies to deliver AI services than to develop them in-house. In such cases, as elsewhere in public-private partnerships, the Principles of Public Life are binding on all those who provide services financed by public money.

19 Bureau of Investigative Journalism (2019), Government Data Systems: The Bureau Investigates. Available at: https://www.thebureauinvestigates.com/stories/2019-05-08/algorithms-government-it-systems

"Transparency – and therefore accountability – over the way in which public money is spent remains a very grey area in the UK…People are convinced that the growth of technology in the public sector has hugely important ramifications, but are baffled as to what exactly is going on and who is doing it."[20]

Dr Crofton Black, Government Data Systems: The Bureau Investigates, The Bureau of Investigative Journalism

This lack of transparency poses a clear risk to standards. If public bodies are not sufficiently open about where and how they are using these systems, the public will not be able to scrutinise and hold accountable those institutions which use AI in dubious or controversial ways. Without a well-informed public debate around AI, it is also likely that the UK will lack a common consensus on where and how the technology can be used for good in the public sector.

"Much of the public simply don't yet know enough about how AI or automation works, or where innovations might be used, to make an informed decision on whether they support or oppose them. This creates a vacuum of information, into which negative narratives about Britain's future are just as likely to take root as positive ones."[21]

Mark Kleinman, Professor of Public Policy and Director of Analysis at the Policy Institute, King's College London

2.4. Accountability

Holders of public office are accountable to the public for their decisions and must submit themselves to the scrutiny necessary to ensure this.

2.4.1. How can organisations be accountable when using AI?

Accountability is about holding individuals and organisations responsible for how any AI application is being used in the public sector. Yet applying the principle in this context is complicated by the fact that the outcome of an AI system will not simply be the product of the software itself, or any single decision-maker.

This is because the success or failure of an AI system may be the product of one or several components. In most cases, a system failure will be the result of multiple factors, and responsibility will not be easily apportioned.

All public officials responsible for part of an AI-assisted process will have a degree of professional accountability for their areas of responsibility. However, ultimately, accountability for AI systems lies with senior leadership who oversee AI projects and set governance for its effective and ethical use. It is senior leadership who should be held accountable if their staff are not sufficiently trained, if they have not implemented proper checks on the quality of data, or if they have approved the deployment of a system that prevents public officials from altering the basis on which AI makes a decision. Rather than lying with any single designer, system-builder, or operator of an AI system, accountability has to rest with those who choose to adopt and implement the system as part of their responsibility for public service delivery. It is the role of senior leadership to ensure that suitable governance is in place for any risks a system poses.

20 Same source

21 'People not robots are the key to the fourth industrial revolution', Mark Kleinman, accessed at: https://www.kcl.ac.uk/news/people-not-robots-are-the-key-to-the-fourth-industrial-revolution

Senior leadership should then be held accountable for decisions an AI system takes.

2.4.2. Should AI have full responsibility for decision-making in the public sector?

The issue of responsibility concerns how much human involvement there is in each individual decision taken by an AI system. As machine learning systems will be able to make decisions in many areas of public service delivery without any human involvement, the rise of AI means public bodies will need to assess the extent to which public officials should be involved in a decision-making process. It also raises the question of who or what is ultimately responsible for the outcomes of AI-enabled decisions. Both questions are key for accountability. Public office-holders and the public will need to understand how decisions are made and on the basis of what evidence.

It may become increasingly difficult to assign human responsibility to an automated decision-making process if AI can make decisions autonomously and automatically. In the future AI could be granted legal personhood and be held liable for its own decisions in the same way that a private company is. This would require a radical reworking of the law.

> "When you have a non-human decision-maker, can you always ascribe the outcome to a human? If you cannot then you have a gap where there is no legal liability. One could stretch existing laws around negligence and vicarious liability, but the more independently AI takes decisions, the harder it will be to tie decisions back to human beings."
> **Jacob Turner, Barrister and Author of Robot Rules: Regulating Artificial Intelligence**

Most experts consulted for this review rejected legal personhood for AI. Instead, policymakers, technologists and ethicists all told the Committee that retaining an element of human responsibility was a prerequisite for upholding high ethical standards in AI-enabled public services. This is a common theme in current AI ethics codes, which usually make explicit that AI should be human-centric, uphold human agency and respect human autonomy.

In polling undertaken for this review, there was also a clear preference for upholding a degree of human responsibility in automated decision-making in the public sector. 69% of those polled said that they would be more comfortable with a public body using AI if a human was using their professional expertise to make a final judgement on any decision. Participants in focus groups also took this view and said that the lack of human involvement in a decision-making process would be unnerving.[22]

Retaining a degree of human involvement and responsibility for automated decision-making is also likely to help uphold public standards in practice. There will be more of an incentive for public officials to monitor and check their AI systems if an official has to answer to the public for the outcome of an automated decision. The Committee therefore believes that public officials should be in control of AI, retain some involvement in all automated decision-making processes, and take responsibility for decisions made by AI systems.

The extent and nature of this responsibility will vary. The Committee heard about a number of models for upholding human responsibility in automated decision-making. Some contributors told the Committee that forcing public officials to directly intervene in all simple, automated decisions was neither fair nor plausible, particularly where intervention would be unnecessary or obstructive.

22 Appendix 3, 81

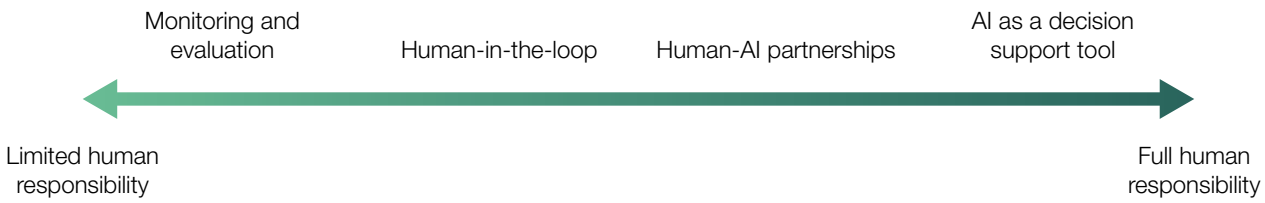
More limited forms of oversight, such as monitoring and evaluation, would likely still be necessary in this context, and would allow public officials to identify and remedy potential problems within the system. This was seen as a fairly limited form of oversight, as an automated decision could still occur without significant human involvement.

The Committee heard from some experts that 'human-in-the-loop' models are helpful for retaining a degree of human control over an automated decision-making system. In a 'human-in-the loop' system, a public official can intervene in the decision-making process of a machine. This means that AI works more or less autonomously, but that a human can observe how different variables are weighted and intervene where necessary. In these systems, humans are likely to be involved in the training process of the algorithm, continuously testing and tuning the data in order to achieve better results.

Others rejected the concept of 'human-in-the-loop' as too AI-centric. Some contributors argued that these models portrayed AI as immutable, because they ask individuals to shoehorn human judgement into machine learning systems so that human responsibility is protected. This was seen as inverting a 'human-centric' approach. Contributors spoke instead of a more interactive partnership between human and machine, where the outcome of an automated decision is the equal product of human and AI involvement.[23] This was seen as useful for eradicating the potential flaws of both human and AI decision-making processes, and for enhancing the quality and accuracy of the decision as a whole.

> "Rather than focusing on the concept of humans-in-the-loop, we need to think carefully about the end-to-end process and ensure that we think about how AI and humans work together to deliver efficiencies and better results."
> **Sana Khareghani, Head, Office for AI**

Those who saw the relationship between human and machine in this way suggested that there should be an element of human control at every stage of the AI process, from design, through procurement, to the deployment of an AI system. Contributors suggested that this whole-systems approach would help mitigate the risk of an accountability gap, where it is unclear which public officials, if any, are responsible for an automated decision.



23 N. R. Jennings, L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden and A. Rogers (2014) "On human-agent collectives" Communications of ACM 57 (12) 80-88

> "If you are saying that there may be some decisions that need to be made so rapidly that the machine makes the decision (if it has been appropriately codified), there is still human accountability at the design stage and in the verification and validation of the AI system before it is put into use. This means you may not have an accountability gap as ultimately a human is still accountable at the design and testing stages."
> **Fiona Butcher, Fellow, Defence Science and Technology Laboratory, Ministry of Defence**

Many contributors took the view that AI should not retain any role in making a final decision, particularly where the adverse effects on an individual could be significant. They suggested instead that AI should be thought of as a decision-support tool, rather than a decision-making system. For example, an AI system that identifies a malignant melanoma should not be seen as making a decision in conjunction with a medical professional, but as making a recommendation to a doctor who retains final discretion on diagnosis and treatment.

Models for upholding human responsibility can be placed on a spectrum, from humans having limited to full responsibility for an automated decision, as shown below. Senior leadership will have to choose which level of responsibility is most appropriate for the application of AI in their organisation (see chapter 5).

2.4.3. Should public bodies provide explanations for AI decisions?

Automated decision-making systems usually work by analysing large quantities of data to find patterns and correlations between variables and outcomes. These patterns are used to generate insight, which is used to inform a recommendation or a decision. While the outcome of an AI system will be clear, the process by which it comes to a decision will often not be. The Committee heard that some more complex forms of machine learning cannot show how they determine which variable caused which outcome, or in some cases, what those variables are. AI systems that are opaque in this way are often referred to as 'black boxes'.

> "The fact that we cannot always explain how an AI system made a decision and whether that process was adequate challenges public servants' ability to make decisions in an open and transparent manner."[24]
> **Leverhulme Centre for the Future of Intelligence, University of Cambridge**

The term 'explainability' is typically used to describe the extent to which an AI system's decision-making process can be understood. The levels of explainability necessary will vary across systems, but to uphold accountability, an AI system will need to provide some kind of explanation of its decision-making process.

It was the view of most contributors to this review that the black box problem is largely an avoidable issue. The Committee heard that most machine learning systems deployed in the public sector will be processing data held in simple, readable formats, such as demographic data. Here, a less complex, more explainable system could be used as it would deliver results of a similar level of accuracy to more complex, unexplainable systems.

24 Written evidence 9 (Leverhulme Centre for the Future of Intelligence, University of Cambridge)

> "If you stick with a simpler model which is inherently interpretable, you are not going to sacrifice that much on accuracy but you are going to keep the benefits of understanding the variables you are using and understanding how the model works."
> **Dr Reuben Binns, Postdoctoral Research Fellow in AI, ICO**

Very complex systems, such as those based on neural networks, can make it hard to follow the logic of the system. However, in such cases, there may be a trade-off between accuracy and explainability. The Committee heard that these technologies are unlikely to be used in the near future in the public sector. Where these systems are used, public officials will need to be able to justify why their need for such complex systems outweighs the requirement for transparency.

Many contributors saw public officials and private companies choosing not to provide an explanation as a greater obstacle than technical capability. The most significant risk for public standards is that officials and companies will fail to include provisions for explainability when designing their systems even though it would be technically possible to do so.

> "I think something we need to be challenging ourselves on is whether the lack of transparency and the lack of explainability is a real necessity for the system or whether it is bad design…sometimes there is a challenge to be made of vendors and people who are building the system."
> **Simon McDougall, Executive Director, Technology Policy and Innovation, ICO**

Our evidence showed this could be for multiple reasons. Building in provisions for explainability could increase the cost of the system.

The Committee also heard that private providers of public services may not want to reveal the intricacies of their systems in order to protect their intellectual property rights and commercial secrets. This 'commercial black box' was cited by some as a greater obstacle to transparency than technical opacity.

Private companies developing AI software consulted for this review told the Committee that they often had the capability to make their products and services more explainable, but that they were rarely asked to do so by those procuring technology for the public sector. The Committee was told that requirements for technical transparency are not usually included in procurement tenders and contracts.

> "Claims about what is technically (im)possible should be treated with caution. Our engagement with industry to date suggests that, if a degree of explainability is made a priority from the outset by its commissioner, it can be built in."[25]
> **Centre for Data Ethics and Innovation**

Overall, the evidence submitted to this review suggests that technical obstacles to public bodies providing explanations for AI-enabled decisions are currently small. It should be possible for citizens to obtain meaningful explanations in policy areas as diverse as healthcare, policing and social care. To achieve this, explainability needs to be considered in the early stages of project development and design, and during procurement processes, by those commissioning the technology for use in the public sector.

25 Written evidence 18 (Centre for Data Ethics and Innovation)

If explanations are provided, AI could present an opportunity to enhance accountability and openness in public services. Understanding the reasons behind human decision-making is often fraught with difficulty. It will not always be possible to understand, for example, how a public official came to their decision about a benefits claimant, or why their judgement was correct. When AI is used alongside human judgement, it may help provide greater clarity over which variables informed a decision.

Given that public bodies will, more often than not, be able to provide explanations for AI decisions, the key question is how, when and to whom explanations should be provided. Public bodies should note that the provision of an explanation appeals to the general public. In polling done for this review, 51% of those polled said that the provision of "an easy-to-understand explanation for the AI software's decision" would make them "much more comfortable" or "a bit more comfortable" with using AI in the public sector.[26]

> "The incorporation of an AI tool into a decision-making process may come with the risk of creating 'substantial' or 'genuine' doubt as to why decisions were made and what conclusions were reached…consideration should be given to the circumstances in which reasons for an explanation of the output may be required."[27]
>
> **Marion Oswald, Senior Fellow in Law and Director of the Centre for Information Rights, University of Winchester**

Overall, the Committee heard that the type of explanation necessary, or indeed whether an explanation was needed at all, was dependent on context. High impact decisions, such as those that have the potential to affect a citizen's rights or grant access to a service, are more likely to require clear explanations that give an account of the rationale, reasons and individual circumstances for a specific automated decision. Low-impact decisions, such as those made to increase administrative efficiency, are less likely to require an explanation beyond a statement that outlines the general functionality of an automated system. The ICO's Project Expl*AI*n report also found that in some contexts, for example in healthcare, accuracy was seen as more important than explainability.[28]

The Committee heard that there were also valid reasons not to disclose how an AI system came to a decision. There were concerns about individuals being able to manipulate systems for desired outcomes if public bodies were too transparent about what variables were used to inform a decision. A regulator, for example, may not want to provide an explanation for an AI system used to identify non-compliance with the law in case companies learn how to evade detection of non-compliance.

These concerns may well be valid reasons not to provide explanations to service users in certain contexts. However, those reasons must be shown to be legitimate and not an excuse to implement unexplainable systems. The burden of proof should always be on a public official to justify in clear terms why the benefits of explainability are outweighed by the possible detriment disclosure could cause. In such cases, this follows the principle of openness.

26 Appendix 3, 81

27 Written evidence 4 (Marion Oswald)

28 The Information Commissioner's Office (ICO) (2019), Project Explain, Interim report. Available at: https://ico.org.uk/media/about-the-ico/documents/2615039/project-explain-20190603.pdf

2.5. Objectivity

Holders of public office must act and take decisions impartially, fairly and on merit, using the best evidence and without discrimination and bias.

2.5.1. Data bias

Data can be collected on multiple aspects of our world, from the speed of a car to somebody's personal preferences. Millions of data points together can reflect more complex scenarios, such as city-wide traffic jams or the voting habits of demographic groups.

But data may not always be representative. It is well understood that AI has the potential to produce discriminatory effects if a data set is in some way flawed or an algorithm operates in a biased way. Machine learning identifies patterns in past data and makes current decisions based on those patterns, so AI systems have the potential to entrench or amplify historic biases. AI systems could exacerbate biases against protected characteristics, such as race or sex, and make discriminatory outcomes against those characteristics more likely.

Imagine a machine learning system deployed by a company to filter job applications by scanning individuals' CVs. The system has 'learned' what makes a successful applicant by processing the CV data of past successful applicants – when the recruitment process was run by humans – to determine what each successful applicant has in common. In theory, the system should identify educational qualifications, relevant experience, and seniority in previous roles as key criteria. It should then use these criteria to filter new applications.

But this theory only holds if educational qualifications, relevant experience and seniority were the determinants of successful applications when humans ran the company's recruitment process. If those humans were biased themselves – say, for example, they favoured male applicants over female applicants – the machine learning system would inevitably replicate that bias. This was the case with an Amazon machine learning system developed in 2014, "which effectively taught itself that male candidates were preferable".[29]

Sampling errors can also produce discriminatory outcomes. A machine learning tool designed to diagnose skin cancer that has only been trained on white skin will be less accurate on other skin colours. This bias in the training data may not be the result of active human prejudice, but it will result in a discriminatory outcome: the system is more likely to misdiagnose BAME people.

> "There is a very old adage in computer science that sums up many of the concerns around AI enabled public services:
>
> 'Garbage in, garbage out'
>
> In other words, if you put poor, partial, flawed data into a computer it will mindlessly follow its programming and output poor, partial, flawed computations. AI is a statistical-inference technology that learns by example. This means if we allow AI systems to learn from 'garbage' examples, then we will end up with a statistical-inference model that is really good at producing 'garbage' inferences."[30]
> **British Computer Society**

29 Maya Oppenheim (2018), 'Amazon scraps "sexist AI" recruitment tool', Independent. Available at: https://www.independent.co.uk/life-style/gadgets-and-tech/amazon-ai-sexist-recruitment-tool-algorithm-aw8579161.html

30 Written evidence 11 (British Computer Society)

Ultimately, AI systems are only as good as the data we put into them. 'Bad' data can contain implicit and explicit racial, gender or ideological biases. When this data is used to train machine learning algorithms, these biases find their way into the AI systems we design, which can result in discriminatory decisions. The Committee heard that machine learning systems could potentially discriminate on the basis of any variable it identifies when processing data.

> "Decision-making, algorithmic or otherwise, can of course also be biased against characteristics which may not be protected in law, but which may be considered unfair, such as socio-economic background. In addition, the use of algorithms increases the chances of discrimination against characteristics that are not obvious or visible. For example, an algorithm might be effective at identifying people who lack financial literacy and use this to set interest rates or repayment terms."[31]
> **Centre for Data Ethics and Innovation, Interim Report on Data Bias**

2.5.2. How will data bias affect objectivity?

The principle of objectivity requires government and public bodies to act and take decisions impartially, without discrimination or bias. Data bias therefore poses a direct risk to public standards. The introduction of automated machine learning systems in areas such as policing, immigration and healthcare risks inadvertently introducing or amplifying discrimination in sensitive policy areas. From a standards perspective, and in the eyes of the law, discrimination via algorithm is no less of an offence than discrimination by a public official.

> "The statistics speak for themselves. We know that you are eight times more likely to be subject to stop and search in the UK if you are black. If you are building an algorithm on these statistics, that is a huge problem."
> **Sandra Wachter, Associate Professor and Senior Research Fellow, Oxford Internet Institute**

Civil liberties groups fear discriminatory machine learning systems are already in use in the UK. In June 2019, the Financial Times reported that the Home Office used an algorithmic tool to stream visa applications. The potential for the streaming tool to replicate historic bias is clear: if officials had previously discriminated against applicants from certain countries, the streaming tool would do so too.[32]

Data bias is a well-known phenomenon that frequently features in the media. In focus groups undertaken for this review, there was evidence that the public are aware of the issue. When given the example of predictive policing software, participants immediately mentioned the risk of biased profiling, despite assumptions that computers are inherently neutral or objective. For most participants data bias was seen as a more significant issue than a lack of explainability or human responsibility. Of the three issues, data bias appeared to have the greatest potential to delegitimise the use of AI in the public sector in the eyes of the general public.

However, policy experts often qualified negative perceptions of data bias with three considerations. They made the point that data bias can be used to measure and reveal discrimination in existing public sector practices. Often, marginalised groups complain of systematic discrimination from public

31 Centre for Data Ethics and Innovation (CDEI) (2019), Interim Report: Review into bias in algorithmic decision-making. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819168/Interim_report_-_review_into_algorithmic_bias.pdf

32 Helen Warrell (2019),'Home Office under fire for using secretive visa algorithm', Financial Times. Available at: https://www.ft.com/content/0206dd56-87b0-11e9-a028-86cea8523dc2

bodies, but lack the statistical tools to measure bias. Machine learning tools will probably be able to reveal that discrimination, and potentially expose new, unknown biases.

> "Some of our existing systems are designed in a way that makes it impossible to measure bias...One of the good things about machine learning technologies is that they have exposed some bias which has always been there."
> **Professor Helen Margetts, Professor of Society and the Internet at the University of Oxford and Director of the Public Policy Programme, The Alan Turing Institute**

Second, it was frequently mentioned that AI systems will be no more biased than the human processes they are replacing. For some, this meant criticism of the discriminatory impact of AI systems was overblown, as AI systems were not likely to be significantly worse than what already exists.

> "Right now we are more likely to be replacing a human process with an AI process. All us humans are bringing a whole suitcase of preconceptions, prejudices and baggage along with us to that decision, some conscious and some unconscious. As we talk around bias in AI – and there is plenty of stuff to talk about – we have to keep in mind we are not moving from a beautiful neutral model."
> **Simon McDougall, Executive Director, Technology Policy and Innovation, ICO**

Third, the Committee heard that if technologists can successfully identify and minimise bias, AI has the potential to be more objective than humans currently are. Contributors cited a famous 'hungry judge' study, which found that judges were more likely to issue harsher decisions just before lunch.[33] AI systems on the other hand do not get hungry. Though confidence in the possibility of eradicating data bias was mixed, some contributors said they could foresee a future where, in some areas, a duty of objectivity could require public bodies to use AI systems rather than humans.

> "I think we have to start from the point of view that we are dealing with biased systems usually anyway. It is one of the hopes of artificial intelligence that it might be able to reduce bias in certain areas and, certainly, provide lots more ways of systematically thinking about measuring that bias."
> **Dr Jonathan Bright, Senior Research Fellow, Oxford Internet Institute**

2.5.3. Mitigating and managing data bias

AI experts suggested a range of methods to manage data bias. Chief among these was the need to ensure diversity in AI teams. A workforce composed of a single demographic is less likely to check for and notice discrimination than diverse teams. At every stage – from the design of a product to its deployment – diversity was seen as a necessity. The Committee heard that while data bias may create discrimination, a lack of diversity will facilitate it.

33 Zoe Corbyn (2011), 'Hungry judges dispense rough justice', Nature, International Weekly Journal of Science. Available at: https://www.nature.com/news/2011/110411/full/news.2011.227.html

"There will be new jobs for humans to work out what machines are doing. And this is where it comes back to diversity – those humans in the loop must be diverse, so they can see the true range of possible impacts the machine is having."
Professor Dame Wendy Hall, Regius Professor of Computer Science, University of Southampton and co-author, UK government AI review

Others suggested that public officials should also be expected to alter or limit the scope and powers of an AI system when it displays a high degree of bias. The impact of bias on the general public can be reduced, for example, by removing a system from the front-line of service delivery. Given the risk to both public trust and public standards, officials should be prepared to remove an AI system entirely if it persistently produces biased results.

While it was seen as implausible to prohibit any system displaying bias, public bodies should always know how their systems are biased and who is most affected by that bias. Once the bias of a system is known, suitable remedial action and mitigation procedures should follow.

"What we might want to say is 'it is unacceptable not to know the ways in which your system is biased, and you are then required to account for how you use and understand the results of that system in that context.' You need to be able to provide a justification and that justification has to be subject to scrutiny and challenge."
Oliver Buckley, Executive Director, CDEI

Data scientists consulted for this review also outlined a number of technical methods that could be deployed to mitigate bias, while voicing scepticism that any AI system could be completely free of bias. As a matter of good practice, technologists recommended programming systems to exclude characteristics like race, gender, or age from predictive models. However, it was widely accepted that this would only have a limited impact. This was due to challenges around proxy characteristics. You could strip out ethnicity, for example, but location could act as an effective proxy, resulting in the same discriminatory effects.

"A draft tool we have looked at (at West Midlands Police) had intelligence information built in as input factors, including things like the number of stop and search counts, and that raised red flags around what that could be a proxy for in that particular region."
Marion Oswald, Senior Fellow in Law and Director of the Centre for Information Rights, University of Winchester

Proxy characteristics can also be extremely subtle and not easily identifiable. A predictive policing model used to predict the likelihood of criminals reoffending could use natural language processing to analyse police interviews. That model could identify a defendant's defensive response to questioning as an indicator of a propensity to reoffend. However, a defensive tone may instead be a response to a more aggressive line of questioning from the interviewing police officer, and police officers may, historically, have been more likely to ask more aggressive questions of male and ethnic minority interviewees.

Furthermore, stripping out certain characteristics may not make a system less biased. In the reoffending model outlined above, one answer could be to strip out gender as a likely predictor of reoffending rates. However, as the CDEI state, "Blinding algorithms to demographic differences and proxies for these differences does not always lead to fairer outcomes...Preventing an algorithm designed to calculate the risk of criminals reoffending from taking into account their sex, would likely result in disproportionately harsher sentences for women overall as women tend to reoffend less often than men. By excluding sex, the algorithm becomes less accurate for women and so, arguably, less fair."[34]

> "I'm not convinced that human cleansing of data adequately answers this problem. When we remove certain data points, how are we sure that we are making a dataset less biased? Whose rules are being used, why and who is saying that those rules are the right ones?"
> **Sana Khareghani, Head, Office for AI**

Some suggested that a better solution was to increase the size and diversity of datasets. Overall, however, there was recognition that more research was needed into technical solutions to data bias.

34 Centre for Data Ethics and Innovation (CDEI) (2019), Interim report: Review into bias in algorithmic decision making, 11. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819168/Interim_report_-_review_into_algorithmic_bias.pdf

Chapter 3: Guidance and ethical principles

3.1. Introduction

While AI poses particular risks and opportunities for openness, accountability and objectivity, there is nothing inherently new about what is needed to govern and manage AI responsibly. The Committee found that artificial intelligence does not necessitate a fundamental reworking of public sector practice. Successful AI governance is a question of clear regulation and proper controls for understanding, managing and mitigating risk.

In this sense, AI is a new challenge that can be solved with effective governance and a traditional risk management approach. The senior leadership of public bodies will first need to assess the risk an AI tool poses to public standards. They will then need to set governance mechanisms that mitigate that risk to a level deemed acceptable for the context AI is used in. Senior leadership will need to justify and be ultimately accountable for any risk mitigation measures their organisations take. By implementing the right processes, policies and management structures, public bodies will remain accountable, open and objective when using AI.

Public sector organisations will not, however, be able to establish sound governance alone. AI poses new challenges around issues such as explainability and responsibility that public sector organisations will not encounter when using conventional digital systems. Public bodies will need clear guidance based on sound ethical principles on how to adapt their governance and management structures for AI. To this end, the Office for AI, the Government Digital Service, and the Turing Institute collectively published A Guide to Using Artificial Intelligence in the Public Sector ('the AI Guide'), a comprehensive set of guidance for public bodies to use.[35]

Separately, the ICO has published its AI Auditing Framework. In collaboration with the World Economic Forum, the Office for AI has also produced specialist guidelines on AI procurement. This chapter assesses the quality, practicality and accessibility of guidance issued so far.

With the establishment of the Office for AI and the CDEI, and the designation of the Alan Turing Institute as the UK's national centre for AI, the UK government has signalled its ambition to become a world leader in responsible innovation. It should be noted that the issuing of the AI Guide is the most significant piece of work published towards this goal. However, the Committee's view is clear: guidance alone is not enough, and clear, well-established regulation is needed to ensure the responsible use of AI in the public sector. The form that AI regulation could take is discussed in chapter 4.

3.2. Ethical principles

> "A hallmark of good governance is the development of shared values, which become part of the organisation's culture, underpinning policy and behaviour throughout the organisation, from the governing body to all staff."[36]
>
> **The Independent Commission on Good Governance**

35 Government Digital Service, Office for AI and the Alan Turing Institute (2019), A Guide to Using Artificial Intelligence in the Public Sector. Available at: https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

36 The Independent Commission on Good Governance (2004), The Good Governance Standard for Public Services, 13. Available at: https://www.jrf.org.uk/report/good-governance-standard-public-services as quoted by this Committee in its report Standards Matter, 22

Ethical principles underpin good public sector governance. AI is no exception. Establishing a clear set of ethical principles covering the use of AI in the public sector reminds public officials to consider public standards when using AI, and to choose a course of action that best adheres to those principles. As artificial intelligence will fundamentally change the way public services are delivered, the public sector needs a clear set of ethical principles specific to the challenges posed by AI.

The final section of the AI Guide, 'Using AI ethically and safely', begins on this basis, establishing a new set of values and principles, known as the FAST Track Principles and the SUM Values.[37] The SUM Values "support, underwrite, and motivate a responsible innovation ecosystem" by outlining the values that underpin the ethical permissibility of an AI project. Those values are respect, connect, care, protect. The FAST principles guide the design and use of AI systems. They are fairness, accountability, sustainability and transparency.

The establishment of ethical principles specifically for AI in the UK public sector is welcome. Academics estimate that over 70 AI ethics codes have been published over the past three years, and contributors emphasised the risk of 'ethics-shopping', where, as Professor Luciano Floridi argues, "private and public actors may shop for the kind of ethics that is best retrofitted to justify their current behaviours, rather than revising their behaviours to make them consistent with a socially accepted ethical framework."[38] A single statement of AI ethical values is a significant step forward in solving this problem.

The SUM Values provide a good starting point for public officials debating whether or not to introduce AI. AI will create new possibilities in prediction, automation and analysis, so it is important that public sector organisations examine the ethical permissibility of their project before deciding to procure or build an AI system. Multiple contributors warned against public bodies taking a "shiny new tool" approach to AI where projects were embarked on without consideration of their long-term social, ethical and environmental impact.

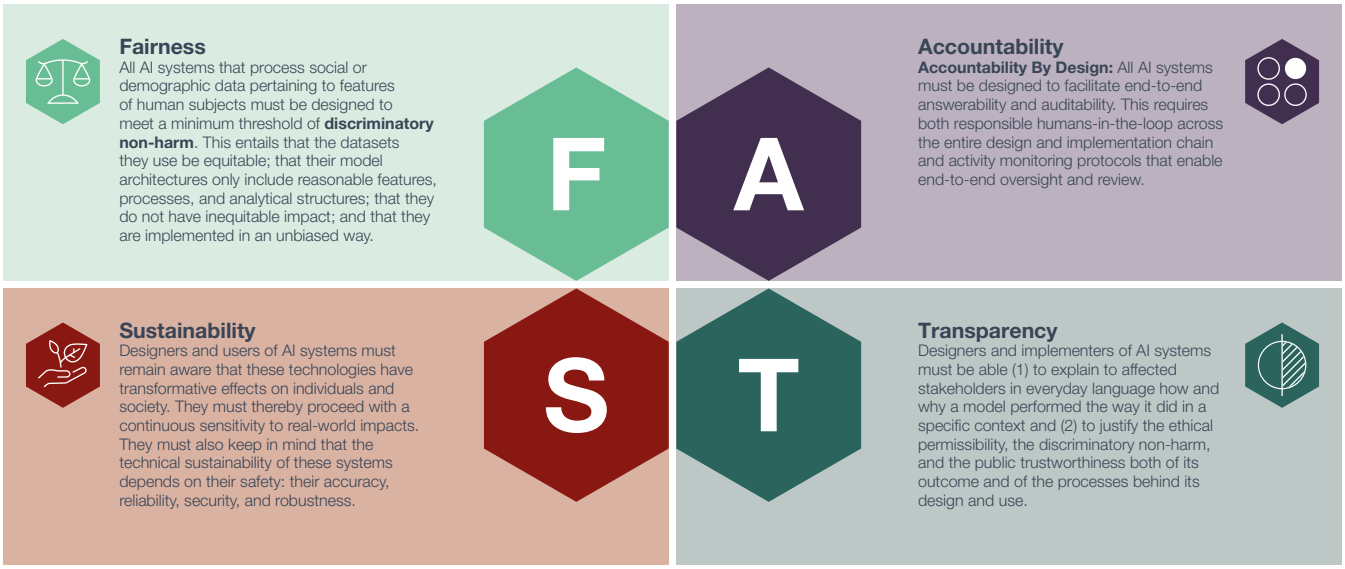
The FAST principles provide a clear, actionable and appropriate guide for public sector behaviour. Fairness, accountability, sustainability and transparency comprehensively cover the five areas of concern identified in chapter two of this review, and complement, rather than contradict, the Seven Principles of Public Life.

The FAST principles certainly need greater distribution and promotion across public life. Many contributors called for a 'Super-Code' of ethical principles and the FAST principles could provide this. However, in order for an overarching set of AI ethical principles to gain traction across the public sector, the principles should be promoted more prominently, and the descriptors should be shorter and clearer. The principles should also be made explicit in all sector-specific AI codes of conduct.

37 Dr David Leslie (2019), Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. Available at: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

38 Luciano Floridi (2019), 'Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical', Philosophy & Technology, 32:185. Available at: https://link.springer.com/article/10.1007/s13347-019-00354-x

FAST Track Principles



In order to become authoritative, the FAST principles must live outside of the AI Guide. In the way that the Seven Principles of Public Life are the defining mission of this Committee, upholding the FAST principles could become the overarching goal of the Office for AI, CDEI, and Turing Institute's public sector work. For principles to shape institutions, they need to be integrated into public sector cultures.

This is all the more urgent as the establishment of the FAST principles has not solved the problem of ethics shopping, outlined above. Currently there are three sets of ethical principles endorsed by UK government bodies. As well as the AI Guide, the Department for Digital, Culture, Media and Sport (DCMS) published the Data Ethics Framework (DEF), which prescribes a number of useful values and practices, while the Centre for Data Ethics and Innovation (CDEI) adheres to OECD Principles on Artificial Intelligence, which the government has also adopted. Each of these three sets has a different focus, and some are more high-level than others, but this multiplicity of principles and codes confuses the landscape and undermines attempts to make any set of ethical principles authoritative. It is also unclear how they work together. For example, although the AI Guide mentions that its principles are intended to supplement the Data Ethics Framework, it is unclear how they work together in practice.

Elevating the reach and status of an authoritative set of principles is also necessary given the prominence of private companies in AI-enabled public service delivery. Private providers may have their own lists of ethical principles that are inappropriate for public service delivery, or may exploit ambiguities in the higher-level and less focused principles adopted by government bodies.

It is noted that the Data Ethics Framework is currently under review, and that the AI Guide is intended as an iterative document. The government should use this opportunity to identify, endorse and promote an authoritative high level set of ethical principles. The public should be able to find easily a clear statement of ethical principles that govern the use of AI in the public sector, and it should be made clear to those on the frontline of service delivery which ethical principles public officials are expected to adhere to. This should include outlining the purpose, scope of application and respective standing of each of the three sets currently in use.

Recommendation 1a:

There are currently three different sets of ethical principles intended to guide the use of AI in the public sector – the FAST SUM Principles, the OECD AI Principles, and the Data Ethics Framework. It is unclear how these work together and public bodies may be uncertain over which principles to follow.

The public needs to understand the high level ethical principles that govern the use of AI in the public sector. The government should identify, endorse and promote these principles and outline the purpose, scope of application and respective standing of each of the three sets currently in use.

Clear and authoritative ethical principles then need to be further elaborated and specified in codes of conduct that are explicit about what is expected of public office-holders in different contexts. It is likely that sector-specific AI ethics codes will be necessary, particularly in high-risk policy areas such as policing, criminal justice, health and social care. Sector-specific codes can help make abstract ethical principles clearer and more tailored to particular professional settings, while retaining the link to the standards expected of public office-holders across the whole of the public sector.

The Code of Conduct for Data-Driven Health and Care Technology produced by the Department of Health and Social Care (DHSC) is a good example of best practice, which integrates AI ethics with pre-existing medical and public standards.[39]

Codes of conduct that elaborate what the principles imply in particular organisations ensure that everyone in the organisation knows what is expected of them. They also inform those holding them to account. This is useful where the application of principles may not be self-evident and where it remains unclear how public officials will uphold these ethical principles in practice. The DHSC code, for example, asks public office-holders to "Generate clear evidence of the effectiveness and economic impact of a product or innovation...an evidence-generation plan should be developed using the evidence standards framework published by the National Institute for Health and Care Excellence (NICE)."[40]

Such codes should not, however, override principles. Behaviour of public office-holders can technically be within the rules set out in a code and yet still offend against underlying principles and values expected of them by the public. Principles and codes must be viewed as complementary rather than alternatives.

Public officials should also be aware that ethics is an ongoing and dynamic practice. Principles and codes serve as a useful guide for those looking to make the right judgement in a particular context, but they are not a substitute for comprehensive ethical risk management. Government departments should be aware that establishing and promulgating an AI ethics code is the beginning, and not the end, of effective AI governance.

39 Department of Health and Social Care (2019), Code of conduct for data-driven health and care technology. Available at: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology

40 Same source

3.3. The Office for AI, GDS and the Turing Institute's 'Understanding artificial intelligence ethics and safety' – The AI Guide

Best practice and guidance is a vital part of any framework designed to uphold ethics. Public bodies will need guidance that shows how ethical principles for AI translate into practice. The Office for AI, the Government Digital Service (GDS), and the Turing Institute published the AI Guide in August 2019. The guidance helps public bodies understand AI, then covers three stages of an AI project, and ends with a longer document on AI ethics.

This guidance is welcome, and it is a comprehensive and valuable resource for public bodies wanting to implement AI while upholding high public standards.

A Guide to Using Artificial Intelligence in the Public Sector

Understanding artificial intelligence

Assessing if AI is the right solution for your users' needs

Planning and preparing for artificial intelligence implementation

Managing your artificial intelligence project

Using artificial intelligence ethically and safely.[41]

Importantly, ethical standards are not restricted to the final section of the AI Guide. The guidance emphasises that ethics must be considered at every stage of an AI process, from assessing if AI is the right solution, through project planning, to system management. Public bodies using this guidance should ensure they follow every section, rather than the section on ethics and safety alone.

The section most relevant to this review is 'Using artificial intelligence ethically and safely', produced by the Turing Institute and summarised on the Office for AI website. This guidance reflects a number of key points made by contributors to this review: AI ethics (and therefore ethics-based governance) is heavily context-specific; ethical principles must be actionable; and ethics-based governance is a continuous process, rather than a one-time event. The guide's integration of ethical issues into a process-based governance (PBG) framework is laudable, and reflects a core conclusion of this review: that high public standards are a product of good governance.

It remains to be seen if the guidance will have a significant impact on AI in the public sector. We were informed that future iterations of the AI Guide would be subject to more extensive publicity. This is vital. Guidance, no matter how good, will leave no mark on the landscape without extensive measures to promote its adoption.

Future iterations of 'Using artificial intelligence ethically and safely' must also be made easier to use and understand. The full document is nearly 100 pages in length and assumes a level of technical awareness above what can be reasonably expected of senior leadership in a local council, school or police force. This undermines the practicality of the guidance, especially as leadership – those setting governance – is the intended audience for this guide. In its current form, implementing the guidance would require oversight from specialist AI policy professionals, but this is not a resource many public sector organisations will have. As it stands, the ethics and safety guidance would work better as a source document for sector-specific guidance and best practice, rather than an authoritative guide for all public sector organisations to follow.

41 Government Digital Service, Office for AI and the Alan Turing Institute (2019), A Guide to Using Artificial Intelligence in the Public Sector. Available at: https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

Good guidance has the potential to change behaviours and shape professional cultures. It is always an important part of any standards regime. Guidance is, however, a non-binding tool. The issuing of good guidance does not constitute the formalisation of public standards for AI. Using Artificial Intelligence in the Public Sector is a high-quality set of guidance, as is 'Using artificial intelligence ethically and safely'. But further work on promotion and accessibility is needed to ensure this guidance has the greatest effect.

Recommendation 1b:

The guidance by the Office for AI, the Government Digital Service and the Alan Turing Institute on using AI in the public sector should be made easier to use and understand, and promoted extensively.

Contributors did question if the Office for AI and GDS were the right organisations to issue guidance, given a potential conflict of interest between promoting AI adoption and upholding ethical standards.

> "The guidelines and advice are the shared responsibility of the Office for AI in BEIS, and the Government Digital Service. The OAI is also responsible for promoting the development of AI technologies and industries, and so has a conflicting interest, and the GDS has wide responsibilities to support digitalization of central government. It seems unlikely that either organisation has the capacity or remit to ensure robust and consistent ethical supervision on broader questions of automated decision system adoption and use in public policy, including their use outside central government."
> **Dr Emma Carmel, Associate Professor, Social and Policy Sciences, University of Bath**

This concern, while valid, does not fully reflect the nature of the Office of AI and GDS, both of which have shown a clear commitment to ethics as well as adoption. This is reflected in the guidance, which makes clear that AI is a data science tool of limited utility for addressing specific problems, and not a universal solution to any public policy challenge.

3.4. The ICO's auditing framework for AI

The ICO has also issued guidance relevant to public standards and AI through its auditing framework, which provides information on how data processors can ensure compliance with data protection requirements for AI under the GDPR.[42] It covers issues of fairness and transparency that mirror this Committee's concerns.

42 ICO (2019), An overview of the Auditing Framework for Artificial Intelligence and its core components. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

Explanations and guidance on each individual risk area is given in a series of blog posts. These are useful, provide clear instructions on how to mitigate risk, and are written in an accessible way for those without technical expertise. It is an important resource for public bodies wanting to uphold public standards in the areas covered. ICO blog posts on fully automated systems and bias and discrimination provide the best user-orientated guidance on these topics seen in the course of this review.

In its analysis of fully automated decision-making models, ICO guidance states:

- human reviewers must be involved in checking the system's recommendation and should not "routinely" apply the automated recommendation to an individual
- reviewers' involvement must be active and not just a token gesture. They should have actual "meaningful" influence on the decision, including the "authority and competence" to go against the recommendation
- reviewers must "weigh-up" and "interpret" the recommendation, consider all available input data, and also take into account other additional factors.[43]

Proposed framework

1. Governance and accountability

Cross-cutting focus areas	Risk appetite	Leadership engagement and oversight	Management and reporting structures	Compliance and assurance capabilities
	Data protection by design and by default	Policies and procedures	Documentation and audit trails	Training and awareness

2. AI-specific risk areas

Fairness and transparency in profiling	Accuracy	Fully automated decision making models	Security and cyber
Trade-offs	Data minimisation and purpose limitation	Exercise of rights	Impact on broader public rights*

*Includes only considerations with scope of an ICO investigation/audit

43 ICO (2019), Automated Decision Making: the role of meaningful human reviews. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-automated-decision-making-the-role-of-meaningful-human-reviews/

Public sector organisations should be aware that the auditing framework for AI is constrained by the ICO's remit, and that the GDPR is not a perfect fit for all public standards concerns. In particular, contributors cautioned against public sector organisations assuming that a Data Protection Impact Assessment covers all ethical issues; specialist AI impact assessment will also be necessary (see chapter 4). Nonetheless, public sector organisations should be confident that the auditing framework provides an authoritative steer on the ethical issues covered by the GDPR, including explainability, automation and responsibility, and bias.[44]

3.5. The Office for AI's guidelines for AI procurement

The Draft Guidelines for AI Procurement published by the Office for AI and the World Economic Forum in September 2019 outline a number of useful ways for public bodies to compel private providers of public services to consider ethics. These include referencing ethical principles such as the government's Data Ethics Framework in invitations to tender, developing strategies to address the ethical limitations of training data, and using ethical considerations as evaluation criteria.[45] The guidelines are a work in progress and undergoing trial at the time of writing.

There are key robust practices you can ask for suppliers to demonstrate when providing AI solutions. The guidance for understanding AI ethics and safety provides a useful framework to identify those. Besides having an ethical framework within their company, robust practices include:

- having an internal AI ethics approach, with examples of how it has been applied to design, develop, and deploy AI solutions
- processes to ensure accountability over outputs of algorithms
- avoiding outputs of analysis which could result in unfair decision making

The Office for AI's Draft Guidelines for AI procurement.[46]

44 ICO (2019), Project Explain, Interim report. Available at: https://ico.org.uk/media/about-the-ico/documents/2615039/project-explain-20190603.pdf

45 Office for AI (2019), Draft Guidelines for AI Procurement. Available at: https://www.gov.uk/government/publications/draft-guidelines-for-ai-procurement/draft-guidelines-for-ai-procurement

46 Same source

The Office for AI's guidelines also helpfully outline how public bodies can reform their own internal practices to ensure that the people commissioning the technology understand the importance of considering public standards, such as transparency and accountability, throughout the procurement process. Their guidance on building multidisciplinary teams is particularly useful. The Committee frequently heard how important it is to combine the expertise of data scientists, data ethics specialists and policy experts when using AI technology, and it is likely that "developing, evaluating and delivering AI invitation-to-tenders will be more effective with diverse teams that understand the interdependent disciplines AI covers."[47] The guidelines' extensive use of the Data Ethics Framework (DEF) is also welcome, as this ensures consistency and continuity across an AI project for public sector organisations using the DEF in AI deployment.

Private providers of public services are subject to the Seven Principles of Public Life, and this area has been the focus of the Committee's attention before. As this past work has shown, when awarding contracts public bodies should consider the ethical behaviour and culture of a company, as well as whether the AI product meets ethical standards. The procurement process should be used to convey to private companies that they have ethical obligations throughout the entire course of a contract and that ethics is not a one-off event, nor one that can be devolved to the public sector purchaser. The Committee's recommendations for reform of the procurement process are discussed in chapter 4.

47 Same source

Chapter 4: Regulating AI

4.1. Introduction

Recent guidance published by the Office for AI (in partnership with the Turing Institute and GDS) and the ICO marks a welcome and significant step forward in AI governance across the public sector. This guidance, covered in chapter 3, provides a good starting point for thinking about how public bodies can establish process-based governance mechanisms that safeguard public standards when they are using AI. The Committee believes, however, that guidance alone does not provide a strong enough incentive to change behaviour.

A strong and coherent regulatory framework for AI in the UK public sector is still a work in progress. A comparison between AI in healthcare and AI in policing is instructive. Healthcare practitioners told the Committee they were confident AI could be implemented ethically because medicine operates within a strictly regulated system, where there is already in place a professional system for testing, integrating and challenging new practices and technologies, and clear standards for reporting, research and clinical trials. Experts working in the field of medical AI told this review that new technologies would slot easily into this pre-existing framework.

In contrast, the same established and well-understood regulatory framework does not currently exist in policing. There is no clear process for evaluating, procuring or deploying new technologies such as predictive policing or facial recognition, which are already being used to support decision-making across the UK. In the absence of a clear regulatory framework for policing, safeguards for public standards are left to individual police forces, whose recent attempts at creating ethical AI systems have led to mixed results.[48] Evidence submitted to this review showed that the use of AI in policing is far more representative of the wider public sector than AI in healthcare. AI may be used in areas such as education, social care and welfare, without a proper understanding of the distinctive value added or risks created by AI systems, their impact on citizens, and the extent to which they serve legitimate policy aims.[49] Hence the need for a strong regulatory framework.

Efforts to establish clear regulation for AI are underway. The General Data Protection Regulation 2018 (GDPR) establishes an extensive legal framework for any organisation processing personal data, including provisions for automated processing. Through its strong ethical foundation and fair processing requirements, the GDPR safeguards against many of the standards issues highlighted in this report. The ICO, as the UK's data protection regulator, is currently looking at how the GDPR applies to AI. Their conclusions will form a substantive part of the UK's regulatory landscape for AI.

Other laws, regulations and public bodies are also relevant. The Equality Act 2010 prohibits discrimination against certain protected characteristics, making it the key law safeguarding against data bias. The Centre for Data Ethics and Innovation (CDEI) was established to advise government on AI regulation. Procurement processes act as a form of soft regulation, setting the terms for commercial relationships. Mandatory impact assessments can change public sector behaviour and obligations under the Freedom of Information Act set the terms for transparent disclosure.

48 In October 2019, the ICO issued a formal Opinion on Live Facial Recognition. The Information Commissioner found the current laws, codes and practices relating to LFR will not drive the ethical and legal approach needed to manage the risk.

ICO (2019), Information Commissioner's Opinion: the use of live facial recognition technology by law enforcement in public places. Available at: https://ico.org.uk/media/about-the-ico/documents/2616184/live-frt-law-enforcement-opinion-20191031.pdf?hootPostID=d672132320a2e1a6fea681db20056c9

49 Alexander Babuta, Marion Oswald and Christine Rinik (2018), 'Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges', RUSI Whitehall Report, 3-18. Available at: https://rusi.org/sites/default/files/201809_whr_3-18_machine_learning_algorithms.pdf.pdf

The Committee has concluded, however, that even taken together this regulatory framework is not yet fit for purpose. Though improvements have been made in recent months and years, current regulation – as it is understood and implemented across the public sector – does not provide a strong enough defence against the risks to public standards identified in this report. This chapter hopes to provide some direction on how government should regulate AI to uphold public standards, covering in turn the GDPR; the Equality Act; the CDEI; Procurement and the Digital Marketplace; Impact Assessment; and Transparent Disclosure. This chapter is not a comprehensive examination of AI regulation. It is limited to the areas that most directly affect the three public standards at the core of this review: openness, accountability and objectivity.

4.2. Legal compliance

Any effective system of public sector regulation requires public bodies to take proactive measures to comply with existing legislation and ensure there is a clear basis in law for any activity they undertake. However, there was a widespread perception among contributors to this review that public bodies are introducing AI into service delivery without a clear understanding of the requirements of the law. Concerns were most pressing in law enforcement and the judiciary, where new surveillance capabilities, such as automated facial recognition (AFR), will impact on citizens' rights and freedoms.

Legal experts told the Committee that public bodies were often relying on a tenuous and piecemeal legal basis, often constituted from multiple sources, to legitimate the use of new technology. Contributors criticised the fact that intrusive and controversial technology, which has the potential to reshape society in radical ways, is introduced in this way.

> "[It is] not adequate to employ technical legal arguments to 'cobble together' an 'implicit' lawful basis, given that power, scale and intrusiveness of these technologies create serious threats to the rights and freedoms of individuals, and to the collective foundations or our democratic freedoms."[50]
> **Professor Karen Yeung, Interdisciplinary Professorial Fellow in Law, Ethics and Informatics, University of Birmingham Law School and School of Computer Science**

The validity of these legal bases is already being tested in the courts. In September 2019, the High Court found that the use of facial recognition by South Wales Police was lawful.[51] Some contributors welcomed the use of judicial review to establish legal clarity, viewing it as an important mechanism to establish checks and balances on executive power. Others, however, argued that it is not appropriate for the legislative framework for era-defining technology to be created by judicial review, especially when much of the legislation subject to review was not designed with AI in mind.

Public bodies should not implement AI without understanding the legal framework governing its use. Introducing algorithmic systems into the public sector without a clear legal basis not only undermines public standards, but also the rule of law. Judicial review may create legal clarity but a series of high-profile court cases investigating illegality by public bodies will undermine trust in what can be a potentially beneficial technology.

50 Written evidence 20 (Professor Karen Yeung)

51 Bridges v Chief Constable of South Wales Police [2019] EWHC 2341. Available at: https://www.judiciary.uk/wp-content/uploads/2019/09/bridges-swp-judgment-Final03-09-19-1.pdf

The Law Society, in its report on the use of algorithms in the criminal justice system, recommended that "[t]he lawful basis of all algorithmic systems in the criminal justice system must be clear and explicitly declared in advance."[52] This should apply not only to the criminal justice system, but to the public sector in general. Public bodies should publish a statement on how their use of AI complies with the relevant laws and regulations before they are deployed in public service delivery.

Recommendation 2:

All public sector organisations should publish a statement on how their use of AI complies with the relevant laws and regulations before they are deployed in public service delivery.

4.3. The GDPR

Given that most uses of AI in the public sector will involve the processing of citizens' personal data, the GDPR – which has direct application in UK law through the Data Protection Act 2018 – creates an extensive legal framework for AI. It places a number of obligations on organisations handling personal data and has a strong ethical foundation. The GDPR gives people enhanced protections against unnecessary data collection, and seeks to limit the intrusive use of data through its principles of fairness and privacy by design, which in turn protect a range of further rights.

Article 5 of the GDPR sets out key principles for the processing of personal data. These include lawfulness, fairness and transparency; purpose limitation; accuracy; and accountability. Many of these normal obligations are risk-based and especially pertinent to AI. Insofar as automated decision-making involves the processing of personal data, all of these provisions apply. For example, organisations must identify a lawful basis for collecting and processing personal data. An organisation that does not establish a clear legal basis for the use of AI would not only undermine public standards but would likely be in breach of data protection legislation.

Full knowledge and articulation of purposes for processing are also required by the purpose specification and use limitation principles. These say that personal data should only be collected for specified purposes and then only used for those purposes or purposes that are compatible with the original one. This could provide an effective safeguard for ensuring that AI is only used for the purpose it is meant to serve. However, it is likely that a narrow interpretation of this principle may not prove useful, particularly because AI may yield unforeseen and sometimes unpredictable results.

Data protection law is technology neutral. It does not directly refer to AI or any associated technologies such as machine learning. However, the GDPR does have a significant focus on large scale automated processing of personal data, and several provisions make specific reference to the use of profiling and automated decision-making. This means that it applies to the use of AI to provide a prediction or recommendation about someone. For example, the law requires organisations to handle personal data in ways that people would reasonably expect and not in ways that are unduly detrimental and might cause harm. This would likely require organisations to use AI in ways that are proportionate and not discriminatory.

52 The Law Society (2019), Algorithms in the Criminal Justice System, 61 (Recommendation 5). Available at: https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/

The right to be informed	Articles 13 and 14 of the GDPR give individuals the right to be informed of the existence of solely automated decision-making, meaningful information about the logic involved, and the significance and envisaged consequences for the individual.
The right of access	Article 15 of the GDPR gives individuals the right of access to information on the existence of solely automated decision-making, meaningful information about the logic involved, and the significance and envisaged consequences for the individual. Recital 71 provides interpretive guidance. It says that individuals should have the right to obtain an explanation of a solely automated decision after it has been made, but it is not legally binding.
The right to object	Article 21 of the GDPR gives individuals the right to object to processing of their personal data, specifically including profiling, in certain circumstances.
Rights related to automated decision-making including profiling	Article 22 of the GDPR gives individuals the right not to be subject to a solely automated decision producing legal or similarly significant effects. There are some exceptions to this and in those cases it obliges organisations to adopt suitable measures to safeguard individuals, including the right to obtain human intervention, to express their view, and to contest the decision. Recital 71 also provides guidance for Article 22.[53]

The GDPR is enforced by a specialist data protection regulator, the Information Commissioner's Office (ICO). The ICO also has a number of enforcement powers, which help to safeguard against potential breaches of data protection legislation. The ICO can impose penalty notices, fine organisations up to £20 million or 4% of their annual turnover (whichever is higher) for breaking the law, and issue guidance that must be considered by Courts arbitrating on the GDPR.

The ICO also issues formal opinions. These opinions, though non-binding, carry significant weight and authority and should encourage organisations to comply with the views of the Commissioner. For example, in October 2019 the Information Commissioner issued an opinion on the use of live facial recognition technology (LFR) by law enforcement in public places,[54] following the High Court judgement on South Wales Police Force. It found that sensitive processing happens at each stage of the LFR process and as such it is subject to data protection law, including the EU Law Enforcement Directive.

Issues of openness, responsibility, explanations and accountability examined in this review are all covered by the GDPR. Overall, the GDPR regulates these issues well. Though some legal experts voiced doubts that the law covered the issues of responsibility and explainability sufficiently, the Committee is satisfied that ICO guidance resolves the issues identified.

4.3.1. The GDPR and openness

Articles 13, 14 and 15 of the GDPR cover several elements of openness. Articles 13 and 14 are transparency obligations. These Articles tell organisations what information they must disclose to individuals before processing their data.

53 ICO and the Alan Turing Institute (2019), Explaining decisions made with AI, Draft guidance for consultation. Available at https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf

54 ICO (2019), Information Commissioner's Opinion: the use of live facial recognition technology by law enforcement in public places. Available at: https://ico.org.uk/media/about-the-ico/documents/2616184/live-frt-law-enforcement-opinion-20191031.pdf?hootPostID=d67213232a0a2e1a6fea681db20056c9

This includes the purpose of and lawful basis for processing, and also details of the existence of automated decision-making, including profiling.

These provisions effectively say that if you use AI to make solely automated decisions about people with legal or similarly significant effects, you must tell them what information you use, why it is relevant and what the likely impact is going to be.[55] Article 15 has similar effects but has to be triggered by the data subject. It says that individuals have a right to access information about the processing of their personal data after such processing has taken place, including where solely automated decision-making systems were used.[56] These rights, taken together, provide a clear regulatory obligation for public sector organisations to be transparent about their use of AI.

4.3.2. The GDPR and responsibility

At first glance, Article 22 imposes a general restriction on "solely automated decision-making" and profiling where it results in a decision with "legal or similarly significant effects". In theory this would prevent public bodies from implementing AI where no human has intervened in the decision-making process, creating a strong legal safeguard against the removal of human responsibility in a public sector decision-making process.

Some legal experts told the Committee, however, that Article 22 is less of a safeguard than it appears to be. This is because the word "solely" effectively undermines the provision, as it would permit any automated system subject to a cursory glance by a human operator, even if the human operator did not or could not make any changes or contribute to the operation of the system. The law could allow public officials to circumvent these provisions by rubber-stamping AI decisions with little or no human intervention in the decision-making process.

Data protection experts told the Committee that the applicability of the provision could be improved by using the phrase "solely or predominantly based on", or by using a more detailed definition of automated decision-making, where the nature and type of human involvement is specified.[57]

However, ICO guidance makes clear that a public official automatically approving an AI decision does not constitute sufficient human involvement in the decision-making process. This interpretation is supported by the EU's Article 29 Data Protection Working Party guidelines on automated decision-making, which say that the data controller cannot avoid Article 22 provisions by fabricating human involvement. For example, if someone inputs data to be processed, but has no influence on the decision, it may still be considered solely automated.[58] Given that ICO guidance must be considered by the courts in any AI cases, the Committee is of the view that the law as it currently stands provides an adequate safeguard against fully automated decision-making in the public sector.

55 ICO [online], Guide to the General Data Protection Regulation: The right to be informed. Available at: https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/the-right-to-be-informed-1-0.pdf

56 ICO [online], Guide to the General Data Protection Regulation: The right of access. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-of-access/

57 Sandra Wachter, Brent Mittelstadt, Luciano Floridi (2017), 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection', International Data Privacy Law, 7;2. Available at: https://academic.oup.com/idpl/article/7/2/76/3860948

58 Article 29 Data Protection Working Party (2018), Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, IV, 21. Available at: https://iapp.org/media/pdf/resource_center/W29-auto-decision_profiling_02-2018.pdf

"[H]uman involvement has to be active and not just a token gesture. The question is whether a human reviews the decision before it is applied and has discretion to alter it or whether they are simply applying the decision taken by the automated system."[59]

What does the GDPR say about automated decision-making and profiling? ICO

4.3.3. The GDPR and explanations

There is legal uncertainty around the right to explanation, which is said to exist under the GDPR. If this provision were to exist, it would grant citizens a legally mandated and meaningful right to explanation for decisions made by automated systems. This would be a promising legal mechanism in the broader pursuit by government of accountability and transparency in AI-enabled public service delivery. However, some legal experts told the Committee that such a right is unlikely to exist because there is nothing in the legally binding provisions of the GDPR that mandates a right to an explanation, and the idea of a right to an explanation only exists in non-binding recitals to the law. This means the law, in this regard, runs the risk of being toothless.

To provide guidance on explanations and clarify the law, the ICO and the Turing Institute are undertaking Project Expl*AI*n. In their guidance, they take an alternative view, stating that "the reference to an explanation of an automated decision after it has been made in Recital 71 makes clear that such a right is implicit in Articles 15 and 22." Contributors to this review also emphasised that administrative law and the right to an appeal in UK law creates a strong legal incentive to provide an explanation for any public sector decision.

"You need to be able to give an individual an explanation of a fully automated decision to enable their rights, to obtain meaningful information, express their point of view and contest the decision."[60]

ICO Guidance, Why Explain AI, Project Expl*AI*n

The Committee is satisfied that ICO guidance provides a sufficient regulatory safeguard for the provision of explanations in public sector decision-making. ICO guidance should be considered authoritative and public bodies should provide explanations accordingly.

4.3.4. The GDPR and accountability

The GDPR includes its own explicit accountability principle, which says that organisations are responsible for the way that they use personal data and must have in place appropriate mechanisms for demonstrating compliance with GDPR principles. Article 24 of the GDPR says that organisations need to implement technical and organisational measures that are risk-based and proportionate to meet the requirements of accountability. Organisations are advised and in some cases required to: implement data protection policies; take a "data protection by design" approach; document processing activities; and carry out data protection impact assessments (DPIAs). AI that processes personal data will have to comply with these requirements, and the GDPR therefore provides a strong regulatory impetus for organisational accountability.

59 ICO [online], Guide to the General Data Protection Regulation: What does the GDPR say about automated decision-making and profiling? Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-does-the-gdpr-say-about-automated-decision-making-and-profiling/

60 ICO and the Alan Turing Institute (2019), Explaining decisions made with AI, Draft guidance for consultation. Available at https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf

Data impact assessments are used to analyse, identify and minimise data protection risks. Irrespective of whether there is a new formal mechanism for AI risk assessment (see section 4.7), a DPIA is almost always going to be mandatory where public bodies are using AI to make decisions. This is because Article 35(1) says that organisations must carry out a DPIA where the type of processing is likely to result in a high risk to the rights and freedoms of individuals. This includes profiling, the large scale use of sensitive data and public monitoring, and is likely to include most, if not all, processing of personal data by innovative technology.[61]

4.4. The Equality Act

Data bias could cause AI to produce decisions and policy outcomes that are discriminatory. Civil rights groups have criticised predictive policing models in particular, fearing that the use of AI could introduce discriminatory practice. Decisions may be made by algorithm without due consideration to policies and practices intended to safeguard those with protected characteristics, enhance diversity and improve outcomes for marginalised people. From a standards perspective, there is no reason to view discrimination resulting from biased data differently from discrimination resulting from human bias. Both undermine the Nolan Principle of objectivity.

> "Although predictive policing is simply reproducing and magnifying the same patterns of discrimination that policing has historically reflected, filtering this decision-making process through complex software that few people understand lends unwarranted legitimacy to biased policing strategies that disproportionately focus on BAME and lower income communities."[62]
> **Policing by Machine, Liberty**

Biased decision-making may also violate non-discrimination law. The Equality and Human Rights Commission (EHRC) has statutory powers to enforce the Equality Act 2010, which prohibits discrimination against nine protected characteristics.[63] The Act also established the Public Sector Equality Duty (PSED) in 2011, which mandates public bodies to take a proactive approach to fighting inequality. There is nothing in the Equality Act that specifically refers to AI or automated decision-making. However, evidence from anti-discrimination lawyers outlined a number of ways in which the law's provisions against direct and indirect discrimination could apply to AI.

61 ICO [online], Guide to the General Data Protection Regulation: When do we need to do a DPIA? Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/when-do-we-need-to-do-a-dpia/

62 Liberty (2019), Policing by Machine: predictive policing and the threat to our rights. Available at: https://www.libertyhumanrights.org.uk/sites/default/files/LIB%2011%20Predictive%20Policing%20Report%20WEB.pdf

63 Equality Act 2010, section 4. Available at: http://www.legislation.gov.uk/ukpga/2010/15/section/4
Protected characteristics defined by the Act are: age; disability; gender reassignment; marriage and civil partnership; pregnancy and maternity; race; religion or belief; sex; and sexual orientation

"In 2017, Durham Constabulary started to implement a Harm Assessment Risk Tool (HART), which utilised a complex machine learning algorithm to classify individuals according to their risk of committing violent or non-violent crimes in the future. This classification is created by examining an individual's age, gender and postcode. This information is then used by the custody officer, so a human decision maker, to determine whether further action should be taken. In particular, whether an individual should access the Constabulary's Checkpoint programme which is an "out of court" disposal programme.

There is potential for numerous claims here. A direct age discrimination could be brought by individuals within certain age groups who were scored negatively. Similarly, direct sex discrimination claims could be brought by men, in so far as their gender leads to a lower score than comparable women. Finally, indirect race discrimination or direct race discrimination claims could be pursued on the basis that an individual's postcode can be a proxy for certain racial groups. Only an indirect race discrimination claim would be susceptible to a justification defence in these circumstances."[64]
AI Law Hub

Contributors to the review mentioned that the Public Sector Equality Duty (PSED), if used properly, was the single best tool available to deal with data bias. This is because it requires organisations to consider how they could positively contribute to the advancement of equality, and requires "equality considerations to be reflected in the design of policies and delivery of services".[65] Equality Impact Assessments are not required by law, but are often used by public bodies to facilitate compliance with the PSED. They are used to identify possible negative impacts of decisions on individuals and groups with protected characteristics and plan mitigating action accordingly. They are also used to identify opportunities to advance equality within the policies, strategies and services of a public authority.

"Public bodies must consider the Public Sector Equality Duty when they make decisions about how they fulfil their public functions and deliver their services. When moving towards automated decision making the PSED provides an opportunity for equality considerations to be built into decision-making processes as they are developed."
Rebecca Hilsenrath, Chief Executive, Equality and Human Rights Commission

However, there is uncertainty around how the legislation applies in practice to automated decision-making in the public sector. There is currently no bespoke regulatory guidance outlining what public bodies introducing AI systems need to do to comply with the Equality Act 2010. Public bodies introducing AI systems need to know how the Act applies to discriminatory outcomes enabled by automated decision-making. They need specific guidance on how to comply with the legislation, as well as guidance on how to measure bias and mitigate its effects, particularly given the widespread belief among AI experts that data bias cannot or should not be completely eradicated.

64 Robin Allen QC and Dee Masters [online], 'UK's existing equality and human rights framework', AI Law Hub. Available at: https://ai-lawhub.com/framing-the-debate/#criminal
On Durham Constabulary's HART model see: Alexander Babuta (2018), 'Innocent Until Predicted Guilty? Artificial Intelligence and Police Decision-Making', RUSI Newsbrief, 38;2. Available at: https://rusi.org/sites/default/files/20180329_rusi_newsbrief_vol.38_no.2_babuta_web.pdf

65 Equality and Human Rights Commission [online], Public Sector Equality Duty Guidance. Available at: https://www.equalityhumanrights.com/en/advice-and-guidance/public-sector-equality-duty

Through Project ExplAIn, the ICO and the Turing Institute are developing extensive guidance on explanations and the GDPR. The Committee believes that a similar project is necessary on data bias and the Equality Act. The Equality and Human Rights Commission should develop guidance on data bias in partnership with the Turing Institute and the CDEI.

Recommendation 3:

The Equality and Human Rights Commission should develop guidance in partnership with both the Alan Turing Institute and the CDEI on how public bodies should best comply with the Equality Act 2010.

Though this project should focus on the Equality Act as it currently stands, some contributors suggested that a fundamental rethink of anti-discrimination law may be needed in the long term. The use of machine learning raises new issues that current anti-discrimination law may not cover. How will we detect cases of discrimination when a citizen may not even know if a decision has been made on the basis of a protected characteristic? Can discrimination law have any effect if discrimination occurs via proxy characteristics but we cannot identify what those proxies are? What forms of algorithmic profiling count as discrimination? Government should remain open to a revision of anti-discrimination law if the current legal framework cannot answer these questions convincingly.

4.5. Regulatory assurance body

Some contributors to this review suggested that a new system of ethical regulation for the use of AI in the public sector was necessary. The Committee heard that a statutory arms-length public body, similar to the Human Fertilisation and Embryology Authority (HFEA), could have a role in licensing technology and leading on standards, review and assessment.[66] Justice of The Supreme Court, The Right Hon Lord Sales also called for an independent regulator of algorithms that would be staffed by technical experts, lawyers and ethicists. He argued that issues around AI are so large and impenetrable that an expert commission on algorithms is necessary to safeguard against the legal and ethical challenges posed by AI. Lord Sales said that this is particularly pertinent because government currently lacks the technical capacity to do this well itself.[67]

The Committee agrees with the rationale for extra regulatory scrutiny and independent advice on the issues associated with AI. However, most contributors to this review argued that a single AI regulator was impractical. The Committee heard that any system of ethical regulation for AI in the public sector would require sectoral-based review to account for the context specific risks and opportunities of automated decision-making across policy areas. A new AI regulator would inevitably overlap with existing regulatory bodies, who will already have to regulate AI within their sectors and remits. As such, the Committee believes that the UK does not need a new regulator.

66 Written evidence 12 (Dr Emma Carmel)

67 Lord Sales, Justice of the UK Supreme Court (2019), 'Algorithms, Artificial Intelligence and the Law', The Sir Henry Brooke Lecture for BAILII. Available at: https://www.supremecourt.uk/docs/speech-191112.pdf

> "People often say 'Let's have a new regulator. Let's have a new, shiny one.' Actually, there is a lot of expertise already in the regulators because they are having to deal with this kind of thing in markets which they are there to regulate. We ought to build on that and use the expertise we have got."
> **Professor Helen Margetts, Professor of Society and the Internet, University of Oxford and Director of the Public Policy Programme, The Alan Turing Institute**

Instead, the Committee is of the view that existing regulators should be aware of how automated technology will impact their sectors, and adapt their practices accordingly. However, given the complexity of this technology and that expertise is not necessarily well established in this area, it is unlikely that regulators will be able to meet the challenges posed by AI without guidance from a central body. AI will create unforeseen issues for regulation, where technical knowledge and expertise will be necessary. There is clear space in the regulatory landscape for a "regulatory assurance" body, which provides advice to individual regulators and government on the issues associated with AI, and identifies any regulatory gaps. This body would not act as a regulator, but it would need full independence from government to advise objectively and without political interference.

The Centre for Data Ethics and Innovation (CDEI) has many of the necessary skills to fulfil this role and the Committee supports the government's published intention for CDEI to oversee the regulatory landscape, analysing and anticipating gaps in governance and regulation that could impede the ethical deployment of AI, and to advise government accordingly.[68] The Committee also believes that the CDEI has a role to play in advising individual regulators, as well as government, on the issues associated with this technology. The Committee supports the government's intention to place the centre on a statutory footing to safeguard its independence.[69] However, the specific roles and functions of the CDEI remain unclear. The government must clarify its purpose and assure that appropriate safeguards are in place so that it can fulfil its intended role as a regulatory assurance body.

68 Centre for Data Ethics and Innovation, Government Response to Consultation, November 2018. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/757509/Centre_for_Data_Ethics_and_Innovation_-_Government_Response_to_Consultation.pdf

69 Same source

Recommendation 4:

Given the speed of development and implementation of AI, we recommend that there is a regulatory assurance body, which identifies gaps in the regulatory landscape and provides advice to individual regulators and government on the issues associated with AI.

We do not recommend the creation of a specific AI regulator, and recommend that all existing regulators should consider and respond to the regulatory requirements and impact of the growing use of AI in the fields for which they have responsibility.

The Committee endorses the government's intention for CDEI to perform a regulatory assurance role. The government should act swiftly to clarify the overall purpose of CDEI before setting it on an independent statutory footing.

4.6. Procurement and the Digital Marketplace

Contributors to this review emphasised the importance of 'ethics by design'. Some ethical requirements will require technical solutions, which will need to be specified in the commissioning and design of any project. For example, to build an AI system that is accountable, public bodies may need to 'build in' the capacity for it to produce explanations for its decisions. This makes procurement a crucial point in the AI lifecycle where provisions for ethical standards must be set. It is important from the start of any project that the business, technology and procurement are aligned around what the preferred outcomes will be.

Evidence gathered for this review indicates that most public bodies will use external suppliers to build and manage their AI systems. This raises additional issues over and above those where AI is built and managed in-house. In its 2014 report, 'Ethical Standards for Providers of Public Services' and later in its 2018 report 'The Continuing Importance of Ethical Standards for Public Service Providers', the Committee called for public service providers to recognise that the Nolan Principles apply to them. Private providers of public services cannot delegate responsibility for standards to public bodies, and they should have in place provisions for ensuring high ethical standards in public service delivery, irrespective of whether they are using AI. Government also has a responsibility to manage third-party contracts in a way that engenders high ethical standards. Conversely public authorities cannot outsource their risk to suppliers.

"The Cabinet Office should reinforce the message that the Seven Principles of Public Life apply to any organisation delivering public services.

The Cabinet Office should ensure that ethical standards reflecting the Seven Principles of Public Life are addressed in contractual arrangements, with providers required to undertake that they have the structures and arrangements in place to support this.

Commissioners of services should include a Statement of Intent as part of the commissioning process or alongside contracts where they are extended, setting out the ethical behaviours expected by government of the service providers."[70]

Recommendations from the Committee's 2014 and 2018 reports into providers of public services

In its 2018 report, the Committee found that the government had made some improvements in how it manages the ethical conduct of contractors as part of a broader maturing of outsourcing practices. However, the Committee has not had a formal response from government to the recommendations made in that report, and there appears to be limited progress on introducing formal measures to reinforce the application of ethical standards in the procurement process.

This lack of focus on ethical standards was reflected in evidence collected for this review. Ethical considerations do not appear to play much part in AI procurement across the public sector at present. Public policy officials and private service providers both told the Committee that provisions for ethics are not typically part of tenders or contracts, and that ethics are often considered, if at all, mid-way through the development of an AI system.

"Ethical standards are definitely not part of the procurement process at this point in time."
Ian O'Gara, Accenture

Ethical considerations need to be injected early into the procurement cycle to give them the best chance of surviving the life of the contract. Ethics should be considered at each stage of the procurement process: from strategic planning, through scrutinising tenders and verifying contracts, to monitoring and evaluating the performance of a public service provider.

The procurement process should be designed so that AI products and services that facilitate high standards are preferred, and that it rewards companies that have prioritised ethical practices. As part of the commissioning process, or when contracts are extended, the government should set out the ethical principles that companies providing services to them are expected to exemplify. Adherence to ethical standards should be part of the evaluation process and should be given an appropriate weighting. Companies that show a commitment to these ethical behaviours should be scored more highly than those that do not. This would help ensure that the suppliers who think about ethics, and who build ethics into their systems, have a competitive advantage. In doing so, government will leverage its significant purchasing power to incentivise private providers to build ethical AI.

Several contributors to this review also suggested that public bodies should ensure that provisions for ethical standards are written into service delivery contracts. This was seen as particularly important given the potential for private companies to cite commercial confidentiality or trade secrets as reasons to withhold information about how their algorithms work. This would undermine

70 The Committee on Standards in Public Life (2018), The Continuing Importance of Ethical Standards for Public Service Providers. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/705884/20180510_PSP2_Final_PDF.pdf

accountability by making access to explanations and the auditing of AI systems more difficult.

> "Assertions of commercial confidentiality should not be accepted as an insurmountable barrier to appropriate rights of access to the [algorithmic] tool and its workings for the public sector body, particularly where the tool's implementation will impact fundamental rights. Government procurement contracts relating to AI and machine learning should not only include source code escrow provisions, but rights for the public sector party…as standard."[71]
>
> **Marion Oswald, Senior Fellow in Law and Director of the Centre for Information Rights, University of Winchester**

Recommendation 5:

Government should use its purchasing power in the market to set procurement requirements that ensure that private companies developing AI solutions for the public sector appropriately address public standards.

This should be achieved by ensuring provisions for ethical standards are considered early in the procurement process and explicitly written into tenders and contractual arrangements.

Centralised procurement tools should also be improved. The Cabinet Office informed the Committee that the Crown Commercial Service's Digital Marketplace is responsible for around 25% of public sector technology procurement of common goods and services. As it stands, the marketplace contains no provisions to support ethical standards. In order to advertise their products or services on the marketplace, private companies need only to fill out a tick-box questionnaire, with no reference to managing standards.

This represents a missed opportunity. The marketplace could offer a range of tools to help providers assess if AI products will support or undermine public standards. Canada, for example, operates a register of responsible AI companies.[72] The marketplace could also allow AI products and services to be classified according to certain features, such as explainability. Such tools would help public bodies navigate the range of products and services offered. In discussions with the Committee, Crown Commercial Service (CCS) officials expressed a desire for the marketplace to play a more active role in the procurement process.

A new specialist AI framework, including separate streams for machine learning and robotic process automation, is currently under development. Before the launch of its new AI framework, CCS should consider what tools it can introduce to the marketplace to best help public bodies find AI products and services that meet their ethical requirements. The shaping of this will be determined by engaging with the market, both suppliers and departments, to get their views prior to designing the new commercial vehicle.

71 Written evidence 4 (Marion Oswald)

72 Available at: https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/list-interested-artificial-intelligence-ai-suppliers.html

Recommendation 6:

The Crown Commercial Service should introduce practical tools as part of its new AI framework that help public bodies, and those delivering services to the public, find AI products and services that meet their ethical requirements.

4.7. Impact assessment

Contributors cited the absence of a compulsory standards risk management tool as a major gap in the UK's national AI governance framework. Multiple public policy experts told this review that mandatory impact assessments should fill this gap.

This was for four reasons. First, mismanagement of AI systems could seriously undermine public standards. Impact assessments would inform public bodies what level of risk their AI system could pose, and allow those authorities to set risk-based governance accordingly.

> "Public servants must be incentivised in some way to carry out impact assessments and act upon their results, without being constrained from adopting beneficial innovation."[73]
> **Centre for Data Ethics and Innovation**

Second, impact assessments were deemed necessary because AI is new technology that most public bodies have little experience in. Risks such as data bias would be new and unfamiliar to most, and so impact assessment would push these issues to the fore.

Third, impact assessments were seen as important for accountability. Proper accountability depends on public bodies being aware of the risks of their AI systems, so that authorities can be assessed against any mitigation measures they take.

Fourth, impact assessments are necessary because AI tools are likely to have a major impact on citizens and we need to be certain that their interests and rights are protected. Impact assessment is one major element in meeting the responsibility of due diligence.

Though some standards issues are covered by EIAs (Equality Impact Assessments) and DPIAs (Data Protection Impact Assessments), neither was seen as comprehensively covering all relevant standards issues. In particular, contributors cautioned against using DPIAs as a proxy for ethical risk assessment.

In contrast, multiple contributors spoke favourably about the Canadian model of Algorithmic Impact Assessment. Following a Treasury Directive on Automated Decision Making,[74] the Canadian government introduced a mandatory algorithmic impact assessment for automated decision systems.[75] The assessment consists of an electronic survey that covers the social, environmental, and human rights impact of an AI system, as well as provisions for data quality and human responsibility. It then generates a risk score for the automated system. Though some contributors noted flaws with the specific wording of the Canadian model, it was applauded as an overall framework for upholding ethical standards.

73 Written evidence 18 (Centre for Data Ethics and Innovation)

74 Available at: https://docs.google.com/document/d/1LdciG-UYeokx3U7ZzRng3u4T3lHrBXXk9JddjjueQok/edit

75 Available at: https://open.canada.ca/aia-eia-js/?lang=en

"The AIA provides designers with a measure to evaluate AI solutions from an ethical and human perspective, so that they are built in a responsible and transparent way. For example, the AIA can ensure economic interests are balanced against environmental sustainability.

The AIA also includes ways to measure potential impacts to the public, and outlines appropriate courses of action, like behavioral monitoring and algorithm assessments."[76]

Canadian Government Video on AIA

Will you have documented processes in place to test datasets against biases and other unexpected outcomes? This could include experience in applying frameworks, methods, guidelines or other assessment tools.

Will you be developing a process to document how data quality issues were resolved during the design process?

Will you be making this information publicly available?

Will you undertake a Gender Based Analysis Plus of the data?[77]

Questions on data quality taken from Canada's Algorithmic Impact Assessment

Alternatively, the section in the Turing AI Guide 'Using AI ethically and safely' favoured a Stakeholder Impact Assessment (SIA). An SIA encourages public bodies to identify affected stakeholders, analyse the fairness of desired outcomes, and examine the possible impacts of an AI system on the individual and society.

In contrast with the Canadian box-ticking approach, the SIA offers more open-ended questions and allows public bodies to develop their own sector-specific and use case-specific questions.

Both the AIA and the SIA would help public bodies navigate the full range of ethical risks that AI poses. The SIA in particular is designed to point towards effective mitigation measures. The stipulation that an SIA is carried out at three stages of an AI project lifecycle – problem formulation, pre-implementation and reassessment after deployment – ensures that its outcomes will inform project design and lead to remedial action.

Goal-Setting and Objective-Mapping

How are you defining the outcome (the target variable) that the system is optimising for? Is this a fair, reasonable, and widely acceptable definition? Does the target variable (or its measurable proxy) reflect a reasonable and justifiable translation of the project's objective into the statistical frame?

Is this translation justifiable given the general purpose of the project and the potential impacts that the outcomes of its implementation will have on the communities involved?[78]

Questions taken from the UK government guidance's Stakeholder Impact Assessment

76 Available at: https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html

77 Available at: https://open.canada.ca/aia-eia-js/?lang=en

78 Dr David Leslie (2019), Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. Available at: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

The Committee believes that for impact assessments to be effective they must meet three conditions. First, an impact assessment should be mandatory for any machine learning system before it is deployed. The Committee heard that too often in the public sector impact assessments are undertaken as a rubber-stamping activity after a project has already been approved. Given the absence of regulatory clarity, an optional impact assessment would mean that restraints on machine learning systems in the UK public sector remained weak. In line with the SIA model, the impact assessment should also be repeated at later stages as an AI system develops.

Second, as an AI impact assessment is a tool of ethical review, it should not be set by an organisation which already has a vested interest in implementing an AI system. In this regard, we do not favour the SIA's provision to allow public bodies to write their own use-case specific questions. In the words of one expert consulted, a decentralised approach would allow "departments to mark their own homework".

Third, impact assessments should be published. The Office for AI guidance advises public bodies to publish the details of their SIA pre-deployment, which would include details of any mitigation measures taken. This is vital for scrutiny and accountability, as it would allow members of the public to assess how far a public authority has managed the risk an AI system poses adequately. The government should make publication of an AI impact assessment mandatory.

Contributors noted that public sector organisations using AI would be likely to trigger the legislative requirement to undertake a data protection impact assessment, and that the creation of an extra form of impact assessment would create an additional administrative burden for public officials. Government should consider how an AI impact assessment could be integrated into the DPIA process in order to streamline this process for public sector organisations.

Recommendation 7:

Government should consider how an AI impact assessment requirement could be integrated into existing processes to evaluate the potential effects of AI on public standards. Such assessments should be mandatory and should be published.

4.8. Transparent disclosure

The Committee saw two public sector AI projects during the course of this review that demonstrated transparency. The West Midlands Police and Crime Commissioner's Ethics Committee, which advises on data science projects proposed by their Data Analytics Lab, publishes its minutes in full, including where they have criticised police practice.[79] Moorfields Eye Hospital, who have been working in partnership with DeepMind Health (now Google Health) since 2016, also have a useful section on their website dedicated to their machine learning project, including a Q&A and latest updates.[80]

79 Available at: https://www.westmidlands-pcc.gov.uk/ethics-committee/

80 Available at: https://www.moorfields.nhs.uk/landing-page/deepmind-health-research-partnership

Contributors working on AI projects across the public sector told the Committee that negative and sensationalist media coverage often made public bodies wary of being transparent. Policy experts also told the Committee that the biggest incentive towards transparency was often the personal ethical commitment of those working with AI in the public sector. This can lead to quite a fragmented approach to public standards. The Committee heard that central coordination around transparency is required because it is not currently mandated by any regulation or institution.

In its report on Algorithms in the Criminal Justice System, the Law Society recommended the creation of a register of algorithmic systems in criminal justice in the UK. The Committee is of the view that such a register could be expanded beyond criminal justice, if a sensible threshold is set. In discussions with the Committee, the Centre for Data Ethics and Innovation expressed an interest in overseeing such a register.

> "We note the recommendation by the Law Society that a national register of automated decision making tools in use in criminal justice be established. Subject to appropriate exceptions, thresholds and safeguards, this would appear to support the Nolan Principles and would facilitate impact assessment of public sector ADMTs. Such a register may be appropriate in other parts of the public sector."[81]
> **Centre for Data Ethics and Innovation**

However, it is likely that the establishment of a central register, even if restricted to AI systems above a high threshold, would be an extensive and potentially overwhelming bureaucratic challenge, particularly given the predicted scale of AI across public life. There is no guarantee that such a register would be properly accessible to the public. Similar registers, such as those currently used to collect procurement data, were criticised by contributors to this review for being poorly formatted, incomplete and difficult to search.

Having a centralised register of AI systems would also be counter-intuitive to the general public, who would likely go to the website of a public body to find information about how they operate, rather than central government. The Committee is of the view that any system intended to increase transparency should not focus on the creation of a centralised database.

There are already requirements for proactive disclosure under the Freedom of Information Act 2000 (FOI Act). There is a statutory obligation on public bodies under Section 19 of the Act to proactively publish information that the public are likely to be interested in. It is the duty of every public authority to adopt and maintain a publication scheme, approved by the Information Commissioner, which makes information about their business activities available to the public.[82] As the regulator, the ICO provides public bodies with an approved model publication scheme that specifies categories of information that should be published. This includes information about income and expenditure; tendering, procurement and contracts; and decision-making processes.[83] Public bodies that use AI will still need to proactively disclose this information to the public. In theory, this should encourage openness and transparency.

81 Written evidence 18 (Centre for Data Ethics and Innovation)

82 Freedom of Information Act 2000, section 19. Available at: http://www.legislation.gov.uk/ukpga/2000/36/section/19

83 ICO [online], Freedom of Information Act: Model publication scheme. Available at: https://ico.org.uk/media/for-organisations/documents/1153/model-publication-scheme.pdf

The Committee heard, however, that the proactive disclosure requirements of the FOI Act have limited use in the current framework, not least because the legislation is outdated. The ICO told us that publication schemes are not necessarily useful for enforcing transparency, particularly because it is difficult to assess the nature and extent of compliance across the public sector. They said that more could be done to encourage proactive disclosure in other ways, by promoting openness and transparency by design, for example. This would require public bodies to think about what information they should proactively disclose, as well as the implications of not being transparent, from the start of the AI commissioning process.

It is unlikely, however, that an expectation on public bodies to think about openness is enough to change behaviour. Public bodies will need guidance to help them think through openness and transparency implications. The Committee recommends that government should set clear guidelines for public bodies on what information they should proactively disclose about their AI systems.

These guidelines should make explicit the features of an AI system that warrant transparency, such as the processing of personal data for predictive analytics, or the potential impact of a system on an individual. They should also specify how public organisations should make information available to the public.

Recommendation 8:

Government should establish guidelines for public bodies about the declaration and disclosure of their AI systems.

Members of the public also have a general right of access to information under Section 1 of the FOI Act. This says that any person making a request for information to a public authority is entitled to be informed by the public authority whether it holds information of the description specified in the request, and if that is the case, to have that information communicated to them.[84] This does not extend to private providers of public services. The Committee has previously recommended that the government should hold a consultation on extending the application of the FOI Act to private providers where information relates to the performance of a contract with the government in the delivery of public services.[85] The increasing use of private sector companies to deliver AI-enabled public services adds urgency to the Committee's 2018 recommendation.

84 Freedom of Information Act 2000, section 1. Available at: http://www.legislation.gov.uk/ukpga/2000/36/section/1

85 The Committee on Standards in Public life (2018), The Continuing Importance of Ethical Standards for Public Service Providers. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/705884/20180510_PSP2_Final_PDF.pdf

Chapter 5: The role of public bodies

5.1. Introduction

Decisions on adopting and implementing artificial intelligence in the public sector lie with individual government departments and public bodies. Individual police forces or NHS trusts, for example, commission and operate AI systems in their organisation. Each body will need to establish suitable governance mechanisms to manage the ethical risks associated with AI and address regulatory compliance.

In January 2019, Singapore's Personal Data Protection Commission published a proposed model framework for the governance of AI systems in Singapore.[86] The framework is a useful starting point for thinking about the kinds of mechanisms that public sector organisations in the UK should adopt when using AI technology. It states that the risks associated with AI can be managed by adapting existing governance structures to incorporate values, risks and responsibilities relating to algorithmic decision-making. This includes setting clear roles and responsibilities for the ethical deployment of AI and putting in place internal controls to address the risks involved in using AI to make decisions.

The Committee shares the view that effective governance of AI in the public sector does not require a radical overhaul of traditional risk management. Public sector organisations should already have in place governance frameworks that identify, assess and mitigate risk and establish clear responsibilities for decision making. They are also already subject to rigorous scrutiny and checks by external bodies to ensure that they are operating in accordance with their mandates. Therefore, it should not be a huge step for public bodies to put in place effective risk management structures to ensure the robust governance of AI.

This chapter looks first at the risks that need to be managed before deployment, when public bodies are contemplating using AI and are developing AI systems for public service delivery. It then covers five areas of governance key to risk management when deploying AI:

- setting responsibility
- internal and external oversight
- monitoring and evaluation
- appeal and redress
- training and education

Recommendations made here reflect the issues of most concern to the Committee and are intended to supplement public sector guidance discussed in chapter 3.

5.2. Legal and legitimate AI

Before deploying AI, public sector organisations need to demonstrate that the benefits of using the technology outweigh the risks. They will also need to ensure that they are using AI in ways that are legal and legitimate and do not undermine individual rights.

As a first step, policy experts emphasised that public bodies should carefully consider the appropriateness of using artificial intelligence in any given context. On a case by case basis, public sector organisations will need to justify why they are using an algorithm; consider whether the potential impact on individuals is necessary and proportionate; and demonstrate how the tool will improve the current system.[87] Office for AI guidance makes clear that the decision to use AI should always be based on user need.[88] Contributors to this review said that public officials should be prepared to walk away from experimental AI where

86 Personal Data Protection Commission Singapore (2019), A proposed model Artificial Intelligence governance framework (for public consultation). Available at: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf

87 Written evidence 2 (Jamie Grace)

88 Office for AI and GDS (2019), Assessing if artificial intelligence is the right solution. Available at: https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution

there is no clear benefit to the public and where the potential infringement of individual rights cannot be shown to be necessary and proportionate.[89] The Committee agrees with this judgement.

> "You can imagine a scenario where things go wrong because the public sector has implemented some AI technology because it is shiny, cool and exciting rather than helpful."
> **Eddie Copeland, Director, London Office of Technology and Innovation (LOTI)**

All providers of public services must also publish a statement on how their use of AI complies with relevant laws and regulations, as the Committee recommends in chapter 4 (recommendation 2). This is of particular importance given that these technologies can interfere with individual rights and freedoms, and do so at scale, operating in ways that are often difficult for individuals to understand, challenge or contest.

Where AI automates the processing of personal data, public bodies will also need to demonstrate that the data processed by the algorithm is fairly and lawfully obtained, processed and retained, and only used for legitimate purposes, as stipulated under the GDPR. These issues will probably need to be considered through some form of data protection and/or AI impact assessment (as discussed in chapter 4).

5.3. System design

Experts consulted for this review said that good system design could help to mitigate some of the risks to standards identified in this report. For example, it may be possible to build into a system a degree of technical transparency, or provisions for monitoring and evaluating an AI system's performance. To do this, public bodies will need to anticipate how public standards may be affected by any new system before it is introduced and subsequently in deployment.

Public bodies should start by conducting an AI impact assessment, as discussed in chapter 4. This will help public organisations assess how their proposed AI system could affect public standards such as openness, accountability and objectivity. During the course of this review, some public officials expressed concerns that impact assessments could be used retrospectively after the details of a project were already set, as a tick box exercise to show compliance rather than the proper consideration of ethical risk. Such an approach should be avoided as it could embed avoidable risk into the design of an AI system.

Where standards risks are identified it is essential that project development teams alter the design of their systems. For example, if a project risks amplifying bias, public bodies may want to consider broadening their dataset to dilute the effects of that bias. Similarly, if a system is highly automated and risks undermining the principle of accountability, public bodies should consider redesigning the system so that human involvement in the decision-making process is active and meaningful. There is no one-size-fits-all answer to AI system design, as this will be highly context and risk dependent. In the case of automation and accountability, for example, it may be acceptable to automate a system that sends citizens text reminders to pay their council tax, but it would not be appropriate to automate a predictive policing system that grants or denies parole to prisoners.

89 Written evidence 4 (Marion Oswald)

Contributors also emphasised that the type of standards risk identified at the project design stage should inform decisions as to whether to procure an AI system from external providers or to build one in-house. One expert cited West Midlands Police as an example of good practice: by having their own in-house Data Lab, developers understand the ethical constraints of a policing context and apply that understanding when designing their AI systems.

If the risk to public standards remains high despite any mitigation measures taken, then public bodies should not shy away from moderating or constraining the intended use of an AI system. In some cases, it may be that a project should not proceed from design to deployment, even if significant expenses have already been incurred. One expert said that an AI system could come with a "health warning" if there was a high risk the product was biased because it had only been trained on certain populations.

Finally, contributors emphasised the importance of considering standards iteratively as a project progresses from design through to deployment. This is because some systems will undergo a process of continuous testing and redesign to optimise performance, and because AI systems can act in ways that are unpredictable and unexpected. The values underlying what is acceptable and unacceptable in different contexts may also change over time.

Public bodies should not only mitigate standards risks at the project design stage, but continue to monitor risks to standards at all stages of the AI lifecycle and throughout the duration of an AI project. Standards review will need to occur every time a substantial change to the design of an AI system is made.

Recommendation 9:

Providers of public services, both public and private, should assess the potential impact of a proposed AI system on public standards at project design stage, and ensure that the design of the system mitigates any standards risks identified.

Standards review will need to occur every time a substantial change to the design of an AI system is made.

5.4. Diversity

The field of AI is at risk of replicating or perpetuating historical biases and existing structures of inequality in society (see 2.5). In April 2019, a report by the AI Now Institute said that biased AI systems can largely be attributed to the lack of diversity within the AI industry.[90]

Public bodies must maximise diversity at all stages of the AI process to help tackle issues of bias and discrimination within AI systems. There needs to be diversity in the workforce and in training and education, so that biases, whether conscious or unconscious, are less likely to be programmed into AI systems. This includes those building and developing AI systems, and those who have responsibility for AI at various stages of deployment (see 5.5.1). An increased access to a wider range of skills and perspectives at each stage of the process will help public bodies to better consider the impact of AI systems on public standards, and to mitigate the risks identified. Datasets used to train machine learning algorithms will also need to be diverse, so that they work accurately and objectively on different individuals and populations.

90 Sarah Myers West, Meredith Whittaker and Kate Crawford (2019), 'Discriminating Systems: Gender, Race and Power in AI', AI Now Institute. Available at: https://ainowinstitute.org/discriminatingsystems.pdf

Recommendation 10:

Providers of public services, both public and private, must consciously tackle issues of bias and discrimination by ensuring they have taken into account a diverse range of behaviours, backgrounds and points of view. They must take into account the full range of diversity of the population and provide a fair and effective service.

5.5. Deployment of an AI system

Once public sector organisations have assessed their proposed AI systems to show that they are necessary, proportionate and lawful, and have designed their systems in ways that help to mitigate the ethical risks identified, they need to set effective governance mechanisms for its use. Even well designed AI systems will pose risks to openness, accountability and objectivity, and so public bodies will need to put in place sound risk management and other internal controls to address those risks in the day-to-day management of the system.

The Committee has identified five areas of governance necessary for upholding public standards in this context: setting responsibility; monitoring and evaluation; internal and external oversight; appeal and redress; and training and education.

5.5.1. Setting responsibility

In most AI systems, there will not be a single person responsible for the whole system; rather responsibility will be allocated across a range of individuals who engage with the system at various stages of deployment. The key question for public bodies is how responsibility for, and oversight of, AI is allocated across an organisation.

Responsibility, and ultimately human control, will be shared by individuals from across an organisation, including individuals who operate AI systems, project managers who monitor entire AI systems and senior leadership who oversee the policy for which AI is being used. Responsibility could be distributed as in the table below.

Senior leadership	Project managers	Individuals operating AI systems
Make decision to introduce an AI system	Oversee end-to-end AI system process	Check input data
Set governance mechanisms for AI system	Assess the impact of the AI on groups of data subjects	Identify any false positives or system errors
Assess how the AI impacts their policy area as a whole	Monitor and evaluate the system, and make improvements where necessary	Accept or reject decision recommendations

As the Office for AI set out in their guidance, any allocation of responsibility should be clearly documented, so public officials are fully aware of their roles and responsibilities and it is clear to all officials interacting with an AI system where responsibility lies. Such a record will help facilitate accountability for the system.

> "Humans must be ultimately responsible for decisions made by any system...Good governance will require for each use case, a specific understanding of the appropriate division of responsibilities."[91]
> **Centre for Data Ethics and Innovation**

Once responsibility is set, senior leadership must ensure that public officials have the capacity to exercise their responsibility in a meaningful way. They must be properly trained and provided with the resources and guidance needed for them to discharge their duties. Fundamentally, officials must have both the knowledge and the power to implement change, otherwise any designated responsibility is meaningless.

> "The person [needs to have] both the agency and the knowledge necessary to make changes to the system's behaviour and to intervene when it seems like something is going to go wrong."
> **Dr Brent Mittelstadt, Research Fellow and British Academy Postdoctoral Fellow, Oxford Internet Institute**

Public bodies should also seek to ensure that their organisational culture encourages and empowers their officials to use their professional knowledge and expertise in confirming automated decisions.

Similarly, private sector providers should make sure that any system they build meets the requirements for human responsibility set by the public authority.

Recommendation 11:

Providers of public services, both public and private, should ensure that responsibility for AI systems is clearly allocated and documented, and that operators of AI systems are able to exercise their responsibility in a meaningful way.

5.5.2. Monitoring and evaluation

Public bodies deploying AI should establish monitoring systems and processes to evaluate and identify issues relating to the performance of the technology. It is not acceptable for a public organisation introducing AI to assume the technology will always function as intended, particularly because machine learning systems are often vulnerable to flaws like inaccuracy, and can operate in unique and unexpected ways that can have unintended consequences.

Some contributors to this review argued that public bodies may not be aware AI systems could be inaccurate, often citing facial recognition as an example. At Notting Hill Carnival in 2017, facial recognition technology used by the Metropolitan Police was said to be wrong 98% of the time, and more likely to misidentify ethnic minorities and women.[92] The possibility of inaccuracy underscores the importance of monitoring and evaluation, particularly in a context like facial recognition, where the consequences of misidentifying an individual can be significant. Public bodies cannot assume that their AI systems will work as well in real life as they

91 Written evidence 18 (Centre for Data Ethics and Innovation)

92 Vikram Dodd (2018), 'UK police use of facial recognition technology a failure, says report', The Guardian. Available at: https://www.theguardian.com/uk-news/2018/may/15/uk-police-use-of-facial-recognition-technology-failure

do in a data science lab. They will also need to judge what an acceptable error rate is for their specific use of AI, according to the probability and severity of harm to an individual in a particular context.

Contributors also told the Committee that monitoring is vital to prevent unintended consequences. Even where AI is introduced with good intentions, poor quality data or a lack of knowledge about how an AI system operates will lead to unwanted outcomes. Public bodies should periodically re-test and validate their models on different demographic groups to observe whether any groups are being systematically advantaged or disadvantaged, so that they can update their AI systems where necessary.

Many machine learning systems also refine the way that they process data to improve accuracy over time. Such refinements may distort the original goal of the AI system, so public bodies will need to monitor whether an AI system is achieving its intended purpose. The continuous refinement of AI systems could also be a problem if the system is deployed in an environment where the user can alter its performance and does so maliciously. For example, Microsoft's chatbot Tay was designed to learn from interactions it had with real people on Twitter in 2016. When users decided to feed it racist and offensive information, it learned to interact that way itself.[93] Chatbots that dispense advice on behalf of public bodies will have significant effects on citizens. These systems will need to be subject to consistent monitoring and evaluation to ensure they are not corrupted by human interaction, either intentionally or by accident.

> "Another concern is when you have systems that continue to learn through interaction with the user. There is the potential for a user to either maliciously poison the training data or to be mischievous in the way that they train the system thereby influencing the way it develops in the future."
> **Fiona Butcher, Fellow, Defence, Science and Technology Laboratory, Ministry of Defence**

Deploying AI is a process, not a single event. Once established, public bodies will need to keep a close eye on their AI systems to ensure that they continue to operate as intended. Deriving the best outcomes from AI will require a continuous process of tweaking and moderating the way an AI system operates.

Recommendation 12:

Providers of public services, both public and private, should monitor and evaluate their AI systems to ensure they always operate as intended.

93 Dave Lee (2016), 'Tay: Microsoft issues apology over racist chatbot fiasco', BBC News. Available at: https://www.bbc.co.uk/news/technology-35902104

5.5.3. Internal and external oversight

Human oversight of AI is a standards imperative. To ensure that public bodies remain accountable for automated decision-making, there needs to be internal control over the AI system, its decision-making process and its outcomes. Senior leadership should oversee the entire end-to-end AI process, to ensure that potential violations of human rights and public standards do not occur. While monitoring and evaluation requires a detailed look at data input and output, proper oversight involves leadership taking a bird's eye view of an entire system, including its design, governance and outcomes.

To have complete control over their AI systems, senior leadership need to have oversight over the whole AI process, from the point of data entry to the implementation of an AI-assisted decision.

Currently, a number of civil society organisations exercise, at a distance, oversight of AI systems. Organisations such as Liberty and Big Brother Watch have been prominent in scrutinising live facial recognition and predictive policing technologies. While these organisations are a vital part of democratic accountability, some contributors expressed doubts that civil society alone can provide meaningful oversight when AI is deployed across government at scale. Other external oversight mechanisms, such as ethics committees, will probably be necessary, as will good regulation, as discussed in chapter 4.

> "It is unclear whether civil society organisations have the capacity to engage in meaningful oversight, particularly given the rapidity with which different systems are being deployed across the sector and across the world."[94]
> **Law Society Report, Algorithms in the Criminal Justice System**

Oversight is important for proper accountability as the perspectives of those running AI systems on a day-to-day basis might be quite limited. For example, an individual whose role it is to decide whether to accept or reject an AI decision may not be aware of how data input has affected the outcome. Similarly, those building datasets to feed into an algorithm may not be aware of how their input selections could adversely affect decisions made later on. Effective oversight will help public bodies identify misuse and other unintended consequences of AI.

Specialist oversight bodies are useful tools for ensuring that difficult ethical issues relating to AI are given proper consideration. In high-risk areas such as health, policing or criminal justice, the use of an independent ethics committee would help ensure that issues around openness, accountability and objectivity are considered by individuals with the necessary knowledge and expertise. They also provide an independent level of assurance and are less likely to be subject to conflicts of interest. Some contributors to the review also noted that formal ethics committees could help build public trust for new technologies.

94 The Law Society (2019), Algorithms in the Criminal Justice System. Available at: https://www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/

"We use oversight bodies to assure ourselves that we have consent from the public because we know that the people who are most likely to be adversely affected by AI are less likely to come forward and present their views. We use oversight bodies, scrutiny panels and independent advisory groups to be representative of those communities."
Superintendent Chris Todd, West Midlands Police

Public bodies need to choose oversight mechanisms that are appropriate for the systems they are developing. If the risk to individuals is low, internal oversight of AI by senior leadership may be sufficient. In other, more risky policy areas, external scrutiny may be necessary. One contributor suggested that public bodies could appoint an independent ethics officer to oversee these considerations.

Contributors also suggested that public bodies should be required to act on recommendations made by independent oversight bodies, so that they had real powers of scrutiny. The Committee agrees, and is of the view that public bodies should act on any recommendations made by independent oversight bodies and set oversight mechanisms that allow their AI systems to be properly scrutinised.

Recommendation 13:

Providers of public services, both public and private, should set oversight mechanisms that allow for their AI systems to be properly scrutinised.

5.5.4. Appeal and redress

To remain accountable for their decisions, public bodies need to enable people to challenge decisions and to seek redress using procedures that are independent and transparent. This is the case whether AI is involved in the decision-making process or not. This is because public bodies and organisations carrying out public functions have to act in accordance with public and administrative law principles, and must act lawfully, rationally, proportionately and fairly.[95] Public law allows citizens to contribute to a public body's decision-making process, through consultation, and to challenge individual decisions where they have been made.

Many public bodies have complaints procedures that individuals can follow. Where complaints cannot be resolved, individuals usually have access to independent and impartial advice through an ombudsman scheme, and almost all decisions made by public bodies that have an impact on citizens carry a statutory right of appeal. This means that decisions will need to be explained and justified to a tribunal or other independent body, irrespective of whether AI is used. Appeals can generally look at whether the decision was made in accordance with the law and make findings of fact. Individuals can also ask public bodies to review their decisions in certain circumstances. If there is no right of appeal, complaints procedure, ombudsman scheme or review process – or if those things do not adequately address the problem – individuals may be able to challenge a decision by judicial review.

95 Public Law Project (2018), An introduction to public law. Available at: https://publiclawproject.org.uk/wp-content/uploads/2018/07/An-introduction-to-public-law-1.pdf

The existing appeals process should be utilised for those wishing to appeal against automated decisions. Public bodies should continue to make available fair and transparent avenues of redress for individuals who have been adversely affected by a decision, even if that decision was automated. Whether AI is used or not, public bodies should continue to uphold existing principles of administrative justice. For example, public bodies should: (1) make users and their needs central when making decisions; (2) enable people to challenge decisions and seek redress in ways that are independent, fair and transparent; and (3) keep people informed and empower them to resolve their problems as quickly and comprehensively as possible.[96] Public bodies should also ensure that their mechanisms for redress continue to be proportionate and efficient, and lead to well-reasoned, lawful and timely outcomes.

Public bodies will need to be able to explain and justify decisions made by AI technology. This means that they need to be auditable and transparent enough to satisfy a proper process of appeal and redress. Audits are necessary to discover how AI systems work and make decisions. Public bodies need to be able to track the process by which a system was designed, procured and deployed, and should be able to trace the way an automated decision was made. A decision that adversely affects an individual may be down to the failure of any one of these stages or a combination of them. A meaningful process of redress should enable public bodies to find out what failed and how that failure can be rectified.

Recommendation 14:

Providers of public services, both public and private, must always inform citizens of their right and method of appeal against automated and AI-assisted decisions.

5.5.5. Training and education

Contributors to this review consistently emphasised the importance of training and education. Public officials will need new skills and knowledge to ensure that high public standards are upheld in an AI-enabled public sector. Those using AI at all levels will need to be taught how AI works and be educated about the ethical risks of AI systems.

Contributors told this review that the risks to standards were greater if the decision to introduce AI was poorly informed. Without the right knowledge, senior management may be unaware, for example, of the potential for AI to amplify historic bias in their policy area.

Working with the right skills to assess AI

When identifying whether AI is the right solution, it's important that you work with:

- specialists who have a good knowledge of your data and the problem you're trying to solve, such as data scientists
- at least one domain knowledge expert who knows the environment where you will be deploying the AI model results.[97]

Office for AI Guidance, Assessing if artificial intelligence is the right solution

96 Administrative Justice and Tribunals Council (2010), Principles for Administrative Justice

97 Government Digital Service and Office for AI (2019), Assessing if artificial intelligence is the right solution. Available at: https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution

Those commissioning public services will need to know the technical capabilities of AI systems to assess the risks to standards posed by AI. Commissioners of public services will require a reasonable level of technical knowledge to judge whether the level of explainability a system offers matches the need for explanations in their policy area, for example.

Those operating AI systems will need to understand how their precise system operates to identify errors, ensure data is input correctly, and exercise discretion when implementing AI-enabled decisions. If operators of AI systems are not suitably trained, it would be unreasonable to hold them accountable for accepting or rejecting an AI decision.

> "From the perspective of the judiciary or the courts, I think education is the starting point… we are going to have to do a lot of work to develop effective training, knowledge systems and skills systems, to enable judges as well the Court Service staff to understand the implications of the operations of the systems."
> **John Sorabji, Principal Legal Adviser to the Lord Chief Justice and Master of the Rolls**

Training and education should happen before an AI system is deployed, but it should not be a one-off event. AI experts told this review that, like any new technology, AI is still in a period of rapid change. Individual systems themselves would be continuously upgraded and their capabilities enhanced. Training and education will have to keep up with these changes. It is important, therefore, that training and education is an ongoing process throughout the lifecycle of an AI system and not a one-off event.

Recommendation 15:

Providers of public services, both public and private, should ensure their employees working with AI systems undergo continuous training and education.

Citizens will also need to be informed about how artificial intelligence will change the way they engage with public services. Government should publicise information on citizens' data rights and facilitate better public understanding of how AI-enabled public services will operate. Public engagement on AI will help increase trust in government innovation and ensure citizens do not feel disempowered by new technology. It is part of the CDEI's remit to lead on public engagement and the Committee believes this should be a vital part of the Centre's role in public life.

Appendix 1: About the Committee on Standards in Public Life

The Committee on Standards in Public Life (CSPL, the Committee) advises the Prime Minister on ethical standards across the whole of public life in England. It monitors and reports on issues relating to the standards of conduct of all public office-holders. The Committee is an advisory non-departmental public body sponsored by the Cabinet Office. The Chair and members are appointed by the Prime Minister.

The Committee was established in October 1994, by the then Prime Minister, with the following terms of reference:

"To examine current concerns about standards of conduct of all holders of public office, including arrangements relating to financial and commercial activities, and make recommendations as to any changes in present arrangements which might be required to ensure the highest standards of propriety in public life."

The remit of the Committee excludes investigation of individual allegations of misconduct.

On 12 November 1997, the terms of reference were extended by the then Prime Minister: "To review issues in relation to the funding of political parties, and to make recommendations as to any changes in present arrangements."

The terms of reference were clarified following the Triennial Review of the Committee in 2013. The then Minister for the Cabinet Office confirmed that the Committee "should not inquire into matters relating to the devolved legislatures and governments except with the agreement of those bodies", and that "the Government understands the Committee's remit to examine 'standards of conduct of all holders of public office' as encompassing all those involved in the delivery of public services, not solely those appointed or elected to public office."

The Committee is a standing committee. It not only conducts inquiries into areas of concern about standards in public life, but can also revisit those areas to monitor whether and how well its recommendations have been put into effect.

Membership of the Committee for the period of this review

Lord (Jonathan) Evans KCB DL, Chair
The Rt Hon Dame Margaret Beckett DBE MP
The Rt Hon Jeremy Wright QC MP (from 21 November 2019)
The Rt Hon Simon Hart MP (until July 2019)
Dr Jane Martin CBE
Jane Ramsey
Dame Shirley Pearce DBE
Monisha Shah
The Rt Hon Lord (Andrew) Stunell OBE

Chair of Committee's Research Advisory Board

Professor Mark Philp

Secretariat

The Committee is assisted by a Secretariat consisting of Lesley Bainsfair (Secretary to the Committee), Amy Austin (Senior Policy Adviser), Ally Foat (Senior Policy Adviser), Nicola Richardson (Senior Policy Adviser), Aaron Simons (Senior Policy Adviser) and Lesley Glanz (Executive Assistant). Press support is provided by Maggie O'Boyle.

Appendix 2: Terms of reference

The terms of reference for the Committee's review into artificial intelligence and standards are to:

1. Consider whether existing frameworks and regulations are sufficient to ensure that standards are upheld as technologically assisted decision-making is adopted more widely in the public sector, including:

    a. examining the current use of artificial intelligence and associated advanced technologies in the public sector
    b. exploring how standards may be affected by the widespread introduction of these technologies into the public sector
    c. examining what safeguards and considerations of standards are currently in place in technology procurement processes in the public sector
    d. examining what safeguards and considerations of standards are currently in place in the deployment of AI and advanced technologies within the public sector
    e. examining what safeguards and considerations of standards are currently in place in private sector organisations developing AI services intended for use in the public sector.

2. Examine how provisions for standards can be built into the development, commissioning and deployment of new technologies in the public sector.

3. Consider to what extent the use of artificial intelligence and associated advanced technology has implications for our understanding and formulation of the Seven Principles of Public Life.

4. Make recommendations for how standards can be maintained in the public sector where advanced technologies are increasingly used for service delivery, including best practice guidance and regulatory change where necessary.

Appendix 3: Methodology

The Committee used a range of methods as part of its evidence gathering for its review, including:

- 50 individual stakeholder meetings and conference calls
- 3 roundtable seminars
- 19 written submissions
- polling and focus group research
- desk research, including a review of relevant academic texts, think tank reports, government and parliamentary reviews, and media coverage
- attending AI roundtables and conferences hosted by external organisations.

Stakeholder meetings

The Committee and Secretariat held 50 meetings and conference calls with individual stakeholders.

Name	Organisation
Carly Kind and Olivia Varley-Winter	Ada Lovelace Institute
Professor Edward Harcourt	Arts and Humanities Research Council (AHRC)
Emily Commander	Arts and Humanities Research Council (AHRC)
Tabitha Goldstaub	AI Council
Dr Adrian Weller, Dr David Leslie, Dr Florian Ostmann and Dr Ricardo Silva, Dr Brent Mittelstadt	The Alan Turing Institute
Dr David Halpern and Aisling Ní Chonaire	Behavioural Insights Team
Silkie Carlo	Big Brother Watch
Gillian Stamp	Bioss International
Crofton Black	The Bureau of Investigative Journalism
Roger Taylor, Alex Lawrence-Archer, Oliver Buckley, Bethan Charnley and Michael Birtwistle	Centre for Data Ethics and Innovation (CDEI)
Richard Thomas CBE and Bojana Bellamy	Centre for Information Policy Leadership
Chief Rabbi Ephraim Mirvis	Chief Rabbi of the United Hebrew Congregations of the Commonwealth
Niall Quinn	Crown Commercial Service
Joe Baddeley, Sam Roberts and Natalia Domagala	Department for Digital, Culture, Media and Sport (DCMS)
Rebecca Hilsenrath and Andrew Harding	Equality and Human Rights Commission (EHRC)
Steve Unger	Flint Global
Jacob Turner	Fountain Court Chambers
Matthew Cain, Robert Miller, Liz Harrison and Suki Binjal	Hackney Council
Apollo Gerolymbos	London Fire Brigade

Name	Organisation
Christophe Prince	Home Office
Lord Clement-Jones	House of Lords APPG on AI
Professor Nick Jennings	Imperial College London
Simon McDougall	Information Commissioner's Office
Chief Constable Alan Pughsley	Kent Police
Christina Blacklaws and Alexandra Cardenas	The Law Society
Dr Rune Nyrup	Leverhulme Centre for the Future of Intelligence, University of Cambridge
Eddie Copeland	London Office of Technology (LOTI)
Professor Richard Susskind	IT Adviser to the Lord Chief Justice
Dr Pearse Keane	Moorfields Eye Hospital
Sarah Wilkinson	NHS Digital
Matthew Gould	NHSX
Jacob Beswick, Tim Cook and Sabine Gerdon	Office for AI
Professor Sandra Wachter	Oxford Internet Institute, University of Oxford
Dr Jonathan Bright	Oxford Internet Institute, University of Oxford
Professor Luciano Floridi	Oxford Internet Institute, University of Oxford
James Loft and Nisha Deo	Rainbird Technologies
Alexander Babuta	Royal United Service Institute (RUSI)
Simon Dennis	SAS Institute
Zee Kin Yeong	Singapore Infocomm Media Development Authority
Ed Humpherson	UK Statistics Authority
Professor Alastair Denniston	University of Birmingham
Professor Karen Yeung and Professor Andrew Howes	University of Birmingham
Professor Charles Raab	University of Edinburgh
Professor Dame Wendy Hall	University of Southampton
Superintendent Iain Donnelly	West Midlands Police
Thomas McNeil	Strategic Adviser to the West Midlands Police and Crime Commissioner

Roundtable seminars

The Committee held three roundtable seminars in London as part of this review. Transcripts of the roundtables are available on the Committee's website.

Roundtable for Practitioners, Government and Public Service Providers, held on 23 May 2019, at 1 Horse Guards Road, London

Lord (Jonathan) Evans	Chair, Committee on Standards in Public Life (CSPL)
Dame Shirley Pearce	Independent Member, Committee on Standards in Public Life (CSPL)
Jane Ramsey	Independent Member, Committee on Standards in Public Life (CSPL)
Monisha Shah	Independent Member, Committee on Standards in Public Life (CSPL)
Oliver Buckley	Executive Director, Centre for Data Ethics and Innovation (CDEI)
Fiona Butcher	Science and Technology Lab, Ministry of Defence
Bethan Charnley	Innovation Policy Lead, Government Digital Service
Jimmy Elliott	General Counsel, SAS Institute
Sabine Gerdon	Project Lead, Office for AI
Sana Khareghani	Head of the Office for AI
Alex Lawrence-Archer	Chief Operating Officer, CDEI
Simon McDougall	Executive Director for Tech, Policy and Innovation, ICO
Ian O'Gara	Digital Strategy Director (Public Sector), Accenture
Marion Oswald	Senior Fellow, Department of Law, University of Winchester

Roundtable for Academics and Policy Experts, held on 29 May 2019, at Imperial College London.

Lord (Jonathan) Evans	Chair, Committee on Standards in Public Life (CSPL)
Jane Ramsey	Independent Member, Committee on Standards in Public Life (CSPL)
Monisha Shah	Independent Member, Committee on Standards in Public Life (CSPL)
Professor Mark Philp	Chair, Research Advisory Board, Committee on Standards in Public Life (CSPL)
Professor Nick Jennings	Vice Provost, Imperial College
Alexander Babuta	Research Fellow, National Security Studies, RUSI
Professor Alan Brown	Professor in Digital Economy, University of Exeter
Alexandra Cardenas	Head of Commercial and Technology Law, The Law Society
Jamie Grace	Senior Lecturer in Law, Sheffield Hallam University
Professor Edward Harcourt	Director of Research, AHRC, UKRI
Professor Philip Howard	Director, Oxford Internet Institute
Samantha McGregor	Head of Creative Industries, Digital Arts and Humanities, AHRC, UKRI
Professor Charles Raab	Professorial Fellow, University of Edinburgh
Chief Superintendent Chris Todd	NPCC lead for Data Analytics, and Director of Intelligence at West Midlands Police
Peter Wells	Director of Public Policy, Open Data Institute

Roundtable for Academics and Policy Experts, held on 5 June 2019, at Admiralty House, London

Lord (Jonathan) Evans	Chair, Committee on Standards in Public Life (CSPL)
Dame Shirley Pearce	Independent Member, Committee on Standards in Public Life (CSPL)
Monisha Shah	Independent Member, Committee on Standards in Public Life (CSPL)
Dr Reuben Binns	Postdoctoral Fellow in AI, ICO
Dr Jonathan Bright	Senior Research Fellow and Political Scientist, Oxford Internet Institute, University of Oxford
Professor Lizzie Coles-Kemp	Professor for Information Security, Royal Holloway University
Emily Commander	Head of Public Policy, AHRC, UKRI
David Evans	Director of Public Affairs, Goodfaith
Professor Anthony Finkelstein	UK Government Chief Scientific Adviser for National Security, and Chair of Software Science, UCL
Professor Andrew Howes	Head of Computer Science, University of Birmingham
Professor Helen Margetts	Professor Helen Margetts, Director for Public Policy, The Alan Turing Institute
Dr Brent Mittelstadt	Research Fellow, Oxford Internet Institute, University of Oxford
Professor Paul Nightingale	Director of Strategy and Operations, Economic and Social Research Council
Dr John Sorabji	Principal Legal Adviser to the Lord Chief Justice and Master of the Rolls
Andrew Yell	Global Supplier Manager, Farnell

Written submissions

The Committee received written submissions and additional written material from 20 individuals and organisations. No formal public consultation was held for this review.

Crofton Black, The Bureau of Investigative Journalism
Dr Emma Carmel, University of Bath
British Computer Society (BCS)
Carnegie UK Trust
Centre for Data Ethics and Innovation (CDEI)
Centre for Information Policy Leadership (CIPL)
Chartered Institute of Public Relations (CIPR)
Mission and Public Affairs Council, The Church of England
Robin Allen QC and Dee Masters, Cloisters Chambers
Ditto AI
David Evans, Good Faith Partnership
The Information Commissioner's Office (ICO)
Jamie Grace, Senior Lecturer in Law, Sheffield Hallam University
Dr Rune Nyrup, Dr Jess Whittlestone and Professor Stephen Cave, Leverhulme Centre for the Future of Intelligence, University of Cambridge
Christopher Marsh
MedConfidential
Marion Oswald, Senior Fellow, Department of Law, University of Winchester
The Royal College of Physicians
SAS Institute
Professor Karen Yeung, University of Birmingham

Polling and focus group research

The Committee commissioned Deltapoll to run quantitative and qualitative research. Polling and focus groups examined attitudes towards AI. One focus group was held with members of the general public and one with front-line public sector officials. A report, focus group transcripts, and full data tables are available on the Committee's website.

DeltaPoll Survey Results

Sample Size: 2,016 GB Adults
Fieldwork: 14-17 June 2019

QV1 I would like you to think about the form of artificial intelligence (AI) that is advanced computer data analytics. This type of AI is computer software that analyses millions of data points of information, finds patterns within that data, and uses those patterns to come to a conclusion or insight about something within our world.

How comfortable or uncomfortable would you be, if at all, if decisions in the government and public sector were made using AI in each of the following scenarios?

QV1_1 AI is used to devise a care plan for a 7 year old with special educational needs.

Reaction	%
Very comfortable	9
Quite comfortable	25
Quite uncomfortable	30
Very uncomfortable	21
Don't know	15
Comfortable (All)	34
Uncomfortable (All)	51
Net comfort	-17

QV1_2 AI is used to evaluate if a prisoner should be released from jail, by predicting the chance the prisoner will reoffend.

Reaction	%
Very comfortable	6
Quite comfortable	16
Quite uncomfortable	28
Very uncomfortable	39
Don't Know	12
Comfortable (All)	22
Uncomfortable (All)	67
Net comfort	-45

QV1_3 AI is used to understand medical scans and diagnose cancer.

Reaction	%
Very comfortable	15
Quite comfortable	38
Quite uncomfortable	21
Very uncomfortable	15
Don't Know	11
Comfortable (All)	53
Uncomfortable (All)	36
Net comfort	+17

QV1_4 AI is used to identify fraud in immigration checks.

Reaction	%
Very comfortable	21
Quite comfortable	43
Quite uncomfortable	15
Very uncomfortable	11
Don't know	10
Comfortable (All)	64
Uncomfortable (All)	26
Net comfort	+38

QV1_5 AI is used to predict if petty criminals are likely to commit serious gun or knife crime.

Reaction	%
Very comfortable	12
Quite comfortable	28
Quite uncomfortable	26
Very uncomfortable	21
Don't know	12
Comfortable (All)	40
Uncomfortable (All)	47
Net comfort	-7

QV1_6 AI is used to scan CVs for unqualified applicants.

Reaction	%
Very comfortable	14
Quite comfortable	39
Quite uncomfortable	21
Very uncomfortable	12
Don't know	13
Comfortable (All)	53
Uncomfortable (All)	33
Net comfort	+20

QV2 Thinking of the previous scenarios shown, how confident are you that the government and public sector will use AI in an ethical way?

Reaction	%
Very confident	5
Quite confident	26
Not very confident	37
Not confident at all	16
Don't know	15
Confident (All)	31
Not Confident (All)	53
Net Confidence	-22

QV3 Thinking of the above scenarios, which, if any, of the following would make you more comfortable with AI being used?

QV3_1 There is an easy-to-understand explanation for the AI software's decision.

Reaction	%
This would make me much more comfortable with AI being used	14
This would make me a bit more comfortable with AI being used	37
This would make no difference to me	33
This would make me less comfortable with using AI	3
Don't know	13
More comfortable (All)	51

QV3_2 A human operator always has the final say on whether to accept or reject an AI decision.

Reaction	%
This would make me much more comfortable with AI being used	31
This would make me a bit more comfortable with AI being used	38
This would make no difference to me	18
This would make me less comfortable with using AI	4
Don't know	9
More comfortable (All)	69

QV3_3 The AI has been deemed acceptable by a government ethics regulator.

Reaction	%
This would make me much more comfortable with AI being used	10
This would make me a bit more comfortable with AI being used	26
This would make no difference to me	42
This would make me less comfortable with using AI	11
Don't know	11
More comfortable (All)	36

QV3_4 You have the right to appeal against an AI decision to a human specialist.

Reaction	%
This would make me much more comfortable with AI being used	25
This would make me a bit more comfortable with AI being used	41
This would make no difference to me	20
This would make me less comfortable with using AI	5
Don't know	9
More comfortable (All)	66

QV3_5 The AI is known to have a 95% accuracy and success rate.

Reaction	%
This would make me much more comfortable with AI being used	16
This would make me a bit more comfortable with AI being used	38
This would make no difference to me	28
This would make me less comfortable with using AI	7
Don't know	11
More comfortable (All)	54

QV3_6 You understand clearly how the AI works.

Reaction	%
This would make me much more comfortable with AI being used	19
This would make me a bit more comfortable with AI being used	32
This would make no difference to me	30
This would make me less comfortable with using AI	4
Don't know	14
More comfortable (All)	51

Committee on Standards in Public Life
Room GC.07, 1 Horse Guards Road, London, SW1A 2HQ
Tel: 020 7271 2685
Email: public@public-standards.gov.uk

February 2020

INSIDE AI

Black-Boxed Politics:

Opacity is a Choice in AI Systems

Katarzyna Szymielewicz
Jan 17 · 23 min read

Written by: Agata Foryciarz, Daniel Leufer, Katarzyna Szymielewicz

Illustrations by: Olek Modzelewski

Artificial intelligence captures our imagination like almost no other technology: from fears about killer robots to dreams of a fully-automated, frictionless future. As numerous authors have documented, the idea of creating artificial, intelligent machines has entranced and scandalized people for millennia. Indeed, part of what makes the history of 'artificial intelligence' so fascinating is the mix of genuine scientific achievement with myth-making and outright deception.

A certain amount of hype and myth making can be harmless, and might even help to fuel real progress in the field. However, the fact that 'AI systems' are now being integrated into essential public services and other high-risk processes means that we must be especially vigilant about combatting misconceptions about AI.

At various points throughout 2019, we saw users of Amazon's Alexa, Google's Assistant, and Apple's Siri being shocked to discover that recordings of their private family conversations were being reviewed by real living humans. This was hardly surprising to anyone familiar with how these voice assistants are trained. But to the majority of customers, who do not question the presentation of these systems as 100% automated, it came as a shock that poorly paid overseas workers had access to what were often intimate and sensitive conversations. Concerns about how such systems operate are only sharpened when we see contracts between Amazon and Britain's National Health Service for Alexa to provide medical advice and, of course, to thereby have access to patient data.

The myths and misconceptions of the black box

> *There are many myths and misconceptions about AI, but in cases where these systems are being used in sensitive, high-risk scenarios such as public health and criminal justice, arguably the most damaging misconception is that these systems are 'black boxes' about which we simply cannot know anything.*

Worse still, the claim is often made that non-interpretable models have superior performance (which, as Cynthia Rudin points out in her work, is often false) and that demanding explanations will lead to a reduction in accuracy and effectiveness. We should thus simply trust, and even learn to embrace, the black box.

But what precisely do we mean when we say that an AI system is a black box? In their recently published guidance on 'Explaining AI decisions' (hereafter ICO-AT Guidance), the UK's ICO and Alan Turing Institute define black boxes as "any AI system whose inner workings and rationale are opaque or inaccessible to human understanding" and then provide a detailed explanation of the interpretability of different technical approaches[1]. There are, however, different causes of such opacity.

On the technical side, we typically refer to certain approaches to AI (such as neural networks) as black boxes because they operate in a manner which is simply too complex for a human to follow in detail or retrace after the fact. However, an AI system can also become a 'black box' because of its proprietary nature: we are prevented from knowing about how it works due to concerns about protecting trade secrets, or to stop the system 'being gamed' (the possibility of which is often a sign of a badly designed, arbitrary system). There are, of course, proprietary deep learning systems which combine both forms of opacity.

> *What we want to demonstrate in this article is that neither form of opacity, the technical or proprietary, is entirely inevitable, and that even when some essential element of opacity remains about a system, there are a whole host of transparency measures which can and should be taken to remove as much opacity as possible.*

Rebranding Statistics

Before we get into the technical and even legal sources of AI opacity, there is one fundamental source of opacity that we must address: the very definition of AI. The term 'artificial intelligence' is currently used to refer to a bewildering range of technologies. At the most mundane end of the scale, the hype around AI has led

companies to describe things as banal as recommendation features in Microsoft PowerPoint as AI, while at the most outrageous end of the scale we have speculations about non-value aligned Superintelligent AI overlords engaging in conflicts with alien species. What, if anything, unites the range of technologies that now fall under this label?

There is arguably one basic idea underlying the project of creating 'artificial intelligence': to make a machine that can accomplish a complex task. Such complex tasks are typically ones that require human intelligence to solve, but there is no reason to define AI solely in terms of mimicking human intelligence. We may want to develop machines to replicate how the human brain works as a way to understand our own minds better, but also to create machines that can solve problems which would be impossible for humans and which have an entirely 'inhuman intelligence'. In highly speculative cases where the goal would be to develop AI system with actual self-awareness or consciousness, we could potentially speak of the AI system itself having aims, goals or intentions. Of course, nothing of the sort is on the horizon, and remains the stuff of science fiction.

> *The AI models which are the subject of today's hype, meanwhile, are just statistical models (also known as machine learning models) — not unlike those which have been in use by social scientists, biologists, statisticians and psychologists for years. Those statistical models have, for decades, been used for tasks such as predicting future values of financial stocks, estimating the effects of treatments on health outcomes based on collected data and recognizing written text from images.*

The main difference in this new "AI era" is the increased efficacy that some machine learning models were able to achieve, thanks to technological advances in processing power and access to large collections of data. These models include neural networks — a class of algorithms which have been studied since the 1950s, but which only recently found successful applications to tasks such as image recognition and machine translation. However, other statistical models which have been around for decades — such as linear or logistic regression, decision trees, support vector machines — are now being rebranded as "AI", leading to increased enthusiasm in their efficacy — while their uses and capabilities remain largely identical.

> *Many well-known applications which have been publicly discussed because of exhibited bias were in fact *not* technical black boxes (i.e. did not rely on neural networks or other highly complex machine learning techniques).*

In such cases we are looking at much less sophisticated, interpretable techniques such as linear and logistic regression, decision trees/rule lists or case-based reasoning. There is certainly some level of mystification going on when these techniques are all lumped under the umbrella term of AI, and the first question we should ask when presented with 'AI systems' is precisely what techniques they are employing.

It is a common mistake made by non-expert commentators and journalists to apply the same 'black box' narrative to simple and complex systems alike. As a result, designers of simple systems also get excused for the lack of transparency. In many cases, the public is kept in the dark not because the inner workings of the system are obscure but because transparency would threaten trade secrets or expose controversial choices made by the owners of the 'AI system'.

This dynamic is a good reason in itself to question the 'black box' narrative and educate the public, so that not all statistical models land in the same black box. Bearing in mind then that today's AI, and indeed the only type of AI that is on any realistic horizon of development, is nothing more than advanced statistical models, let's first examine how non-technical factors can turn potentially interpretable AI systems into black boxes.

iBorderCTRL: when even the ethics report is opaque

A perfect example of this type of non-technical opacity can be seen in the case of iBorderCtrl, a project funded by the European Commission's Horizon 2020 funding initiative. iBorderCtrl claims to offer an AI-based lie detector service to help police the borders of the European Union. Numerous commentators have pointed to the lack of scientific basis for such technology, and Pirate Party MEP and civil liberties activist Patrick Breyer has launched a freedom of information request, which the European Commission has tried to dismiss.

Breyer requested that the relevant authorities make public certain documents, including an ethics assessment, but they refused "on the grounds that the ethics report and PR strategy are "commercial information" of the companies involved and of "commercial value". In this case, an extremely controversial and scientifically dubious technology (AI lie detection) is being used in an extremely sensitive context (migration) and is being funded by public money. It seems to defy common sense that any company providing such technology could be allowed to escape public scrutiny here.

> *Neither the European Commission nor the developers of iBorderCTRL have provided any evidence to suggest that there is any technical reason why we cannot inspect how their software works.*

However, this system becomes a 'black box' because the software is being protected from scrutiny due to concerns about trade secrets. This seems particularly egregious in a case where the technology is so controversial and the risks of discrimination and injustice are so high. Moreover, such a flagrant disregard for transparency is totally at odds with the idea of 'Trustworthy AI' which the EU is so keen to promote.

In contrast to this approach of keeping ethics assessments hidden from public scrutiny, the Council of Europe has recommended that human rights impact assessments (HRIAs) must be conducted by public authorities and must be publicly available:

> *Public authorities should not acquire AI systems from third parties in circumstances where the third party is unwilling to waive restrictions on information (e.g. confidentiality or trade secrets) where such restrictions impede or frustrate the process of (i) carrying out HRIAs (including carrying out external research/review), and (ii) making HRIAs available to the public[2].*

Such measures would do a great deal to eliminate unnecessary opacity from the use of AI systems by public authorities.

The sad reality, however, is that we not only have no access to ethics reports, HRIAs, or the technical specifications of such systems — we do not even know if and where such systems are in use. Much of our knowledge of what systems are actually in use comes from ground work done by investigative journalists and civil society organisations such as Algorithm Watch, a German NGO which has produced a report mapping the use of 'automated decision making' in several European countries.

Eliminating these unnecessary obstacles to transparency would be a basic first step that public authorities could take. Knowing what systems are being used in the public realm (and often being funded by public money) is a necessary condition for effective oversight, but there are also many more ways in which opacity can be tackled in AI systems.

Why we need to prioritize explainability in AI systems

The discussion about what explanations should be required and at what level of detail is ongoing.

> *In their draft guidance on 'Explaining AI decisions' (hereafter ICO-AT Guidance), the UK's ICO and Alan Turing Institute advise that the intelligibility and interpretability of an AI model should be prioritised from the outset and that end-to-end transparency and accountability should be optimised above other parameters. In other words, whenever possible, a company or a public institution that aims to automate decisions that will affect humans should use a model that can be interpreted — not a technical black box.*

Making this choice from the outset would address some of our problems — making it easier for those using AI systems as decision support to reason about their limitations, for those affected by the systems to dispute their incorrect decisions, and for society to have better oversight over the way in which they are used. But we can safely assume that there will be companies and public institutions alike willing to make a different choice: for dubious beliefs about opaque systems automatically leading to improved accuracy; for PR purposes to attract funding and interest because their company is using the most advanced models; out of convenience by just letting the neural network do all the work to avoid complex preprocessing work; or even for the convenience of evading scrutiny by deliberately picking an opaque model. What can we, the concerned public, do with these cases? Should we simply acknowledge that there will always be systems which function in opaque ways?

Indeed, it can be extremely difficult and time consuming to reverse engineer exactly *how* an AI system has made a decision — for example, how exactly the combinations of the individual pixel patterns of an image lead to the system deciding that it has seen a cat. Machine learning practitioners use the term 'local explanation' to refer to an interpretation of individual predictions or classifications. Such explanations can have different levels of granularity, and would typically approximate the process through which the algorithm generates a response, rather than attempting to reconstruct it in all its minutiae.

To provide such an explanation, one can focus on identifying specific input variables that had the most influence in generating a particular prediction or classification — often a difficult task if a prediction or classification was produced by a neural network (although for some neural network models, tools do already exist that can help generate local explanations). This difficulty is often used as a general caveat to rebuff any further conversation about the logic behind an AI system or its fairness. But we should never accept an obstacle in producing "local", case-by-case explanations as a reason to not produce any explanation at all.

> *Instead, following the ICO-AT Guidance once more, we can demand that owners of AI systems that qualify as 'black boxes' provide supplementary explanations which can shed light on the logic behind the results and behaviour of their system.*

These should include internal model explanations, including model type and architecture, the data used to train the models, the results of "stress tests" — describing the model sensitivity under a variety of cases — as well as the results of internal reverse-engineering, providing examples of local explanations for chosen inputs whenever possible. Frameworks for generating such descriptions (including Model Cards for Model Reporting or Datasheets for Datasets) are widely known among data scientists and often used for companies' internal purposes.

The ICO-AT Guidance shows that a lot can be done to explain the inner workings of any AI-assisted system, if those involved are just willing to make the effort. We will not go into the technical details of how to produce internal or post-hoc explanations in this text, as it has been extensively covered by the ICO-Turing Guidance.

The point that we want to reiterate, however, echoing the reasoning of the UK authorities, is that **there are numerous decisions that owners or designers of an AI-assisted system need to make**, starting with the framing of the problem they want to solve and ending with the choice of model they will use and its evaluation method. These decisions should be well-documented and justified because **this is where the conversation about the values embedded into AI systems really starts: with human decisions regarding the design and optimisation of the whole system**.

How human decisions shape every AI system

The key point that we are making here is that explanations of automated decisions need not hinge on the general public understanding how algorithmic systems function.

> *In order to maintain a level of scrutiny over an AI system we do not have to understand every step in the machine learning process. What we do have to understand, however, are the choices, assumptions and trade-offs made by the people who designed this system — which all shape the behavior of the algorithm. This level of explainability can be achieved without opening the technical black box.*

Every algorithm and every ML-assisted decision-making system is designed by humans to achieve certain goals. These goals are not defined by 'artificial intelligence', but rather by humans who designed it, who decide why an automated decision-making

system is needed in the first place and what problem or question it is supposed to solve. Such a system can be employed to help humans find patterns in vast amounts of data, which is what AI systems do best. AI systems can also be employed to support humans in making judgements, based on predictions — or to replace a human decision-maker completely. There are relatively benign examples of automating judgement, such as spam detection or the choice of targeted advertising for consumer goods, and more controversial ones such as choosing targeted housing or job advertisements or detecting hate speech — where historical patterns of discrimination are more likely to be replicated, and where potential errors have serious consequences.

Finally, and most controversially, these systems can also be used to predict certain outcomes, including future human behaviour. Assuming that those outcomes follow clear patterns and that we have sufficient and appropriate data to discover them, this may be an achievable task. Such ideal conditions are rarely present, however — leading many to be opposed to the use of AI systems to predict things such as criminal recidivism or to replace human interviewers in choosing the ideal candidates for jobs.

Regardless of the task or function attributed to an AI system, engineers always begin with a problem to be solved: identifying cats and dogs in a huge dataset composed of pictures, for the sake of producing better search results; predicting the current interest or mood of a person browsing the internet, in order to maximise the chance that he or she will click on suggested content and remain engaged; or predicting whether a defendant is likely to commit a crime in the future.

Unpacking what the system really does: technical choices

> *In the case of each and every ML system, its goals (what it has been set to achieve/optimise for) can be extrapolated from a set of technical and design decisions. In order to understand how a machine learning system functions — and under what circumstances it is likely to fail — we need to first understand the task it was given or the question it was designed to answer.*

We can use sophisticated methods to 'interrogate' the system to detect the actual task it was given, but we can also simply ask the humans who designed it to tell us. What we want to argue is that this information should be available to the public — without having to ask for it.

So let us assume that we know *why* an AI assisted system was implemented — to identify customers most likely to buy a certain product; to maximize click-through rate

or to maximize how long users spend on a website. Are we satisfied? Not yet. There is usually a long path between defining an overall objective and calibrating the system so that it delivers specific, desirable outcomes. In order to verify whether the system achieves its original aim, and is unlikely to behave in undesirable (e.g. discriminatory) ways, we have to understand and verify key technical and design choices.

Whenever we are dealing with a complex task or question, **system designers have to make certain assumptions when translating (or "operationalizing") a general goal into mathematical formulas or functions.** It is one thing to say "we want users to stay engaged with online content", it is another thing to define what "engagement" really means, how it is measured and what data about an individual can be used to predict his or her online behavior. It is one thing to define a goal as "matching social benefits with individuals who really need them", and another thing to actually define who is "in need", formulate the optimal allocation of outcomes, and encode unacceptable worst-case outcomes, acceptable tradeoffs and evaluation procedures in order to produce a fair result.

> *A general goal behind an AI system must be translated from business or political language into the language of mathematical formulas. These decisions and choices made "internally" by data scientists are no less meaningful than the choice of a general goal — just as the details of how a public policy was implemented are no less meaningful than the overall objective of that policy.*

In fact, a success or failure of an AI system in fulfilling its general goal depends to a large extent on these smaller, "internal" decisions, often made by data scientists alone, and never communicated to the public. We argue that these technical decisions are especially crucial to understand when we reason about fairness and impact of ML systems that are applied to humans.

What are these choices? To illustrate a typical decision-making process in the development of an AI-system, we imagined the story of a data scientist tasked with building a model meant to help hospital administrators select people who will be enrolled in a health management program.

One such model — used in hospitals across the United States — has recently fallen under public scrutiny after a Science article revealed that the model systematically recommends healthier White patients over Black patients with higher medical need for enrollment. The effect, as the authors explain, can be traced back to the choice to

define patients with high health needs as those who also generate high healthcare costs — an assumption that is known to be wrong for Black patients, on whose treatment American hospitals spend less.

Our story is fictionalized and simplified, but, as a matter of fact, similar scenarios happen every day.

The Story of a Data Scientist

Jasmine is a data scientist working for a large university hospital. She works closely with the hospital management, working on multiple projects — analyzing trends in spending and medical procedure data and building statistical models to help the management and doctors gain a better insight into how redirecting resources to different patients and departments will affect spending, patient health and employee satisfaction.

One day, Jasmine is in a meeting with the management, where they discuss a newly established government program which provides the hospital with additional resources to help manage the health of patients with significant health needs. The program offers monthly meetings with a nutritionist, physical therapy, weekly, free-of-charge psychotherapy, as well as a personal program coordinator who is available 24/7 to support the patient and help them navigate their healthcare. The program was established to support elderly and diabetic patients, but it is at each hospital's discretion to select patients who will enter the program. There are 50 spots, for over 1,200 patients served by the hospital.

The management is very excited about the additional resources, but one of the senior doctors brings up that the selection of 50 most needy patients may be challenging. Should they select those with poorest health? Those who do not have relatives or spouses who help manage their healthcare? Should they concentrate on the elderly or on diabetic patients, or make space for young people with eating disorders? How do they make sure the selection process is uniform across the hospital? Additionally, since the program starts four months after the patient selection has to be made, how can they best assign patients who will still need support at that later date? It seems difficult and time-consuming to coordinate this selection process across all the doctors in the hospital.

"There is a system we use in the hospital which helps us predict the costs a patient will generate in the following year", says Jasmine. She suggests this system could be used in this case. Her team could also build a custom one, which could predict specific health conditions. It would take some more time, but could reflect the goals of the management better, and save the hospital money in the long run.

The management agrees that a custom model would be more appropriate in this case — they want to avoid using the proprietary system,

especially given the recent Science article which showed using cost prediction as a proxy for health leads to an unfair allocation of resources between people of different ethnicities.

But some key questions remain. The meeting time is coming to an end, but Jasmine asks everybody to stay a few extra minutes and writes out a few bullet points on the whiteboard. What is the definition of a "needy" patient we can all agree on? A model will not stand ambiguity, she says — and it is best that doctors, not data scientists, make the decision. What potential errors do we need to make sure to avoid? Models are known to perpetuate discrimination, and the hospital cannot afford to be called out as discriminatory for not assigning women, people of color or poorer patients to the program, she warns. What are the acceptable trade offs — should we prioritize finding all patients in need, even if some will be included incorrectly, or should we avoid including healthy patients, knowing this will lead to us missing some sick patients? And what data should we be using — will using sensitive data in the analysis make it more difficult for management to access and use the model, given privacy regulations?

All these questions need to be answered in the course of Jasmine's team's work, and the answers they settle on will get translated into mathematical formulas that will define the functioning of the system.

Let's now break those decisions — and their mathematical formulations — into separate components to see how they are embedded throughout the process of building an automated prediction model.

There will be trade-offs. The process of translating real-world problems into mathematical formulas always includes simplifications, assumptions and tradeoffs. AI metrics must correspond to quantifiable phenomena, which do not always correspond to human definitions of success in a given scenario, and their reliability is limited by the mathematical properties of the models we use. For example, we have to accept that the system calibrated to achieve maximum accuracy will often not perform equally well in making sure that no group is discriminated against (as in the case of the COMPAS algorithm). As Cathy O'Neil put it in A RedTail interview: "You can't minimize false positives, maximize accuracy and minimize false negatives all at once. There are always trade-offs, so this is a way of surfacing those trade-offs in ways that it doesn't take a math PhD to do."

> *No matter how well-intentioned system owners are, and how much they might wish for a fair outcome, the outcome the system delivers will be determined not by their wishful thinking but by mathematical and statistical constraints.*

Data does not always reflect the real world. AI systems are as good as the data they have been trained on. Above we explain how human decisions shape an AI system but also show the limitations of what can be done.

> *Even the best model will not perform well if it was trained on datasets that are only remotely connected to the phenomena the system is supposed to predict.*

We have to keep in mind that many phenomena that people may be interested in predicting — such as "health level" or "recidivism" — do not have simple definitions corresponding to measurements present in datasets. In these cases, data scientists depend on proxy variables — 'close replacements' for the variables of interests, such as the number of active chronic conditions or number of arrests (we can measure if people were rearrested after being released on bail, but 'being arrested' is not a good substitute

for 'having committed a crime,' especially when we take over-policing of minorities into account).

If there is bias, there was a human decision behind it. Many studies that look at "discriminatory" algorithms explain bias by bringing external factors such as poor data quality ("the only data that was available to train the system reflected the same bias" or "data scientists missed data about a certain group") and a flawed choice of objective. Although such reasons may seem extrinsic, arising from external realities that designers of the system in question could not control, the truth is that there is a human decision behind each of these choices.

Owners of AI systems and the data scientists they employ are responsible for the choices made at each stage of development: for the choice of the data, even if that choice was very limited; for deploying the system despite the fact that they could not avoid bias; for not revising their main objective, regardless of the fact that the fair outcome they hoped for could not be achieved; and, finally, they are responsible for choosing to use an automated system in the first place, despite being aware of its limitations and possible consequences.

Not an isolated decision, but a decision-making cycle. Every well-managed process of designing and training an AI system involves rethinking objectives, improving the quality of data, changing the model and re-calibrating the system so that it does what it was supposed to be doing — or abandoning it altogether if it does not achieve desired objectives. It is not a one-way process.

Designers and data scientists have to iterate, come back and rethink the design. Although there are no legal requirements for logging the reasons behind design decisions, in practice this circular process is quite common and consistent between data scientists. However, the decision about when to stop this iterative process is not merely a technical or business decision.

Indeed, in cases where the AI system will have significant impact on people's lives, this decision becomes political. For example, how can a software engineer alone decide that a predictive model is sufficiently good in not discriminating against minorities?

> *How can such decisions be left to the intuitions of people without expertise in these topics and, most importantly, how can it be that there is no possibility of public scrutiny about such significant decisions?*

Conclusion: If the 'technical' is political, we need democratic control over these decisions

At the end of her book, *Automating Inequality*, Virginia Eubanks quotes a data science evangelist who dreams of replacing civil servants with AI systems:

"The information and insights will be immediate, real time, bespoke and easy to compare over time. And, ideally, agreed by all to be **perfectly apolitical**."

What would it mean to replace civil servants, who are responsible for incredibly sensitive decisions with a massive impact on people's lives, with *perfectly apolitical* AI systems? It would in fact mean **hiding politics in the black box**. It would mean obscuring the fact that arbitrary human decisions went into constructing these systems at every stage of their development, and it would mean masking these decisions — whether they be about granting bail or deciding if a child should be separated from their parents — beneath a facade of objectivity. In the process, such decisions are seemingly depoliticised, and **this depoliticisation is very much in the interest of those who wish to avoid accountability for their political choices**.

Enough cases of dangerous errors leading to discrimination and safety concerns have been reported in recent years to understand the potential implications, which in the most acute circumstances may mean life or death. A lot has been said about their negative impact after it happened, but key choices made by the owners of these systems or their teams — with regard to (sources of) data, the model used or the loss function that led to such results — were not debated in public. As Ruha Benjamin points out, we have a tendency to see glitches and failures in such systems as "a fleeting interruption of an otherwise benign system," as merely technological problems, whereas we should see them as "rather a kind of signal of how the system operates."[3]

> *If these decisions had been transparent from the start instead of only when things go wrong, and the key stakeholders had been involved and consulted, many of these negative impacts could have been avoided.*

Many designers of AI systems care deeply about the impact of their work, and are often drawn to AI by imagining the potential of using these technologies for social good. Researchers and practitioners alarmed by publicized cases of machine learning bias now take extra steps to ensure their algorithms are fair. However, while admirable, these motivations are not enough — especially as claims of "social good" and "fairness" implicate normative, political choices, as there is no single agreed upon definition of

"social good" and there are at least 21 definitions of "fairness" — many of them mutually exclusive. While in low-stakes scenarios those decisions may be innocuous, they become political in nature once they are made in the context of systems that will impact human lives.

While we might never be able to agree on the definition of "good" or "fair" AI, we should agree on the process of evaluating and discussing the political choices made by the designers of these systems. Currently the field lacks the language and perspective to evaluate and debate these decisions. This allows computer scientists to make broad claims about solving social challenges while avoiding rigorous engagement with the social and political impacts — which often leads to adding complexity to, if not exacerbating, the very challenges they hoped to address.

> *If we agree that in the context of AI implementations technical often becomes political, the next step is to call for transparency of these decisions and some form of democratic oversight.*

While full public scrutiny over every design decision in every AI system being designed today is neither feasible nor necessary, we can — and should — demand transparency in the technical choices behind AI systems used to make decisions that affect humans so that we can see, and potentially challenge, the political decisions that inform technical choices — particularly when used in the public sector. Without such transparency, we are denied the possibility of political participation, as contentious decisions are hidden inside the 'black box' narrative.

—

Initiatives and guidelines that inspired us:

- Explaining AI decisions in practice
- Model Cards for Model Reporting
- Datasheets for Datasets
- Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead
- Accountability of AI Under the Law: The Role of Explanation
- A Framework for Understanding Unintended Consequences of Machine Learning

- Problem Formulation and Fairness
- Roles for Computing in Social Change
- The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning
- The fallacy of inscrutability

Footnotes

[1] See Part 2: Explaining AI in practice, p.40–48 for more details on the interpretability of various types of models.

[2] Unboxing artificial intelligence: 10 steps to protect human rights, p.8.

[3] Ruha Benjamin, *Race after Technology*, p.165.

EUROPEAN
COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

White Paper on Artificial Intelligence
A European approach to excellence and trust

Artificial Intelligence is developing fast. It will change our lives by improving healthcare (e.g. making diagnosis more precise, enabling better prevention of diseases), increasing the efficiency of farming, contributing to climate change mitigation and adaptation, improving the efficiency of production systems through predictive maintenance, increasing the security of Europeans, and in many other ways that we can only begin to imagine. At the same time, Artificial Intelligence (AI) entails a number of potential risks, such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes.

Against a background of fierce global competition, a solid European approach is needed, building on the European strategy for AI presented in April 2018[1]. To address the opportunities and challenges of AI, the EU must act as one and define its own way, based on European values, to promote the development and deployment of AI.

The Commission is committed to enabling scientific breakthrough, to preserving the EU's technological leadership and to ensuring that new technologies are at the service of all Europeans – improving their lives while respecting their rights.

Commission President Ursula von der Leyen announced in her political Guidelines[2] a coordinated European approach on the human and ethical implications of AI as well as a reflection on the better use of big data for innovation.

Thus, the Commission supports a regulatory and investment oriented approach with the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology. The purpose of this White Paper is to set out policy options on how to achieve these objectives. It does not address +the development and use of AI for military purposes.The Commission invites Member States, other European institutions, and all stakeholders, including industry, social partners, civil society organisations, researchers, the public in general and any interested party, to react to the options below and to contribute to the Commission's future decision-making in this domain.

1. INTRODUCTION

As digital technology becomes an ever more central part of every aspect of people's lives, people should be able to trust it. Trustworthiness is also a prerequisite for its uptake. This is a chance for Europe, given its strong attachment to values and the rule of law as well as its proven capacity to build safe, reliable and sophisticated products and services from aeronautics to energy, automotive and medical equipment.

Europe's current and future sustainable economic growth and societal wellbeing increasingly draws on value created by data. AI is one of the most important applications of the data economy. Today most data are related to consumers and are stored and processed on central cloud-based infrastructure. By contrast a large share of tomorrow's far more abundant data will come from industry, business and the public sector, and will be stored on a variety of systems, notably on computing devices working at the edge of the network. This opens up new opportunities for Europe, which has a strong position in

[1] AI for Europe, COM/2018/237 final
[2] https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf.

digitised industry and business-to-business applications, but a relatively weak position in consumer platforms.

Simply put, AI is a collection of technologies that combine data, algorithms and computing power. Advances in computing and the increasing availability of data are therefore key drivers of the current upsurge of AI. Europe can combine its technological and industrial strengths with a high-quality digital infrastructure and a regulatory framework based on its fundamental values to **become a global leader in innovation in the data economy and its applications** as set out in the European data strategy[3]. On that basis, it can develop an AI ecosystem that brings the benefits of the technology to the whole of European society and economy:

- for **citizens** to reap new benefits for example improved health care, fewer breakdowns of household machinery, safer and cleaner transport systems, better public services;
- for **business** development, for example a new generation of products and services in areas where Europe is particularly strong (machinery, transport, cybersecurity, farming, the green and circular economy, healthcare and high-value added sectors like fashion and tourism); and
- for services of **public interest**, for example by reducing the costs of providing services (transport, education, energy and waste management), by improving the sustainability of products[4] and by equipping law enforcement authorities with appropriate tools to ensure the security of citizens[5], with proper safeguards to respect their rights and freedoms.

Given the major impact that AI can have on our society and the need to build trust, it is vital that European AI is grounded in our values and fundamental rights such as human dignity and privacy protection.

Furthermore, the impact of AI systems should be considered not only from an individual perspective, but also from the perspective of society as a whole. The use of AI systems can have a significant role in achieving the Sustainable Development Goals, and in supporting the democratic process and social rights. With its recent proposals on the European Green Deal[6], Europe is leading the way in tackling climate and environmental-related challenges. Digital technologies such as AI are a critical enabler for attaining the goals of the Green Deal. Given the increasing importance of AI, the environmental impact of AI systems needs to be duly considered throughout their lifecycle and across the entire supply chain, e.g. as regards resource usage for the training of algorithms and the storage of data.

A common European approach to AI is necessary to reach sufficient scale and avoid the fragmentation of the single market. The introduction of national initiatives risks to endanger legal certainty, to weaken citizens' trust and to prevent the emergence of a dynamic European industry.

This White Paper presents policy options to enable a trustworthy and secure development of AI in Europe, in full respect of the values and rights of EU citizens. The main building blocks of this White Paper are:

---

[3] COM(2020) 66 final.

[4] AI and digitalisation in general are critical enablers of Europe's Green deal ambitions. However, the current environmental footprint of the ICT sector is estimated at more than 2% of all global emissions. The European digital strategy accompanying this White Paper proposes green transformation measures for digital.

[5] AI tools can provide an opportunity for better protecting EU citizens from crime and acts of terrorism. Such tools could, for example, help identify online terrorist propaganda, discover suspicious transactions in the sales of dangerous products, identify dangerous hidden objects or illicit substances or products, offer assistance to citizens in emergencies and help guide first responders.

[6] COM(2019) 640 final.

- The policy framework setting out measures to align efforts at European, national and regional level. In partnership between the private and the public sector, the aim of the framework is to mobilise resources to achieve an **'ecosystem of excellence'** along the entire value chain, starting in research and innovation, and to create the right incentives to accelerate the adoption of solutions based on AI, including by small and medium-sized enterprises (SMEs).
- The key elements of a future regulatory framework for AI in Europe that will create a unique **'ecosystem of trust'.** To do so, it must ensure compliance with EU rules, including the rules protecting fundamental rights and consumers' rights, in particular for AI systems operated in the EU that pose a high risk[7]. Building an ecosystem of trust is a policy objective in itself, and should give citizens the confidence to take up AI applications and give companies and public organisations the legal certainty to innovate using AI. The Commission strongly supports a human-centric approach based on the Communication on Building Trust in Human-Centric AI[8] and will also take into account the input obtained during the piloting phase of the Ethics Guidelines prepared by the High-Level Expert Group on AI.

The European strategy for data, which accompanies this White Paper, aims to enable Europe to become the most attractive, secure and dynamic data-agile economy in the world – empowering Europe with data to improve decisions and better the lives of all its citizens. The strategy sets out a number of policy measures, including mobilising private and public investments, needed to achieve this goal. Finally, the implications of AI, Internet of Things and other digital technologies for safety and liability legislation are analysed in the Commission Report accompanying this White Paper.

2. CAPITALISING ON STRENGTHS IN INDUSTRIAL AND PROFESSIONAL MARKETS

Europe is well placed to benefit from the potential of AI, not only as a user but also as a creator and a producer of this technology. It has excellent research centres, innovative start-ups, a world-leading position in robotics and competitive manufacturing and services sectors, from automotive to healthcare, energy, financial services and agriculture. Europe has developed a strong computing infrastructure (e.g. high-performance computers), essential to the functioning of AI. Europe also holds large volumes of public and industrial data, the potential of which is currently under-used. It has well-recognised industrial strengths in safe and secure digital systems with low-power consumption that are essential for the further development of AI.

Harnessing the capacity of the EU to invest in next generation technologies and infrastructures, as well as in digital competences like data literacy, will increase Europe's technological sovereignty in key enabling technologies and infrastructures for the data economy. The infrastructures should support the creation of European data pools enabling trustworthy AI, e.g. AI based on European values and rules.

Europe should leverage its strengths to expand its position in the ecosystems and along the value chain, from certain hardware manufacturing sectors to software all the way to services. This is already happening to an extent. Europe produces more than a quarter of all industrial and professional service robots (e.g. for precision farming, security, health, logistics.), and plays an important role in developing and using software applications for companies and organisations (business-to-business applications such as Enterprise Resource Planning, design and engineering software) as well as applications to support e-government and the "intelligent enterprise".

[7] Although further arrangements may need to be put in place to prevent and counter misuse of AI for criminal purposes, this is outside the scope of this white paper.

[8] COM(2019) 168.

Europe leads the way in deploying AI in manufacturing. Over half of the top manufacturers implement at least one instance of AI in manufacturing operations[9].

One reason for Europe's strong position in terms of research is the EU funding programme that has proven instrumental in pooling action, avoiding duplications, and leveraging public and private investments in the Member States. Over the past three years, EU funding for research and innovation for AI has risen to €1.5 billion, i.e. a 70% increase compared to the previous period.

However, investment in research and innovation in Europe is still a fraction of the public and private investment in other regions of the world. Some €3.2 billion were invested in AI in Europe in 2016, compared to around €12.1 billion in North America and €6.5 billion in Asia[10]. In response, Europe needs to increase its investment levels significantly. The Coordinated plan on AI[11] developed with Member States is proving to be a good starting point in building closer cooperation on AI in Europe and in creating synergies to maximise investment in the AI value chain.

3. SEIZING THE OPPORTUNITIES AHEAD: THE NEXT DATA WAVE

Although Europe currently is in a weaker position in consumer applications and on online platforms, which results in a competitive disadvantage in data access, major shifts in the value and re-use of data across sectors are underway. The volume of data produced in the world is growing rapidly, from 33 zettabytes in 2018 to an expected 175 zettabytes in 2025[12]. Each new wave of data brings opportunities for Europe to position itself in the data-agile economy and to become a world leader in this area. Furthermore, the way in which data are stored and processed will change dramatically over the coming five years. Today 80% of data processing and analysis that takes place in the cloud occurs in data centres and centralised computing facilities, and 20% in smart connected objects, such as cars, home appliances or manufacturing robots, and in computing facilities close to the user ("edge computing"). By 2025 these proportions are set to change markedly[13].

Europe is a global leader in low-power electronics which is key for the next generation of specialised processors for AI. This market is currently dominated by non-EU players. This could change with the help of initiatives such as the European Processor Initiative, which focuses on developing low-power computing systems for both edge and next generation high-performance computing, and the work of the Key Digital Technology Joint Undertaking, proposed to start in 2021. Europe also leads in neuromorphic solutions[14] that are ideally suited to automating industrial processes (industry 4.0) and transport modes. They can improve energy efficiency by several orders of magnitude.

Recent advances in quantum computing will generate exponential increases in processing capacity[15]. Europe can be at the forefront of this technology thanks to its academic strengths in quantum computing, as well as European industry's strong position in quantum simulators and programming environments for quantum computing. European initiatives that aim to increase the availability of quantum testing and experimentation facilities will help apply these new quantum solutions to a number of industrial and academic sectors.

---

[9] Followed by Japan (30%) and the US (28%). Source: CapGemini (2019).
[10] 10 imperatives for Europe in the age of AI and automation, McKinsey (2017).
[11] COM(2018) 795.
[12] IDC (2019).
[13] Gartner (2017).
[14] Neuromorphic solutions means any very large-scale system of integrated circuits that mimic neuro-biological architectures present in the nervous system.
[15] Quantum computers will have the capacity to process in less than seconds many fold larger data sets than today's highest performance computers allowing for the development of new AI applications across sectors.

In parallel, Europe will continue to lead progress in the algorithmic foundations of AI, building on its own scientific excellence. There is a need to build bridges between disciplines that currently work separately, such as machine learning and deep learning (characterised by limited interpretability, the need for a large volume of data to train the models and learn through correlations) and symbolic approaches (where rules are created through human intervention). Combining symbolic reasoning with deep neural networks may help us improve explainability of AI outcomes.

4. AN ECOSYSTEM OF EXCELLENCE

To build an ecosystem of excellence that can support the development and uptake of AI across the EU economy and public administration, there is a need to step up action at multiple levels.

A. WORKING WITH MEMBER STATES

Delivering on its strategy on AI adopted in April 2018,[16] in December 2018 the Commission presented a Coordinated Plan - prepared together with the Member States - to foster the development and use of AI in Europe[17].

This plan proposes some 70 joint actions for closer and more efficient cooperation between Member States, and the Commission in key areas, such as research, investment, market uptake, skills and talent, data and international cooperation. The plan is scheduled to run until 2027, with regular monitoring and review.

The aim is to maximise the impact of investment in research, innovation and deployment, assess national AI strategies and build on and extend the Coordinated Plan on AI with Member States:

- *Action 1: The Commission, taking into account the results of the public consultation on the White Paper, will propose to the Member States a revision of the Coordinated Plan to be adopted by end 2020*

EU-level funding in AI should attract and pool investment in areas where the action required goes beyond what any single Member State can achieve. The objective is to attract over €20 billion[18] of total investment in the EU per year in AI over the next decade. To stimulate private and public investment, the EU will make available resources from the Digital Europe Programme, Horizon Europe as well as from the European Structural and Investment Funds to address the needs of less-developed regions as well as rural areas.

The Coordinated Plan could also address societal and environmental well-being as a key principle for AI. AI systems promise to help tackling the most pressing concerns, including climate change and environmental degradation. It is also important that this happens in an environmentally friendly manner. AI can and should itself critically examine resource usage and energy consumption and be trained to make choices that are positive for the environment. The Commission will consider options to encourage and promote AI solutions that do this together with the Member States.

B. FOCUSING THE EFFORTS OF THE RESEARCH AND INNOVATION COMMUNITY

[16] Artificial Intelligence for Europe, COM(2018) 237.
[17] Coordinated Plan on Artificial Intelligence, COM(2018) 795.
[18] COM(2018) 237.

Europe cannot afford to maintain the current fragmented landscape of centres of competence with none reaching the scale necessary to compete with the leading institutes globally. It is imperative to create more synergies and networks between the multiple European research centres on AI and to align their efforts to improve excellence, retain and attract the best researchers and develop the best technology. Europe needs a lighthouse centre of research, innovation and expertise that would coordinate these efforts and be a world reference of excellence in AI and that can attract investments and the best talents in the field.

The centres and the networks should concentrate in sectors where Europe has the potential to become a global champion such as industry, health, transport, finance, agrifood value chains, energy/environment, forestry, earth observation and space. In all these domains, the race for global leadership is ongoing, and Europe offers significant potential, knowledge and expertise[19]. Equally important is to create testing and experimentation sites to support the development and subsequent deployment of novel AI applications.

- *Action 2: the Commission will facilitate the creation of excellence and testing centres that can combine European, national and private investments, possibly including a new legal instrument. The Commission has proposed an ambitious and dedicated amount to support world reference testing centres in Europe under the Digital Europe Programme and complemented where appropriate by research and innovation actions of Horizon Europe as part of the Multiannual Financial Framework for 2021 to 2027.*

C. SKILLS

The European approach to AI will need to be underpinned by a strong focus on skills to fill competence shortages.[20] The Commission will soon present a reinforcement of the Skills Agenda, which aims to ensure that everyone in Europe can benefit from the green and digital transformations of the EU economy. Initiatives could also include the support of sectoral regulators to enhance their AI skills in order to effectively and efficiently implement relevant rules. The updated Digital Education Action Plan will help make better use of data and AI-based technologies such as learning and predictive analytics with the aim to improve education and training systems and make them fit for the digital age. The Plan will also increase awareness of AI at all levels of education in order to prepare citizens for informed decisions that will be increasingly affected by AI.

Developing the skills necessary to work in AI and upskilling the workforce to become fit for the AI-led transformation will be a priority of the revised Coordinated Plan on AI to be developed with Member States. This could include transforming the assessment list of the ethical guidelines into an indicative "curriculum" for developers of AI that will be made available as a resource for training institutions. Particular efforts should be undertaken to increase the number of women trained and employed in this area.

In addition, a lighthouse centre of research and innovation for AI in Europe would attract talent from all over the world due to the possibilities it could offer. It would also develop and spread excellence in skills that take root and grow across Europe.

[19] The future European Defence Fund and Permanent Structured Cooperation (PESCO) will also provide opportunities for research and development in AI. These projects should be synchronized with the wider EU civilian programmes devoted to AI.

[20] https://ec.europa.eu/jrc/en/publication/academic-offer-and-demand-advanced-profiles-eu

- *Action 3: Establish and support through the advanced skills pillar of the Digital Europe Programme networks of leading universities and higher education institutes to attract the best professors and scientists and offer world-leading masters programmes in AI.*

Beyond upskilling, workers and employers are directly affected by the design and use of AI systems in the workplace. The involvement of social partners will be a crucial factor in ensuring a human-centred approach to AI at work.

D. FOCUS ON SMES

It will also be important to ensure that SMEs can access and use AI. To this end, the Digital Innovation Hubs[21] and the AI-on-demand platform[22] should be strengthened further and foster collaboration between SMEs. The Digital Europe Programme will be instrumental in achieving this. While all Digital Innovation Hubs should provide support to SMEs to understand and adopt AI, it will be important that at least one innovation hub per Member State has a high degree of specialisation in AI.

SMEs and start-ups will need access to finance in order to adapt their processes or to innovate using AI. Building on the forthcoming pilot investment fund of €100 million in AI and blockchain, the Commission plans to further scale up access to finance in AI under InvestEU[23]. AI is explicitly mentioned among the eligible areas for the use of the InvestEU guarantee.

- *Action 4: the Commission will work with Member States to ensure that at least one digital innovation hub per Member State has a high degree of specialisation on AI. Digital Innovation Hubs can be supported under the Digital Europe Programme*.
- *The Commission and the European Investment Fund will launch a pilot scheme of €100 million in Q1 2020 to provide equity financing for innovative developments in AI. Subject to final agreement on the MFF, the Commission's intention is to scale it up significantly from 2021 through InvestEU.*

E. PARTNERSHIP WITH THE PRIVATE SECTOR

It is also essential to make sure that the private sector is fully involved in setting the research and innovation agenda and provides the necessary level of co-investment. This requires setting up a broad-based public private partnership, and securing the commitment of the top management of companies.

- *Action 5: In the context of Horizon Europe, the Commission will set up a new public private partnership in AI, data and robotics to combine efforts, ensure coordination of research and innovation in AI, collaborate with other public-private partnerships in Horizon Europe and work together with the testing facilities and the Digital Innovation Hubs mentioned above.*

[21] ec.europe.eu/digital-single-market/en/news/digital-innovation-hubs-helping-companies-across-economy-make-most-digital-opportunities.
[22] www.Ai4eu.eu.
[23] Europe.eu/investeu.

F. PROMOTING THE ADOPTION OF AI BY THE PUBLIC SECTOR

It is essential that public administrations, hospitals, utility and transport services, financial supervisors, and other areas of public interest rapidly begin to deploy products and services that rely on AI in their activities. A specific focus will be in the areas of healthcare and transport where technology is mature for large-scale deployment.

- *Action 6: The Commission will initiate open and transparent sector dialogues giving priority to healthcare, rural administrations and public service operators in order to present an action plan to facilitate development, experimentation and adoption. The sector dialogues will be used to prepare a specific 'Adopt AI programme' that will support public procurement of AI systems, and help to transform public procurement processes themselves.*

G. SECURING ACCESS TO DATA AND COMPUTING INFRASTRUCTURES

The areas for action set out in this White Paper are complementary to the plan presented in parallel under the European data strategy. Improving access to and the management of data is fundamental. Without data, the development of AI and other digital applications is not possible. The enormous volume of new data yet to be generated constitutes an opportunity for Europe to position itself at the forefront of the data and AI transformation. Promoting responsible data management practices and compliance of data with the FAIR principles will contribute to build trust and ensure re-usability of data[24]. Equally important is investment in key computing technologies and infrastructures.

The Commission has proposed more than €4 billion under the Digital Europe Programme to support high-performance and quantum computing, including edge computing and AI, data and cloud infrastructure. The European data strategy develops these priorities further.

H. INTERNATIONAL ASPECTS

Europe is well positioned to exercise global leadership in building alliances around shared values and promoting the ethical use of AI. The EU's work on AI has already influenced international discussions. When developing its ethical guidelines, the High-Level Expert Group involved a number of non-EU organisations and several governmental observers. In parallel, the EU was closely involved in developing the OECD's ethical principles for AI[25]. The G20 subsequently endorsed these principles in its June 2019 Ministerial Statement on Trade and Digital Economy.

In parallel, the EU recognises that important work on AI is ingoing in other multilateral fora, including the Council of Europe, the United Nations Educational Scientific and Cultural Organization (UNESCO), the Organisation for Economic Co-operation and Development's (OECD), the World Trade Organisation and the International Telecommunications Union (ITU). At the UN, the EU is involved in the follow-up of the report of the High-Level Panel on Digital Cooperation, including its recommendation on AI.

The EU will continue to cooperate with like-minded countries, but also with global players, on AI, based on an approach based on EU rules and values (e.g. supporting upward regulatory convergence, accessing key resources including data, creating a level playing field). The Commission will closely monitor the policies of third countries that limit data flows and will address undue restrictions in

[24] Findable, Accessible, Interoperable and Reusable as stated in the Final Report and Action Plan from the Commission Expert Group on FAIR data, 2018, https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf.

[25] https://www.oecd.org/going-digital/ai/principles/

bilateral trade negotiations and through action in the context of the World Trade Organization. The Commission is convinced that international cooperation on AI matters must be based on an approach that promotes the respect of fundamental rights, including human dignity, pluralism, inclusion, non-discrimination and protection of privacy and personal data[26] and it will strive to export its values across the world[27]. It is also clear that the responsible development and use of AI can be a driving force to achieve the Sustainable Development Goals and advance the 2030 Agenda.

5. AN ECOSYSTEM OF TRUST: REGULATORY FRAMEWORK FOR AI

As with any new technology, the use of AI brings both opportunities and risks. Citizens fear being left powerless in defending their rights and safety when facing the information asymmetries of algorithmic decision-making, and companies are concerned by legal uncertainty. While AI can help protect citizens' security and enable them to enjoy their fundamental rights, citizens also worry that AI can have unintended effects or even be used for malicious purposes. These concerns need to be addressed. Moreover, in addition to a lack of investment and skills, lack of trust is a main factor holding back a broader uptake of AI.

That is why the Commission set out an AI strategy[28] on 25 April 2018 addressing the socioeconomic aspects in parallel with an increase in investment in research, innovation and AI-capacity across the EU. It agreed a Coordinated Plan[29] with the Member States to align strategies. The Commission also established a High-Level Expert Group that published Guidelines on trustworthy AI in April 2019[30].

The Commission published a Communication[31] welcoming the seven key requirements identified in the Guidelines of the High-Level Expert Group:

- Human agency and oversight,
- Technical robustness and safety,
- Privacy and data governance,
- Transparency,
- Diversity, non-discrimination and fairness,
- Societal and environmental wellbeing, and
- Accountability.

In addition, the Guidelines contain an assessment list for practical use by companies. During the second half of 2019, over 350 organisations have tested this assessment list and sent feedback. The High-Level Group is in the process of revising its guidelines in light of this feedback and will finalise this work by June 2020. A key result of the feedback process is that while a number of the requirements are already reflected in existing legal or regulatory regimes, those regarding transparency, traceability and human oversight are not specifically covered under current legislation in many economic sectors.

On top of this set of non-binding Guidelines of the High-Level Expert Group, and in line with the President's political guidelines, a clear European regulatory framework would build trust among

[26] Under the Partnership Instrument, the Commission will finance a €2.5 million project that will facilitate cooperation with like-minded partners, in order to promote the EU AI ethical guidelines and to adopt common principles and operational conclusions.
[27] President Von der Leyen, A Union that strives for more – My agenda for Europe, page 17.
[28] COM(2018) 237.
[29] COM(2018) 795.
[30] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top
[31] COM(2019) 168.

consumers and businesses in AI, and therefore speed up the uptake of the technology. Such a regulatory framework should be consistent with other actions to promote Europe's innovation capacity and competitiveness in this field. In addition, it must ensure socially, environmentally and economically optimal outcomes and compliance with EU legislation, principles and values. This is particularly relevant in areas where citizens' rights may be most directly affected, for example in the case of AI applications for law enforcement and the judiciary.

Developers and deployers of AI are already subject to European legislation on fundamental rights (e.g. data protection, privacy, non-discrimination), consumer protection, and product safety and liability rules. Consumers expect the same level of safety and respect of their rights whether or not a product or a system relies on AI. However, some specific features of AI (e.g. opacity) can make the application and enforcement of this legislation more difficult. For this reason, there is a need to examine whether current legislation is able to address the risks of AI and can be effectively enforced, whether adaptations of the legislation are needed, or whether new legislation is needed.

Given how fast AI is evolving, the regulatory framework must leave room to cater for further developments. Any changes should be limited to clearly identified problems for which feasible solutions exist.

Member States are pointing at the current absence of a common European framework. The German Data Ethics Commission has called for a five-level risk-based system of regulation that would go from no regulation for the most innocuous AI systems to a complete ban for the most dangerous ones. Denmark has just launched the prototype of a Data Ethics Seal. Malta has introduced a voluntary certification system for AI. If the EU fails to provide an EU-wide approach, there is a real risk of fragmentation in the internal market, which would undermine the objectives of trust, legal certainty and market uptake.

A solid European regulatory framework for trustworthy AI will protect all European citizens and help create a frictionless internal market for the further development and uptake of AI as well as strengthening Europe's industrial basis in AI.

A. PROBLEM DEFINITION

While AI can do much good, including by making products and processes safer, it can also do harm. This harm might be both material (safety and health of individuals, including loss of life, damage to property) and immaterial (loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination for instance in access to employment), and can relate to a wide variety of risks. A regulatory framework should concentrate on how to minimise the various risks of potential harm, in particular the most significant ones.

The main risks related to the use of AI concern the application of rules designed to protect fundamental rights (including personal data and privacy protection and non-discrimination), as well as safety[32] and liability-related issues.

Risks for fundamental rights, including personal data and privacy protection and non-discrimination

[32] This includes issues of cybersecurity, issues associated with AI applications in critical infrastructures, or malicious use of AI.

The use of AI can affect the values on which the EU is founded and lead to breaches of fundamental rights[33], including the rights to freedom of expression, freedom of assembly, human dignity, non-discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation, as applicable in certain domains, protection of personal data and private life, [34] or the right to an effective judicial remedy and a fair trial, as well as consumer protection. These risks might result from flaws in the overall design of AI systems (including as regards human oversight) or from the use of data without correcting possible bias (e.g. the system is trained using only or mainly data from men leading to suboptimal results in relation to women).

AI can perform many functions that previously could only be done by humans. As a result, citizens and legal entities will increasingly be subject to actions and decisions taken by or with the assistance of AI systems, which may sometimes be difficult to understand and to effectively challenge where necessary. Moreover, AI increases the possibilities to track and analyse the daily habits of people. For example, there is a potential risk that AI may be used, in breach of EU data protection and other rules, by state authorities or other entities for mass surveillance and by employers to observe how their employees behave. By analysing large amounts of data and identifying links among them, AI may also be used to retrace and de-anonymise data about persons, creating new personal data protection risks even in respect to datasets that per se do not include personal data. AI is also used by online intermediaries to prioritise information for their users and to perform content moderation. The processed data, the way applications are designed and the scope for human intervention can affect the rights to free expression, personal data protection, privacy, and political freedoms.

Certain AI algorithms, when exploited for predicting criminal recidivism, can display gender and racial bias, demonstrating different recidivism prediction probability for women vs men or for nationals vs foreigners. *Source: Tolan S., Miron M., Gomez E. and Castillo C. "Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia", Best Paper Award, International Conference on AI and Law, 2019*

Certain AI programmes for facial analysis display gender and racial bias, demonstrating low errors for determining the gender of lighter-skinned men but high errors in determining gender for darker-skinned women. Source: *Joy Buolamwini, Timnit Gebru; Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018*.

Bias and discrimination are inherent risks of any societal or economic activity. Human decision-making is not immune to mistakes and biases. However, the same bias when present in AI could have a much larger effect, affecting and discriminating many people without the social control mechanisms that govern human behaviour[35]. This can also happen when the AI system 'learns' while in operation.

[33] Council of Europe research shows that a large number of fundamental rights could be impacted from the use of AI, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5.

[34] The General Data Protection Regulation and the ePrivacy Directive (new ePrivacy Regulation under negotiation) address these risks but there might be a need to examine whether AI systems pose additional risks. The Commission will be monitoring and assessing the application of the GDPR on a continuous basis.

[35] The Commission's Advisory Committee on Equal Opportunities for Women and Men is currently preparing an "Opinion on Artificial Intelligence" analysing inter alia the impacts of Artificial Intelligence on gender equality which is expected to be adopted by the Committee in early 2020. The EU Gender Equality Strategy 2020-2024 also addresses the link between AI on gender equality; The European Network of Equality Bodies (Equinet) will publish a report (by Robin

In such cases, where the outcome could not have been prevented or anticipated at the design phase, the risks will not stem from a flaw in the original design of the system but rather from the practical impacts of the correlations or patterns that the system identifies in a large dataset.

The specific characteristics of many AI technologies, including opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour, may make it hard to verify compliance with, and may hamper the effective enforcement of, rules of existing EU law meant to protect fundamental rights. Enforcement authorities and affected persons might lack the means to verify how a given decision made with the involvement of AI was taken and, therefore, whether the relevant rules were respected. Individuals and legal entities may face difficulties with effective access to justice in situations where such decisions may negatively affect them.

Risks for safety and the effective functioning of the liability regime

AI technologies may present new safety risks for users when they are embedded in products and services. For example, as result of a flaw in the object recognition technology, an autonomous car can wrongly identify an object on the road and cause an accident involving injuries and material damage. As with the risks to fundamental rights, these risks can be caused by flaws in the design of the AI technology, be related to problems with the availability and quality of data or to other problems stemming from machine learning. While some of these risks are not limited to products and services that rely on AI , the use of AI may increase or aggravate the risks.

A lack of clear safety provisions tackling these risks may, in addition to risks for the individuals concerned, create legal uncertainty for businesses that are marketing their products involving AI in the EU. Market surveillance and enforcement authorities may find themselves in a situation where they are unclear as to whether they can intervene, because they may not be empowered to act and/or don't have the appropriate technical capabilities for inspecting systems[36]. Legal uncertainty may therefore reduce overall levels of safety and undermine the competitiveness of European companies.

If the safety risks materialise, the lack of clear requirements and the characteristics of AI technologies mentioned above make it difficult to trace back potentially problematic decisions made with the involvement of AI systems. This in turn may make it difficult for persons having suffered harm to obtain compensation under the current EU and national liability legislation.[37]

Allen and Dee Masters) on "Regulating AI: the new role for Equality Bodies – Meeting the new challenges to equality and non-discrimination from increased digitalisation and the use of AI", expected early 2020.

[36] An example may be the smart watch for children. This product may cause no direct harm to the child wearing it, but lacking a minimum level of security, it can be easily used as a tool to have access to the child. Market surveillance authorities may find it difficult to intervene in cases where the risk is not linked to the product as such.

[37] The implications of AI, Internet of Things and other digital technologies for safety and liability legislation are analysed in the Commission Report accompanying this White Paper.

Under the Product Liability Directive, a manufacturer is liable for damage caused by a defective product. However, in the case of an AI based system such as autonomous cars, it may be difficult to prove that there is a defect in the product, the damage that has occurred and the causal link between the two. In addition, there is some uncertainty about how and to what extent the Product Liability Directive applies in the case of certain types of defects, for example if these result from weaknesses in the cybersecurity of the product.

Thus, the difficulty of tracing back potentially problematic decisions taken by AI systems and referred to above in relation to fundamental rights applies equally to safety and liability-related issues. Persons having suffered harm may not have effective access to the evidence that is necessary to build a case in court, for instance, and may have less effective redress possibilities compared to situations where the damage is caused by traditional technologies. These risks will increase as the use of AI becomes more widespread.

B. POSSIBLE ADJUSTMENTS TO EXISTING EU LEGISLATIVE FRAMEWORK RELATING TO AI

An extensive body of existing EU product safety and liability legislation[38], including sector-specific rules, further complemented by national legislation, is relevant and potentially applicable to a number of emerging AI applications.

As regards the protection of fundamental rights and consumer rights, the EU legislative framework includes legislation such as the Race Equality Directive[39], the Directive on equal treatment in employment and occupation[40], the Directives on equal treatment between men and women in relation to employment and access to goods and services[41], a number of consumer protection rules[42], as well as rules on personal data protection and privacy, notably the General Data Protection Regulation and other sectorial legislation covering personal data protection, such as the Data Protection Law Enforcement Directive[43]. In addition, as from 2025, the rules on accessibility requirements for goods and services, set out in the European Accessibility Act will apply[44]. In addition, fundamental rights need to be respected when implementing other EU legislation, including in the field of financial services, migration or responsibility of online intermediaries.

While the EU legislation remains in principle fully applicable irrespective of the involvement of AI, it is important to assess whether it can be enforced adequately to address the risks that AI systems create, or whether adjustments are needed to specific legal instruments.

---

38 The EU legal framework for product safety consists of the General Product Safety Directive (Directive 2001/95/EC), as a safety net, and a number of sector-specific rules covering different categories of products ranging from machines, planes and cars to toys and medical devices aiming to provide a high level of health and safety. Product liability law is complemented by different systems of civil liability for damages caused by products or services.

39 Directive 2000/43/EC.

40 Directive 2000/78/EC.

41 Directive 2004/113/EC; Directive 2006/54/EC.

42 Such as the Unfair Commercial Practices Directive (Directive 2005/29/EC) and the Consumer Rights Directive (Directive 2011/83/EC).

43 Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data.

44 Directive (EU) 2019/882 on the accessibility requirements for products and services.

For example, economic actors remain fully responsible for the compliance of AI to existing rules that protects consumers, any algorithmic exploitation of consumer behaviour in violation of existing rules shall be not permitted and violations shall be accordingly punished.

The Commission is of the opinion that the legislative framework could be improved to address the following risks and situations:

- *Effective application and enforcement of existing EU and national legislation*: the key characteristics of AI create challenges for ensuring the proper application and enforcement of EU and national legislation. The lack of transparency (opaqueness of AI) makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights, attribute liability and meet the conditions to claim compensation. Therefore, in order to ensure an effective application and enforcement, it may be necessary to adjust or clarify existing legislation in certain areas, for example on liability as further detailed in the Report, which accompanies this White Paper.

- *Limitations of scope of existing EU legislation*: an essential focus of EU product safety legislation is on the placing of products on the market. While in EU product safety legislation software, when is part of the final product, must comply with the relevant product safety rules, it is an open question whether stand-alone software is covered by EU product safety legislation, outside some sectors with explicit rules[45]. General EU safety legislation currently in force applies to products and not to services, and therefore in principle not to services based on AI technology either (e.g. health services, financial services, transport services).

- *Changing functionality of AI systems*: the integration of software, including AI, into products can modify the functioning of such products and systems during their lifecycle. This is particularly true for systems that require frequent software updates or which rely on machine learning. These features can give rise to new risks that were not present when the system was placed on the market. These risks are not adequately addressed in the existing legislation which predominantly focuses on safety risks present at the time of placing on the market.

- *Uncertainty as regards the allocation of responsibilities between different economic operators in the supply chain*: in general, EU legislation on product safety allocates the responsibility to the producer of the product placed on the market, including all components e.g. AI systems. But the rules can for example become unclear if AI is added after the product is placed on the market by a party that is not the producer. In addition, EU product liability legislation provides for liability of producers and leaves national liability rules to govern liability of others in the supply chain.

- *Changes to the concept of safety*: the use of AI in products and services can give rise to risks that EU legislation currently does not explicitly address. These risks may be linked to cyber threats, personal security risks (linked for example to new applications of AI such as to home appliances), risks that result from loss of connectivity, etc. These risks may be present at the time of placing products on the market or arise as a result of software updates or self-learning when the product is being used. The EU should make full use of the tools at its disposal to

[45] For instance software intended by the manufacturer to be used for medical purposes is considered a medical device under the Medical Device Regulation (Regulation (EU) 2017/745).

enhance its evidence base on potential risks linked to AI applications, including using the experience of the EU Cybersecurity Agency (ENISA) for assessing the AI threat landscape.

As indicated earlier, several Member States are already exploring options for national legislation to address the challenges created by AI. This raises the risk that the single market may be fragmented. Divergent national rules are likely to create obstacles for companies that want to sell and operate AI systems in the single market. Ensuring a common approach at EU level would enable European companies to benefit from smooth access to the single market and support their competitiveness on global markets.

Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics

The Report, which accompanies this White Paper, analyses the relevant legal framework. It identifies uncertainties as to the application of this framework with respect to the specific risks posed by AI systems and other digital technologies.

It concludes that the current product safety legislation already supports an extended concept of safety protecting against all kind of risks arising from the product according to its use. However, provisions explicitly covering new risks presented by the emerging digital technologies could be introduced to provide more legal certainty.

- The autonomous behaviour of certain AI systems during its life cycle may entail important product changes having an impact on safety, which may require a new risk assessment. In addition, human oversight from the product design and throughout the lifecycle of the AI products and systems may be needed as a safeguard.
- Explicit obligations for producers could be considered also in respect of mental safety risks of users when appropriate (ex. collaboration with humanoid robots).
- Union product safety legislation could provide for specific requirements addressing the risks to safety of faulty data at the design stage as well as mechanisms to ensure that quality of data is maintained throughout the use of the AI products and systems.
- The opacity of systems based on algorithms could be addressed through transparency requirements.
- Existing rules may need to be adapted and clarified in the case of a stand-alone software placed as it is on the market or downloaded into a product after its placing on the market, when having an impact on safety.
- Given the increasing complexity of supply chains as regards new technologies, provisions specifically requesting cooperation between the economic operators in the supply chain and the users could provide legal certainty.

The characteristics of emerging digital technologies like AI, the IoT and robotics may challenge aspects of the liability frameworks and could reduce their effectiveness. Some of these characteristics could make it hard to trace the damage back to a person, which would be necessary for a fault-based claim in accordance with most national rules. This could significantly increase the costs for victims and means that liability claims against others than producers may be difficult to make or prove.

- Persons having suffered harm caused with the involvement of AI systems need to enjoy the same level of protection as persons having suffered harm caused by other technologies, whilst technological innovation should be allowed to continue to develop.
- All options to ensure this objective should be carefully assessed, including possible amendments to the Product Liability Directive and possible further targeted harmonisation of national liability rules. For example, the Commission is seeking views whether and to what extent it may be needed to mitigate the consequences of complexity by adapting the burden of proof required by national liability rules for damage caused by the operation of AI applications.

From the discussion above, the Commission concludes that – in addition to the possible adjustments to existing legislation – a new legislation specifically on AI may be needed in order to make the EU legal framework fit for the current and anticipated technological and commercial developments.

C. SCOPE OF A FUTURE EU REGULATORY FRAMEWORK

A key issue for the future specific regulatory framework on AI intelligence is to determine the scope of its application. The working assumption is that the regulatory framework would apply to products and services relying on AI. AI should therefore be clearly defined for the purposes of this White Paper, as well as any possible future policy-making initiative.

In its Communication on AI for Europe the Commission provided a first definition of AI[46]. This definition was further refined by the High Level Expert Group[47].

In any new legal instrument, the definition of AI will need to be sufficiently flexible to accommodate technical progress while being precise enough to provide the necessary legal certainty.

For the purposes of this White Paper, as well as of any possible future discussions on policy initiatives, it seems important to clarify the main elements that compose AI, which are "data" and "algorithms". AI can be integrated in hardware. In case of machine learning techniques, which constitute a subset of AI, algorithms are trained to infer certain patterns based on a set of data in order to determine the actions needed to achieve a given goal. Algorithms may continue to learn when in use. While AI-based products can act autonomously by perceiving their environment and without following a pre-determined set of instructions, their behaviour is largely defined and constrained by its developers. Humans determine and programme the goals, which an AI system should optimise for.

> In autonomous driving for example, the algorithm uses, in real time, the data from the car (speed, engine consumption, shock-absorbers, etc..) and from the sensors scanning the whole environment of the car (road, signs, other vehicles, pedestrians etc..) to derive which direction, acceleration and speed the car should take to reach a certain destination. Based on the data observed, the algorithm adapts to the situation of the road and to the outside conditions, including other drivers' behaviour, to derive the most comfortable and safest drive.

The EU has a strict legal framework in place to ensure inter alia consumer protection, to address unfair commercial practices and to protect personal data and privacy. In addition, the acquis contains specific rules for certain sectors (e.g. healthcare, transport). These existing provisions of EU law will continue to apply in relation to AI, although certain updates to that framework may be necessary to reflect the digital transformation and the use of AI (see section B). As a consequence, those aspects that are

[46] COM(2018) 237 final, p. 1: "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.
AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."

[47] High Level Expert Group, A definition of AI, p. 8: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."

already addressed by existing horizontal or sectoral legislation (e.g. on medical devices[48], in transport systems) will continue to be governed by this legislation.

As a matter of principle, the new regulatory framework for AI should be effective to achieve its objectives while not being excessively prescriptive so that it could create a disproportionate burden, especially for SMEs. To strike this balance, the Commission is of the view that it should follow a risk-based approach.

A risk-based approach is important to help ensure that the regulatory intervention is proportionate. However, it requires clear criteria to differentiate between the different AI applications, in particular in relation to the question whether or not they are 'high-risk'[49]. The determination of what is a high-risk AI application should be clear and easily understandable and applicable for all parties concerned. Nevertheless even if an AI application is not qualified as high-risk, it remains entirely subject to already existing EU-rules.

The Commission is of the opinion that a given AI application should generally be considered high-risk in light of what is at stake, considering whether both the sector and the intended use involve significant risks, in particular from the viewpoint of protection of safety, consumer rights and fundamental rights. More specifically, an AI application should be considered high-risk where it meets the following two cumulative criteria:

- First, the AI application is employed in a sector where, given the characteristics of the activities typically undertaken, significant risks can be expected to occur. This first criterion ensures that the regulatory intervention is targeted on the areas where, generally speaking, risks are deemed most likely to occur. The sectors covered should be specifically and exhaustively listed in the new regulatory framework. For instance, healthcare; transport; energy and parts of the public sector.[50] The list should be periodically reviewed and amended where necessary in function of relevant developments in practice;

- Second, the AI application in the sector in question is, in addition, used in such a manner that significant risks are likely to arise. This second criterion reflects the acknowledgment that not every use of AI in the selected sectors necessarily involves significant risks. For example, whilst healthcare generally may well be a relevant sector, a flaw in the appointment scheduling system in a hospital will normally not pose risks of such significance as to justify legislative intervention. The assessment of the level of risk of a given use could be based on the impact on the affected parties. For instance, uses of AI applications that produce legal or similarly significant effects for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities.

The application of the two cumulative criteria would ensure that the scope of the regulatory framework is targeted and provides legal certainty. The mandatory requirements contained in the new regulatory framework on AI (see section D below) would in principle apply only to those applications identified as high-risk in accordance with these two cumulative criteria.

---

[48] For example, there are different safety considerations and legal implications concerning AI systems that provide specialized medical information to physicians, AI systems providing medical information directly to the patient and AI systems performing medical tasks themselves directly on a patient. The Commission is examining these safety and liability challenges that are distinct to healthcare.

[49] EU legislation may categorise "risks" differently to what is described here, depending on the area, such as for example, product safety

[50] The public sector could include areas like asylum, migration, border controls and judiciary, social security and employment services.

Notwithstanding the foregoing, there may also be exceptional instances where, due to the risks at stake, the use of AI applications for certain purposes is to be considered as high-risk as such – that is, irrespective of the sector concerned and where the below requirements would still apply.[51] As an illustration, one could think in particular of the following:

- In light of its significance for individuals and of the EU acquis addressing employment equality, the use of AI applications for recruitment processes as well as in situations impacting workers' rights would always be considered "high-risk" and therefore the below requirements would at all times apply. Further specific applications affecting consumer rights could be considered.

- the use of AI applications for the purposes of remote biometric identification[52] and other intrusive surveillance technologies, would always be considered "high-risk" and therefore the below requirements would at all times apply.

D. TYPES OF REQUIREMENTS

When designing the future regulatory framework for AI, it will be necessary to decide on the types of mandatory legal requirements to be imposed on the relevant actors. These requirements may be further specified through standards. As noted in section C above and in addition to already existing legislation, those requirements would apply to high-risk AI applications only, thus ensuring that any regulatory intervention is focused and proportionate.

Taking into account the guidelines of the High Level Expert Group and what has been set out in the foregoing, the requirements for high-risk AI applications could consist of the following key features, which are discussed in further detail in the subsections below:

- training data;
- data and record-keeping;
- information to be provided;
- robustness and accuracy;
- human oversight;
- specific requirements for certain particular AI applications, such as those used for purposes of remote biometric identification.

To ensure legal certainty, these requirements will be further specified to provide a clear benchmark for all the actors who need to comply with them.

a) Training data

It is more important than ever to promote, strengthen and defend the EU's values and rules, and in particular the rights that citizens derive from EU law. These efforts undoubtedly also extend to the high-risk AI applications marketed and used in the EU under consideration here.

[51] It is important to highlight that other pieces of EU legislation may also apply. For example, when incorporated into a consumer product, the General Product Safety Directive may apply to the safety of AI applications.

[52] Remote biometric identification should be distinguished from biometric authentication (the latter is a security process that relies on the unique biological characteristics of an individual to verify that he/she is who he/she says he/she is). Remote biometric identification is when the identities of multiple persons are established with the help of biometric identifiers (fingerprints, facial image, iris, vascular patterns, etc.) at a distance, in a public space and in a continuous or ongoing manner by checking them against data stored in a database.

As discussed earlier, without data, there is no AI. The functioning of many AI systems, and the actions and decisions to which they may lead, very much depend on the data set on which the systems have been trained. The necessary measures should therefore be taken to ensure that, where it comes to the data used to train AI systems, the EU's values and rules are respected, specifically in relation to safety and existing legislative rules for the protection of fundamental rights. The following requirements relating to the data set used to train AI systems could be envisaged:

- Requirements aimed at providing reasonable assurances that the subsequent use of the products or services that the AI system enables is safe, in that it meets the standards set in the applicable EU safety rules (existing as well as possible complementary ones). For instance, requirements ensuring that AI systems are trained on data sets that are sufficiently broad and cover all relevant scenarios needed to avoid dangerous situations.

- Requirements to take reasonable measures aimed at ensuring that such subsequent use of AI systems does not lead to outcomes entailing prohibited discrimination. These requirements could entail in particular obligations to use data sets that are sufficiently representative, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected in those data sets;

- Requirements aimed at ensuring that privacy and personal data are adequately protected during the use of AI-enabled products and services. For issues falling within their respective scope, the General Data Protection Regulation and the Law Enforcement Directive regulate these matters.

b) Keeping of records and data

Taking into account elements such as the complexity and opacity of many AI systems and the related difficulties that may exist to effectively verify compliance with and enforce the applicable rules, requirements are called for regarding the keeping of records in relation to the programming of the algorithm, the data used to train high-risk AI systems, and, in certain cases, the keeping of the data themselves. These requirements essentially allow potentially problematic actions or decisions by AI systems to be traced back and verified. This should not only facilitate supervision and enforcement; it may also increase the incentives for the economic operators concerned to take account at an early stage of the need to respect those rules.

To this aim, the regulatory framework could prescribe that the following should be kept:

- accurate records regarding the data set used to train and test the AI systems, including a description of the main characteristics and how the data set was selected;

- in certain justified cases, the data sets themselves;

- documentation on the programming[53] and training methodologies, processes and techniques used to build, test and validate the AI systems, including where relevant in respect of safety and avoiding bias that could lead to prohibited discrimination.

The records, documentation and, where relevant, data sets would need to be retained during a limited, reasonable time period to ensure effective enforcement of the relevant legislation. Measures should be

[53] For instance, documentation on the algorithm including what the model shall optimise for, which weights are designed to certain parameters at the outset etc.

taken to ensure that they are made available upon request, in particular for testing or inspection by competent authorities. Where necessary, arrangements should be made to ensure that confidential information, such as trade secrets, is protected.

c) Information provision

Transparency is required also beyond the record-keeping requirements discussed in point c) above. In order to achieve the objectives pursued – in particular promoting the responsible use of AI, building trust and facilitating redress where needed – it is important that adequate information is provided in a proactive manner about the use of high-risk AI systems.

Accordingly, the following requirements could be considered:

- Ensuring clear information to be provided as to the AI system's capabilities and limitations, in particular the purpose for which the systems are intended, the conditions under which they can be expected to function as intended and the expected level of accuracy in achieving the specified purpose. This information is important especially for deployers of the systems, but it may also be relevant to competent authorities and affected parties.

- Separately, citizens should be clearly informed when they are interacting with an AI system and not a human being. Whilst EU data protection legislation already contain certain rules of this kind[54], additional requirements may be called for to achieve the abovementioned objectives. If so, unnecessary burdens should be avoided. Therefore, no such information needs to be provided, for instance, in situations where it is immediately obvious to citizens that they are interacting with AI systems. It is furthermore important that the information provided is objective, concise and easily understandable. The manner in which the information is to be provided should be tailored to the particular context.

d) Robustness and accuracy

AI systems – and certainly high-risk AI applications – must be technically robust and accurate in order to be trustworthy. That means that such systems need to be developed in a responsible manner and with an ex-ante due and proper consideration of the risks that they may generate. Their development and functioning must be such to ensure that AI systems behave reliably as intended. All reasonable measures should be taken to minimise the risk of harm being caused.

Accordingly, the following elements could be considered:

- Requirements ensuring that the AI systems are robust and accurate, or at least correctly reflect their level of accuracy, during all life cycle phases;

- Requirements ensuring that outcomes are reproducible;

- Requirements ensuring that AI systems can adequately deal with errors or inconsistencies during all life cycle phases.

[54] In particular, pursuant to Art. 13(2)(f) GDPR, controllers must, at the time when the personal data are obtained, provide the data subjects with further information necessary to ensure fair and transparent processing about the existence of automated decision-making and certain additional information.

- Requirements ensuring that AI systems are resilient against both overt attacks and more subtle attempts to manipulate data or algorithms themselves, and that mitigating measures are taken in such cases.

e) Human oversight

Human oversight helps ensuring that an AI system does not undermine human autonomy or cause other adverse effects. The objective of trustworthy, ethical and human-centric AI can only be achieved by ensuring an appropriate involvement by human beings in relation to high-risk AI applications.

Even though the AI applications considered in this White paper for a specific legal regime are all considered high-risk, the appropriate type and degree of human oversight may vary from one case to another. It shall depend in particular on the intended use of the systems and the effects that the use could have for affected citizens and legal entities. It shall also be without prejudice to the legal rights established by the GDPR when the AI system processes personal data. For instance, human oversight could have the following, non-exhaustive, manifestations:

- the output of the AI system does not become effective unless it has been previously reviewed and validated by a human (e.g. the rejection of an application for social security benefits may be taken by a human only);

- the output of the AI system becomes immediately effective, but human intervention is ensured afterwards (e.g. the rejection of an application for a credit card may be processed by an AI system, but human review must be possible afterwards);

- monitoring of the AI system while in operation and the ability to intervene in real time and deactivate (e.g. a stop button or procedure is available in a driverless car when a human determines that car operation is not safe);

- in the design phase, by imposing operational constraints on the AI system (e.g. a driverless car shall stop operating in certain conditions of low visibility when sensors may become less reliable or shall maintain a certain distance in any given condition from the preceding vehicle).

f) Specific requirements for remote biometric identification

The gathering and use of biometric data[55] for remote identification[56] purposes, for instance through deployment of facial recognition in public places, carries specific risks for fundamental rights[57]. The

[55] Biometric data is defined as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique authentification or identification of that natural person, such as facial images or dactyloscopic [fingerprint] data." (Law Enforcement Directive, Art. 3 (13); GDPR, Art. 4 (14); Regulation (EU) 2018/1725, Art. 3 (18).

[56] In connection to facial recognition, identification means that the template of a person's facial image is compared to many other templates stored in a database to find out if his or her image is stored there. Authentication (or verification) on the other hand is often referred to as one-to-one matching. It enables the comparison of two biometric templates, usually assumed to belong to the same individual. Two biometric templates are compared to determine if the person shown on the two images is the same person. Such a procedure is, for example, used at Automated Border Control (ABC) gates used for border checks at airports.

[57] For example on people's dignity. Relatedly, the rights to respect for private life and protection of personal data are at the core of fundamental rights concerns when using facial recognition technology. There is also a potential impact on non-discrimination and rights of special groups, such as children, older persons and persons with disabilities. Moreover, freedom of expression, association and assembly must not be undermined by the use of the technology. See: Facial recognition technology: fundamental rights considerations in the context of law enforcement, https://fra.europa.eu/en/publication/2019/facial-recognition.

fundamental rights implications of using remote biometric identification AI systems can vary considerably depending on the purpose, context and scope of the use.

EU data protection rules prohibit in principle the processing of biometric data for the purpose of uniquely identifying a natural person, except under specific conditions[58]. Specifically, under the GDPR, such processing can only take place on a limited number of grounds, the main one being for reasons of substantial public interest. In that case, the processing must take place on the basis of EU or national law, subject to the requirements of proportionality, respect for the essence of the right to data protection and appropriate safeguards. Under the Law Enforcement Directive, there must be a strict necessity for such processing, in principle an authorisation by EU or national law as well as appropriate safeguards. As any processing of biometric data for the purpose of uniquely identifying a natural person would relate to an exception to a prohibition laid down in EU law, it would be subject to the Charter of Fundamental Rights of the EU.

It follows that, in accordance with the current EU data protection rules and the Charter of Fundamental Rights, AI can only be used for remote biometric identification purposes where such use is duly justified, proportionate and subject to adequate safeguards.

In order to address possible societal concerns relating to the use of AI for such purposes in public places, and to avoid fragmentation in the internal market, the Commission will launch a broad European debate on the specific circumstances, if any, which might justify such use, and on common safeguards.

E. ADDRESSEES

In relation to the addressees of the legal requirements that would apply in relation to the high-risk AI applications referred to above, there are two main issues to be considered.

First, there is the question how obligations are to be distributed among the economic operators involved. Many actors are involved in the lifecycle of an AI system. These include the developer, the deployer (the person who uses an AI-equipped product or service) and potentially others (producer, distributor or importer, service provider, professional or private user).

It is the Commission's view that, in a future regulatory framework, each obligation should be addressed to the actor(s) who is (are) best placed to address any potential risks. For example, while the developers of AI may be best placed to address risks arising from the development phase, their ability to control risks during the use phase may be more limited. In that case, the deployer should be subject to the relevant obligation. This is without prejudice to the question whether, for the purpose of liability to end-users or other parties suffering harm and ensuring effective access to justice, which party should be liable for any damage caused. Under EU product liability law, liability for defective products is attributed to the producer, without prejudice to national laws which may also allow recovery from other parties.

Second, there is the question about the geographic scope of the legislative intervention. In the view of the Commission, it is paramount that the requirements are applicable to all relevant economic operators providing AI-enabled products or services in the EU, regardless of whether they are established in the EU or not. Otherwise, the objectives of the legislative intervention, mentioned earlier, could not fully be achieved.

[58] Article 9 GDPR, Article 10 Law Enforcement Directive. See also Article 10 Regulation (EU) 2018/1725 (applicable to the EU institutions and bodies).

F. COMPLIANCE AND ENFORCEMENT

In order to ensure that AI is trustworthy, secure and in respect of European values and rules, the applicable legal requirements need to be complied with in practice and be effectively enforced both by competent national and European authorities and by affected parties. Competent authorities should be in a position to investigate individual cases, but also to assess the impact on society.

In view of the high risk that certain AI applications pose for citizens and our society (see section A above), the Commission considers at this stage that an objective, prior conformity assessment would be necessary to verify and ensure that certain of the above mentioned mandatory requirements applicable to high-risk applications (see section D above) are complied with. The prior conformity assessment could include procedures for testing, inspection or certification[59]. It could include checks of the algorithms and of the data sets used in the development phase.

The conformity assessments for high-risk AI applications should be part of the conformity assessment mechanisms that already exist for a large number of products being placed on the EU's internal market. Where no such existing mechanisms can be relied on, similar mechanisms may need to be established, drawing on best practice and possible input of stakeholders and European standards organisations. Any such new mechanism should be proportionate and non-discriminatory and use transparent and objective criteria in compliance with international obligations.

When designing and implementing a system relying on prior conformity assessments, particular account should be taken of the following:

- Not all requirements outlined above may be suitable to be verified through a prior conformity assessment. For instance, the requirement about information to be provided generally does not lend itself well for verification through such an assessment.

- Particular account should be taken of the possibility that certain AI systems evolve and learn from experience, which may require repeated assessments over the life-time of the AI systems in question.

- The need to verify the data used for training and the relevant programming and training methodologies, processes and techniques used to build, test and validate AI systems.

- In case the conformity assessment shows that an AI system does not meet the requirements for example relating to the data used to train it, the identified shortcomings will need to be remedied, for instance by re-training the system in the EU in such a way as to ensure that all applicable requirements are met.

The conformity assessments would be mandatory for all economic operators addressed by the requirements, regardless of their place of establishment[60]. In order to limit the burden on SMEs, some support structure might be envisaged including through the Digital Innovation Hubs. In addition, standards as well as dedicated online tools could facilitate compliance.

[59] The system would be based on conformity assessment procedures in the EU, see Decision 768/2008/EC or on Regulation (EU) 2019/881 (Cybersecurity Act), taking into account the specificities of AI. See the Blue Guide on the Implementation of EU product rules, 2014.

[60] As regards the relevant governance structure, including the bodies designated to carry out the conformity assessments, see section H below.

Any prior conformity assessment should be without prejudice to monitoring compliance and ex post enforcement by competent national authorities. That holds true in respect of high-risk AI applications, but also in respect of other AI applications subject to legal requirements, although the high-risk nature of the applications at issue may be reason for the competent national authorities to give particular attention to the former. Ex-post controls should be enabled by adequate documentation of the relevant AI application (see section E above) and, where appropriate, a possibility for third parties such as competent authorities to test such applications. This may be especially important where risks to fundamental rights arise, which are context dependent. Such monitoring of compliance should be part of a continuous market surveillance scheme. Governance-related aspects are further discussed in section H below.

Moreover, both for high- risk AI applications and for other AI applications, effective judicial redress for parties negatively affected by AI systems should be ensured. Issues related to liability are further discussed in the Report on the safety and liability framework accompanying this White Paper.

G. VOLUNTARY LABELLING FOR NO-HIGH RISK AI APPLICATIONS

For AI applications that do not qualify as 'high-risk' (see section C above) and that are therefore not subject to the mandatory requirements discussed above (see sections D, E and F above), an option would be, in addition to applicable legislation, to establish a voluntary labelling scheme.

Under the scheme, interested economic operators that are not covered by the mandatory requirements could decide to make themselves subject, on a voluntary basis, either to those requirements or to a specific set of similar requirements especially established for the purposes of the voluntary scheme. The economic operators concerned would then be awarded a quality label for their AI applications.

The voluntary label would allow the economic operators concerned to signal that their AI-enabled products and services are trustworthy. It would allow users to easily recognise that the products and services in question are in compliance with certain objective and standardised EU-wide benchmarks, going beyond the normally applicable legal obligations. This would help enhance the trust of users in AI systems and promote the overall uptake of the technology.

This option would entail the creation of a new legal instrument that sets out the voluntary labelling framework for developers and/or deployers of AI systems that are not be considered as high-risk. While participation in the labelling scheme would be voluntary, once the developer or the deployer opted to use the label, the requirements would be binding. The combination of *ex ante* and *ex post* enforcement would need to ensure that all requirements are complied with.

H. GOVERNANCE

A European governance structure on AI in the form of a framework for cooperation of national competent authorities is necessary to avoid fragmentation of responsibilities, increase capacity in Member States, and make sure that Europe equips itself progressively with the capacity needed for testing and certification of AI-enabled products and services. In this context, it would be beneficial to support competent national authorities to enable them to fulfil their mandate where AI is used.

A European governance structure could have a variety of tasks, as a forum for a regular exchange of information and best practice, identifying emerging trends, advising on standardisation activity as well as on certification. It should also play a key role in facilitating the implementation of the legal framework, such as through issuing guidance, opinions and expertise. To that effect, it should rely on a network of national authorities, as well as sectorial networks and regulatory authorities, at national and EU level. Moreover, a committee of experts could provide assistance to the Commission.

The governance structure should guarantee maximum stakeholders participation. Stakeholders – consumer organisation and social partners, businesses, researchers, and civil society organisations – should be consulted on the implementation and the further development of the framework.

Given already existing structures such as in finance, pharmaceuticals, aviation, medical devices, consumer protection, data protection, the proposed governance structure should not duplicate existing functions. It should instead establish close links with other EU and national competent authorities in the various sectors to complement existing expertise and help existing authorities in monitoring and the oversight of the activities of economic operators involving AI systems and AI-enabled products and services.

Finally, if this option is pursued, the carrying out of conformity assessments could be entrusted to notified bodies designated by Member States. Testing centres should enable the independent audit and assessment of AI-systems in accordance with the requirements outlined above. Independent assessment will increase trust and ensures objectivity. It could also facilitate the work of relevant competent authorities.

The EU enjoys excellent testing and assessment centres and should develop its capacity also in the area of AI. Economic operators established in third countries wanting to enter the internal market could either make use of designated bodies established in the EU or, subject to mutual recognition agreements with third countries, have recourse to third-country bodies designated to carry out such assessment.

The governance structure relating to AI and the possible conformity assessments at issue here would leave the powers and responsibilities under existing EU law of the relevant competent authorities in specific sectors or on specific issues (finance, pharmaceuticals, aviation, medical devices, consumer protection, data protection, etc.) unaffected.

6. CONCLUSION

AI is a strategic technology that offers many benefits for citizens, companies and society as a whole, provided it is human-centric, ethical, sustainable and respects fundamental rights and values. AI offers important efficiency and productivity gains that can strengthen the competitiveness of European industry and improve the wellbeing of citizens. It can also contribute to finding solutions to some of the most pressing societal challenges, including the fight against climate change and environmental degradation, the challenges linked to sustainability and demographic changes, and the protection of our democracies and, where necessary and proportionate, the fight against crime.

For Europe to seize fully the opportunities that AI offers, it must develop and reinforce the necessary industrial and technological capacities. As set out in the accompanying European strategy for data, this also requires measures that will enable the EU to become a global hub for data.

The European approach for AI aims to promote Europe's innovation capacity in the area of AI while supporting the development and uptake of ethical and trustworthy AI across the EU economy. AI should work for people and be a force for good in society.

With this White Paper and the accompanying Report on the safety and liability framework, the Commission launches a broad consultation of Member States civil society, industry and academics, of concrete proposals for a European approach to AI. These include both policy means to boost investments in research and innovation, enhance the development of skills and support the uptake of AI by SMEs, and proposals for key elements of a future regulatory framework. This consultation will

allow a comprehensive dialogue with all concerned parties that will inform the next steps of the Commission.

The Commission invites for comments on the proposals set out in the White Paper through an open public consultation available at https://ec.europa.eu/info/consultations_en. The consultation is open for comments until 19 May 2020.

It is standard practice for the Commission to publish submissions received in response to a public consultation. However, it is possible to request that submissions, or parts thereof, remain confidential. Should this be the case, please indicate clearly on the front page of your submission that it should not be made public and also send a non-confidential version of your submission to the Commission for publication.

Modelbepalingen voor gemeenten voor het rechtvaardig gebruik van Algoritmische toepassingen

Definities

Algoritmische toepassing: software waarmee op geautomatiseerde wijze voorspellingen worden gedaan, besluiten worden genomen en/of adviezen worden gegeven door gebruik te maken van data-analyse, statistiek en/of zelflerende logica.

Beoogde gebruik: het oplossen van de door de Gemeente voorafgaand aan het gebruik van de Algoritmische toepassing gedefinieerde probleem of problemen.

Besluiten: besluiten van de Gemeente die bestuursrechtelijk, privaatrechtelijk en/of feitelijk van aard zijn en die direct of indirect één of meer burgers van de Gemeente, bezoekers van de Gemeente of bedrijven of andersoortige instellingen die in de Gemeente zijn gevestigd in aanmerkelijke mate treffen.

Overeenkomst: de onderhavige overeenkomst waar deze voorwaarden als bijlage integraal onderdeel vanuit maken.

Procedurele transparantie: het verstrekken van informatie over het doel van de Algoritmische toepassing en het proces dat is gevolgd bij de ontwikkeling van de Algoritmische toepassing en de in dat kader gebruikte data, waaronder in ieder geval moet worden begrepen het geven van inzicht in de gemaakte keuzes en de gehanteerde aannames, de gebruikte methode om risico's te identificeren, de geïdentificeerde risico's en de maatregelen die zijn genomen om de risico's te mitigeren en de partijen die betrokken zijn geweest bij het ontwikkelen van de Algoritmische toepassing en hun rol.

Technische transparantie: het verstrekken van informatie aan de hand waarvan de Gemeente inzicht kan krijgen in de technische werking van de Algoritmische toepassing, waaronder in ieder geval kan worden begrepen het verstrekken van de broncode van de Algoritmische toepassing, de technische specificaties die zijn gebruikt bij de ontwikkeling van de Algoritmische toepassing, de bij de ontwikkeling van de Algoritmische toepassing gebruikte data, informatie over de wijze waarop de bij de ontwikkeling van de Algoritmische toepassing gebruikte data zijn verkregen en bewerkt, informatie over de gehanteerde ontwikkelmethode en het doorlopen ontwikkelproces, motivatie van de keuze voor een bepaald model en bijbehorende parameters en informatie over de prestaties van de Algoritmische toepassing.

Uitlegbaar(heid): Het op individueel niveau kunnen uitleggen waarom een Algoritmische toepassing tot een bepaalde beslissing of uitkomst komt. Tenzij Partijen uitdrukkelijk anders overeenkomen, vormt daar in ieder geval onderdeel van dat duidelijk is wat de belangrijkste factoren zijn op basis waarvan een Algoritmische toepassing tot een bepaalde uitkomst is gekomen en welke wijzigingen in de input moeten worden doorgevoerd om tot een andere uitkomst te komen.

Contractuele voorwaarden

Artikel 1 Toepasselijkheid

1.1. Deze voorwaarden zijn van toepassing indien Opdrachtnemer aan de Gemeente een Algoritmische toepassing levert die door de Gemeente wordt gebruikt bij het nemen van Besluiten, het voorbereiden van Besluiten of in het kader van handhavings- of fraudeonderzoek.

1.2. Deze voorwaarden zijn eveneens van toepassing indien Opdrachtnemer aan de Gemeente een Algoritmische toepassing levert die wordt ingezet voor het nemen van beslissingen of het voorbereiden van beslissingen over medewerkers van de Gemeente.

1.3. Deze voorwaarden zijn van toepassing ongeacht of Opdrachtnemer de Algoritmische toepassing in de vorm van een product, als onderdeel van een dienst of als onderdeel van een ontwikkelovereenkomst aan de Gemeente levert.

Artikel 2 Kwaliteit van de data

2.2. Indien en voor zover de Algoritmische toepassing wordt ontwikkeld op basis van data die door de Gemeente aan Opdrachtnemer worden aangeleverd, zal Opdrachtnemer de redelijkerwijs van hem te verwachten maatregelen nemen om ervoor te zorgen dat de bij het ontwikkelen van de Algoritmische toepassing gebruikte data:

a. worden geanalyseerd, gestructureerd en/of bewerkt overeenkomstig gangbare standaarden, onder meer, maar niet uitsluitend, om ongewenste vooringenomenheid (“bias”) in deze data zoveel als redelijkerwijs mogelijk te voorkomen;
b. worden geanalyseerd, gestructureerd en/of bewerkt op een wijze die in overeenstemming is met wet- en regelgeving.

2.2. Indien en voor zover de Algoritmische toepassing wordt ontwikkeld op basis van data die niet door de Gemeente aan Opdrachtnemer worden aangeleverd, zal Opdrachtnemer ervoor zorgen dat de bij het ontwikkelen van de Algoritmische toepassing gebruikte data:
a. worden verzameld, geanalyseerd, gestructureerd en/of bewerkt overeenkomstig gangbare standaarden, onder meer, maar niet uitsluitend, om ongewenste vooringenomenheid (“bias”) in deze data zoveel als redelijkerwijs mogelijk te voorkomen;
b. worden verzameld, geanalyseerd, gestructureerd en/of bewerkt op een wijze die in overeenstemming is met wet- en regelgeving.

2.3. Indien en voor zover Opdrachtnemer de Algoritmische toepassing voorafgaand aan het sluiten van de Overeenkomst reed in eigen beheer heeft ontwikkeld, staat Opdrachtnemer ervoor in dat de in artikel [2.1] en/of artikel [2.2] beschreven maatregelen reeds zijn genomen.

Artikel 3 Rechten op de data

3.1. Alle rechten met betrekking tot de data die in het kader van de Overeenkomst door de Gemeente aan Opdrachtnemer worden verstrekt, komen toe aan de Gemeente. Opdrachtnemer heeft niet het recht deze data te gebruiken voor andere doeleinden dan het uitvoeren van de Overeenkomst. Opdrachtnemer zal deze data op eerste verzoek van de Gemeente vernietigen en/of tot afgifte van de data aan de Gemeente overgaan.

3.2. Alle rechten met betrekking tot de data die in het kader van de uitvoering van de Overeenkomst worden gecreëerd of verzameld, komen toe aan de Gemeente. Tenzij Partijen anders overeenkomen, heeft Opdrachtnemer niet het recht deze data te gebruiken voor andere doeleinden dan het uitvoeren van de Overeenkomst. Opdrachtnemer zal deze data op eerste verzoek van de Gemeente vernietigen en/of tot afgifte van de data aan de Gemeente overgaan.

3.3. Afgifte van de in artikel [3.2 en 3.3] bedoelde data vindt plaats in een door de Gemeente aan te wijzen gangbaar bestandsformaat. Indien Opdrachtnemer voor het omzetten van de data naar het door de Gemeente gewenste bestandsformaat aanvullende werkzaamheden dient te verrichten, zal de Gemeente Opdrachtnemer hiervoor een redelijke vergoeding betalen. Een geschil over de hoogte van de door Gemeente aan Opdrachtnemer te betalen vergoeding, kan voor Opdrachtnemer nimmer reden zijn om zijn verplichtingen uit deze voorwaarden op te schorten.

Artikel 4 Kwaliteit van de Algoritmische toepassing

4.1. Opdrachtnemer verklaart dat de Algoritmische toepassing is ontwikkeld en functioneert op een wijze die in overeenstemming is met wet- en regelgeving.

4.2. Opdrachtnemer verklaart dat de Algoritmische toepassing nauwkeurig, correct en overeenkomstig gangbare standaarden is ontwikkeld en functioneert.

4.3. Opdrachtnemer verklaart dat de Algoritmische toepassing geschikt is voor het Beoogde gebruik.

Artikel 5 Transparantie over de Algoritmische toepassing

5.1. Opdrachtnemer zal op eerste verzoek van de Gemeente aan de Gemeente Procedurele transparantie verschaffen. De Gemeente heeft het recht om de in dat kader door Opdrachtnemer verstrekte informatie openbaar te maken, bijvoorbeeld door opname van de informatie in een register.

5.2. Opdrachtnemer zal op eerste verzoek van de Gemeente aan de Gemeente Technische transparantie verschaffen om de Gemeente in staat te stellen een audit uit te voeren als bedoeld in artikel [8].
De Gemeente zal dergelijke informatie uitsluitend opvragen en gebruiken indien en voor zover dat noodzakelijk is voor de toepassing van artikel [8]. De Gemeente zal op grond van artikel [5.2] aan haar verstrekte bedrijfsvertrouwelijke informatie geheimhouden en na afloop van een audit als bedoeld

in artikel [8] vernietigen, tenzij een op de Gemeente rustende wettelijke verplichting zich tegen geheim houding of vernietiging verzet of de Gemeente de informatie nodig heeft in het kader van een juridische procedure.

5.3. Bij de toepassing van artikel 5.2 kan Opdrachtnemer ervoor kiezen om de broncode van de Algoritmische toepassing niet af te geven aan de Gemeente, maar aan een onafhankelijke derde die namens de Gemeente de in artikel 8 bedoelde audit zal uitvoeren. Eventuele meerkosten die daaruit voortvloeien komen voor rekening van Opdrachtnemer.

5.4. De Gemeente dient te allen tijde de mogelijkheid te hebben om de werking van de Algoritmische toepassing uit te leggen (Uitlegbaarheid). Opdrachtnemer is verplicht aan het Uitlegbaar maken van de Algoritmische toepassing zijn volledige medewerking te geven en alle daarvoor benodigde informatie aan de Gemeente te verstrekken. De Gemeente heeft het recht om de in dat kader door Opdrachtnemer verstrekte informatie openbaar te maken. Bij strijdigheid tussen artikel [5.4] en artikel [5.2] heeft artikel [5.4] voorrang.

5.5. Gedurende de looptijd van de Overeenkomst vormen de in dit artikel beschreven verplichtingen resultaatsverplichtingen en is de Gemeente, tenzij Partijen anders overeenkomen, voor het nakomen van deze verplichtingen geen aanvullende vergoeding verschuldigd aan Opdrachtnemer. Na de looptijd van de Overeenkomst vormen de in dit artikel beschreven verplichtingen inspanningsverbintenissen en is de Gemeente voor de in dat kader door Opdrachtnemer te verrichtten werkzaamheden een redelijke vergoeding verschuldigd.

Artikel 6 Risicomanagementstrategie bij ontwikkeling van de Algoritmische toepassing

6.1. Bij de ontwikkeling van de Algoritmische toepassing zal Opdrachtnemer een gangbare en up-to-date risicomanagementstrategie hanteren. Bij het toepassen van deze risicomanagementstrategie zal Opdrachtnemer de belangrijkste risico’s identificeren die zich kunnen voordoen bij het gebruik van de Algoritmische toepassing door de Gemeente en maatregelen treffen teneinde de geïdentificeerde risico’s beheersbaar te maken. Bij het identificeren van de risico’s zal Opdrachtnemer indien relevant in ieder geval aandacht besteden aan het risico dat niet wordt voldaan aan één of meer van de verplichtingen genoemd in artikel [2] en artikel [4], risico’s die verband houden met non-discriminatie, de mogelijkheid voor de Gemeente om controle te houden over de Algoritmische toepassing en gegevensbescherming.

6.2. Opdrachtnemer zal de in artikel [6.1] beschreven risicomanagementstrategie op dusdanige wijze uitvoeren en documenteren dat de tijdens de in artikel [8] bedoelde audit kan controleren of Opdrachtnemer aan de in artikel [6.1] beschreven verplichting heeft voldaan.

6.3. Indien en voor zover Opdrachtnemer de Algoritmische toepassing voorafgaand aan het sluiten van de Overeenkomst reeds in eigen beheer heeft ontwikkeld, staat Opdrachtnemer ervoor in dat de in artikel [6.1] en artikel [6.2] beschreven handelingen reeds hebben plaatsgevonden.

Artikel 7 Beheer van de Algoritmische toepassing

7.1. Indien Opdrachtnemer de Algoritmische toepassing in de vorm of als onderdeel van een dienst aan biedt of beheer en onderhoud aan de Algoritmische toepassing verricht, staat Opdrachtnemer ervoor in dat de Algoritmische toepassing en de daarbij behorende documentatie gedurende de looptijd van de Overeenkomst aan de in artikel [4] genoemde voorwaarden blijft voldoen.

7.2. Indien Opdrachtnemer de Algoritmische toepassing in de vorm of als onderdeel van een dienst aanbiedt of beheer en onderhoud aan de Algoritmische toepassing verricht, maakt onderdeel van de door Opdrachtnemer te leveren diensten uit dat Opdrachtnemer gedurende de looptijd van de Overeenkomst continue zal monitoren of de in artikel [6.1] bedoelde risico’s nog actueel zijn en of de in artikel [6.1] bedoelde maatregelen effectief zijn. Indien dat niet het geval blijkt te zijn, zal Opdrachtnemer aanvullende maatregelen treffen.

7.3. Onderdeel van de in artikel [7.2] bedoelde verplichting is dat Opdrachtnemer zal informeren indien nieuwe risico’s bekend worden of de in artikel [6.1] bedoelde maatregelen niet effectief blijken te zijn.

Artikel 8 Audit of andersoortige controle

8.1. Opdrachtnemer is te allen tijde verplicht om mee te werken aan een door of namens de Gemeente

uit de voeren audit of andersoortige controle waarin wordt beoordeeld of Opdrachtnemer de in de Overeenkomst gestelde voorwaarden naleeft. Deze medewerking ziet onder meer op het geven van Technische transparantie, het geven van Procedurele transparantie, het inzicht geven in de uitgevoerde risicomanagementstrategie, het beschikbaar stellen van personeel van Opdrachtnemer voor het houden van interviews en het geven van toegang tot de locaties van Opdrachtnemer.

8.2. De Gemeente zal een rapport op (laten) stellen waarin de conclusies van de audit worden vastgelegd. In het rapport zal de Gemeente vastleggen in hoeverre Opdrachtnemer de verplichtingen uit de Overeenkomst naleeft. Indien de Gemeente vaststelt dat Opdrachtnemer de verplichtingen uit dit artikel niet naleeft, is Opdrachtnemer verplicht om binnen de door de Gemeente in het rapport vastgestelde redelijke termijn de door de Gemeente constateerde gebreken te verhelpen. Herstelt Opdrachtnemer de door de Gemeente geconstateerde gebreken niet binnen de in het rapport vastgelegde hersteltermijn, verkeert Opdrachtnemer van rechtswege in verzuim.

8.3. De Gemeente heeft het recht de conclusies van het in artikel [8.2] bedoelde rapport openbaar te maken. Bij strijdigheid tussen artikel [5.2] en artikel [8.3] heeft artikel [8.3] voorrang.

8.4. De Gemeente heeft het recht om maximaal één keer per kalenderjaar een audit uit te (laten) voeren.

8.5. De kosten van de eventueel door de Gemeente in te schakelen auditor komen voor rekening van de Gemeente. Voor eventuele kosten die Opdrachtnemer zal in het kader van de audit zal maken voor andere werkzaamheden dan het verstrekken van Technische transparantie of Procedurele transparantie, zal de Gemeente Opdrachtnemer een redelijke vergoeding betalen. Een geschil over de hoogte van een dergelijke vergoeding kan voor Opdrachtnemer nooit reden zijn om zijn verplichtingen uit deze voorwaarden op te schorten. Een dergelijke vergoeding hoeft de Gemeente niet te betalen indien uit de audit blijkt dat Opdrachtnemer deze voorwaarden op wezenlijke punten niet naleeft of heeft nageleefd.

Artikel 9 Kosten

Tenzij Partijen anders overeenkomen of in deze voorwaarden anders is bepaald, is de Gemeente voor de verplichtingen genoemd in deze voorwaarden geen aanvullende vergoeding verschuldigd aan Opdrachtnemer.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Algemene Rekenkamer
T.a.v. drs. A.P. Visser, President
Lange Voorhout 8
2500 EA 's-Gravenhage

Turfmarkt 147
's-Gravenhage

www.rijksoverheid.nl
www.facebook.com/minbzk
www.twitter.com/minbzk
www.linkedin.com/company/ministerie-van-bzk

Kenmerk
2020-0000720307

Uw kenmerk
20008002 R

Datum 22 DEC 2020
Betreft conceptrapport 'Aandacht voor algoritmes'

Geachte heer Visser,

Hartelijk dank voor uw rapport 'Aandacht voor algoritmes', als resultaat van het interdepartementaal onderzoek naar algoritmes binnen de rijksoverheid.
Vanuit mijn coördinerende verantwoordelijkheid op het gebied van ICT binnen de Rijksdienst bied ik u deze reactie aan, mede namens mijn collega's.

Uw conclusies worden herkend:

- binnen de rijksoverheid is veel aandacht voor het beperken van de privacy risico's voor de rechten en vrijheden van betrokkenen die een rol spelen bij algoritmes;
- u heeft vastgesteld dat automatische besluitvorming alleen plaatsvindt bij algoritmes die eenvoudige administratieve handelingen uitvoeren, zonder enige impact voor de burger;
- vastgesteld is dat de complexe geïnventariseerde algoritmes niet zelf besluiten nemen, maar de uitvoerende functionarissen nadrukkelijk betrokken zijn bij deze besluiten;
- vastgesteld is dat algoritmes voor een onafhankelijk controleur als de Algemene Rekenkamer geen black box zijn: de AR heeft de algoritmes kunnen bekijken en beoordelen;
- onderzocht is wat die algoritmes nu precies wel en niet doen, uw bevindingen dragen bij aan het demystificeren van algoritmes.

Reactie aanbevelingen
Uw aanbevelingen dragen bij aan verbeterde dienstverlening naar de mensen waar de overheid voor werkt en de daarvoor ingerichte werkprocessen.

1. "Uniformiteit en eenduidigheid van begrippen en kwaliteitseisen."
   Aan een eenduidige gemeenschappelijk taal en concrete kwaliteitseisen voor algoritmes wordt gewerkt via onder meer de kennisbundeling en het gestructureerde overleg van de Nederlandse Digitaliseringstrategie (NDS). Een verkenning is uitgevoerd, in samenspraak met de minister voor Rechtsbescherming en de staatssecretaris van EZK, die onder meer heeft

gekeken naar het voorkomen van fragmentatie, de normering van toezicht en het betrekken van publieke en private kennis. De Tweede Kamer is ten tijde van uw onderzoek geïnformeerd[1] en heeft de resultaten van de verkenning besproken met meerdere bewindspersonen.
In deze actie is het zaak een goede balans te zoeken in meerwaarde van rijksbrede eenduidigheid t.o.v. specifieke invullingen per departement en uitvoeringsorganisatie.

2. "Geef de burger inzicht in de toepassing van het algoritme en geef aan waar zij terechtkunnen als ze vragen hebben."
   De opgestelde richtlijnen voor het gebruik van algoritmes door overheden worden verder aangescherpt en geëvalueerd. Daarnaast wordt een model ontwikkeld voor een impact assessment met betrekking tot algoritmes en mensenrechten. Zowel nationale als Europese wetgeving geeft inzicht in de toegepaste voorspellende of voorschrijvende algoritmes.

   In uw rapport wordt Syri als voorbeeld genoemd. Op pagina 5 van dit rapport wordt aangegeven dat het systeem Syri binnen de overheid (door het UWV en de Belastingdienst) werd gebruikt om fraude op te sporen met algoritmes. Omdat deze passage de indruk wekt van een breed en generiek gebruik van Syri in het reguliere toezicht, wordt er aan gehecht de context en het gebruik van Syri te verduidelijken. Syri is een systeem voor vergelijking van gegevensbestanden van verschillende overheidsorganisaties (zowel centraal als decentraal) op grond van de Wet Suwi, dat is ingezet in een beperkt aantal specifieke samenwerkingsprojecten op het terrein van voorkoming en terugdringing van belasting- en socialezekerheidsfraude, overtredingen van arbeidswetgeving en daarmee samenhangende misstanden. Afgelopen 5 februari heeft de rechter zich uitgesproken over gebruik van Syri waaruit bleek dat de privacy van burgers onvoldoende gewaarborgd was. Daarop heeft de overheid het gebruik van Syri meteen stopgezet.

3. "Leg afspraken omtrent de inzet van algoritmes vast en richt de continue monitoring goed in."
   Afspraken omtrent de inzet en monitoring van algoritmes zijn door de intensieve samenwerking van diverse departementen verder vorm gegeven. Concreet resultaat hiervan zijn onder meer het Strategisch actieplan voor kunstmatige intelligentie, een beleidsbrief publieke waarden en waarborgen tegen data-analyses door de overheid.

4. "Draag zorg voor een vertaling van het toetsingskader naar hanteerbare kwaliteitseisen voor algoritmes."
   Het kabinet werkt met de Algemene Rekenkamer en de Auditdienst Rijk aan een vertaling van het toetsingskader naar hanteerbare kwaliteitseisen voor algoritmes. Bij artificiële intelligentie (AI) algoritmes moet ook de

[1] Kamerstuk TK 35212, nr. 5 d.d. 15 oktober 2020

betrouwbaarheid en kwaliteit van data worden meegenomen, omdat AI-algoritmes gebruik maken van data. Dit borgt het kabinet met een aanpak via een agenda en projectgroep om de input van de verkenning en het rekenkameronderzoek, zowel bij de ontwikkeling als de uitwerking, mee te nemen en een breed gedragen agenda op het terrein van normering en toezicht op algoritmen te realiseren. Een nadere analyse is hierbij nodig naar de mogelijke gevolgen voor administratieve belasting en uitvoerbaarheid bij (decentrale) uitvoerders en de samenhangende benodigde implementatietijd met respect voor bestaande structuren en de autonomie van de departementen en uitvoeringsorganisaties.

5. "Betrek meerdere disciplines al bij ontwikkeling van algoritmes."
   Het betrekken van meerdere disciplines is de norm bij de ontwikkeling van beleid, wetgeving, uitvoerende processen en daaruit volgende algoritmes en het toezicht daarop, conform het integraal afwegingskader. De basis voor de in te zetten instrumenten, algoritmes of anderszins, wordt bij wet bepaald en dient te werken binnen het afgesproken kader. Een ideale mix van disciplines voor dergelijke (ontwikkel)processen blijft echter maatwerk, afhankelijk van beschikbare capaciteit, middelen of tijd.

6. "Draag zorg voor het inzicht hebben en houden in het functioneren van de IT General Controls."
   Voor de werking van reguliere systemen als ook algoritmes is het functioneren van de IT General Controls van belang. Mogelijke additionele rapportagewensen of audit verklaringen dragen bij aan inzicht en controle bij de verantwoordelijke opdrachtgever en eigenaar. Met parallelle acties wordt ook het inzicht vanuit de CIO-kolom versterkt. De dynamiek vraagt echter ook om de balans in ogenschouw te houden tussen controle instrumenten en de organisatorische of administratieve belasting die hierbij gepaard gaat.

Met het door u ontwikkelde toetsingskader heeft u een drietal algoritmes in de praktijk getoetst. De bevindingen zijn veralgemeniseerd en suggereren een uniform beeld over de algoritmes. Ik wil daarbij als kanttekening plaatsen dat daarmee voorbeelden van kwalitatief hoogwaardig gebruik van algoritmes binnen de rijksoverheid, waarbij ook ethische voorwaarden als organiserend principe worden gehanteerd, minder in het rapport tot uiting komen.

Naast het belang van bescherming van de rechten en vrijheden van burgers zal, mede vanuit het oogpunt van toezicht, nader inzicht moeten worden verkregen in de experimenteerruimte voor departementen en uitvoeringsorganisaties op het gebied van de inzet van algoritmen, waarbij zowel de (uitvoerings)praktijk als toezichthoudende organisaties van elkaar kunnen leren.

Zoals u ook aangeeft in uw rapport is de samenwerking geïntensiveerd, mede dankzij de denksessie van 22 september 2020 waarbij de departementen en externe experts deelnamen met de rekenkamer en de ADR.

Ik dank uw rekenkamer voor het onderzoek en de bijdrage om hiermee het begrip van en over algoritmes te vergroten. Daarmee zal verder gewerkt worden aan verbeterde dienstverlening en beleidsuitvoering vanuit de rijksoverheid.

Hoogachtend,

De staatssecretaris van Binnenlandse Zaken en Koninkrijksrelaties,

drs. R.W. Knops

Balancing Security and Mobility

Europese Unie
het fonds voor interne veiligheid

Ministerie van Buitenlandse Zaken

Factsheet

Privacy & Information-supported decision-making

Short-stay visas (KVVs)

AbuDhabiColomboKopenhagenLondenLuandaSanJoséSana'aZagrebAbujaChongqingCotonouLjubljanaLuxemburgSanFranciscoSantiagoDeChileYangonAccraChicago

Information for applicants and sponsors

If you want to visit the Netherlands for a short stay and require a visa, or if you want to sponsor someone who is going to apply for a visa, this information is for you.

In this factsheet the Ministry of Foreign Affairs explains how your and which personal data is processed. This information is in addition to the ministry's general privacy statement. We recommend that you also read that statement carefully, especially the Privacy statement regarding short-stay visa applications.

How the Ministry of Foreign Affairs processes applications

The Ministry of Foreign Affairs is responsible for processing and assessing applications for short-stay visa. Due to the large number of applications it receives, and to ensure that it can carry out its work as objectively and thoroughly as possible, the ministry uses a method based on data analysis. This is called information-supported decision-making.

Information-supported decision-making

Information-supported decision-making means that the ministry uses a range of data to assess a visa application.

For example:

- data that the ministry has received directly from you, or your sponsor, when applying for a visa (for example, the application form and the necessary documents);
- data that the ministry has received from third parties in the migration chain, which can be directly linked to you as an applicant or sponsor (for example, about the handling of a similar visa in the past, or about falsified documents); and
- data that the ministry has based on analyses, and that cannot be directly linked to you or your sponsor; also called profiles. These may give information about your visa application based on similar previous applications and provide us with advice on how extensively we need to assess the application.

The ministry uses all of this information to determine whether a visa application meets the legal requirements. The data used for this purpose is processed by the ministry on the basis of our government task: processing and assessing short-stay visa applications.

How does this work in practice?

Information-supported decision-making works by comparing information given in your visa application with information held by the Ministry of Foreign Affairs. This means that in its system the ministry compares the data that it has received from you (or your sponsor) or that it has gathered itself during the visa process. In addition, the information is compared with information that the ministry has received from various other government organisations. Profiling forms an element of information-supported decision-making. The ministry makes use of profiling in order to have as efficiently as possible access to relevant information. By doing so, the visa decision officers receive supporting advice in view of an objective visa decision-making process.

If there is a match between your information provided in the application and the information held in the ministry's system (amongst others a profile), the match can help determine whether an application can be fast-tracked or whether, for example, an interview or extra document check is needed.

Will this affect my visa application?

The criteria for obtaining a short-stay visa are, and always will be, the same for everyone. This is irrespective of information about compliance with the conditions that applied to visa issued in the past or regarding your sponsor. These criteria have been agreed on by Schengen countries and are laid down in the Visa Code.

The final decision to issue a visa or refuse is solely based on the conditions and refusal grounds as laid down in the Visa Code.

Information-supported decision-making only assists with decisions on applications. An extra interview may be recommended on the basis of the results. It is never a reason for refusing a visa in itself.

The information that the ministry uses ensures that a decision can be taken more quickly and more objectively. Ultimately, it is the consular officer who will decide whether or not to issue a visa. It is not a form of automated decision-making.

What data does the Ministry of Foreign Affairs gather and use?

The Ministry of Foreign Affairs uses the following data for information-supported decision-making:

1. **Start of visa application processing:**
   Application form and the required supporting documents
2. **Processing of visa application:**
   - Application form and the required supporting documents
   - Information on previous visa applications, where relevant (kept for a maximum of 5 years)
   - Where applicable, additional information provided on request by the applicant/sponsor or gathered by the ministry (e.g. information obtained during an interview)
   - Results of analyses based on information of similar visa applications (not directly traceable to applicant/sponsor) by way of profiling
3. **Deciding on visa application:**
   Information mentioned above

When processing the visa application, information about the applicant and/or specified sponsor can also be requested from a third party in the immigration system. These could be:

- Royal Military and Border Police
- Immigration and Naturalisation Service (IND)
- Repatriation and Departure Service
- Social Affairs and Employment Inspectorate

How long will this data be kept?

All data relating to the visa process will be deleted from the ministry's system or anonymised after 5 years.

Do you share my data with other organisations?

The Ministry of Foreign Affairs only shares information with other government organisations and/or authorities when this is necessary to facilitate public tasks in the field of border control, supervision and enforcement and return.

This is done under strict conditions, within the applicable laws and regulations, including those relating to data protection.

Right to object

Under the General Data Protection Regulation (GDPR) you are entitled to object to your personal data being processed by the Ministry of Foreign Affairs. The ministry will consider your objection and stop processing personal data if this proves necessary. You can submit an objection to the address given below.

Any questions?

If you have any questions about information-supported decision-making or would like to view your information, correct it or ask that it be removed from the system, please contact:
Ministry of Foreign Affairs
Consular Affairs and Visa Policy Department (DCV)
Postbus 20061
2500 EB Den Haag

Or send an email to DCV-BAO@minbuza.nl.

See the Ministry of Foreign Affairs privacy statement for more information about exercising your rights and contacting the ministry.

Balancing Security and Mobility

EUROPESE UNIE
Het fonds voor interne veiligheid van de Europese Unie

This is a publication by:
Ministry of Foreign Affairs
Postbus 20061 | 2500 EB Den Haag

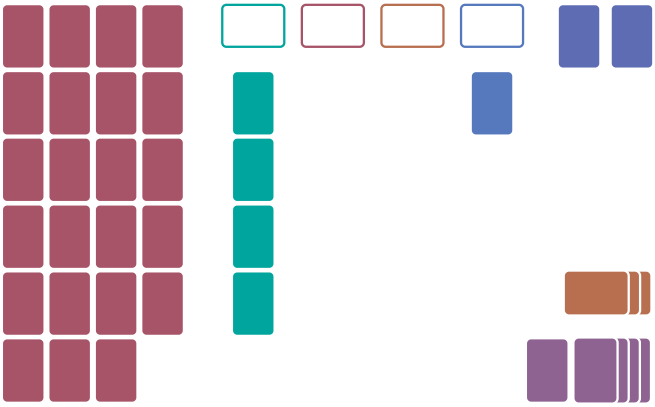
Ethisch Ontwerpspel voor AI

Ethics Inc.

Doel van het spel

De startup ‘Ethics Inc.’ heeft jullie benaderd om een betrouwbare AI-toepassing te ontwikkelen die voldoet aan Europese richtlijnen. Samen vormen jullie een denktank om de startup te adviseren over hoe zij deze toepassing het best kunnen ontwikkelen. Het spel eindigt als er consensus is bereikt na één of meer rondes.

Speloverzicht:



Inhoud van de doos:

1x Spelregelkaart
1x AI-Kaart
4x Voorzitterkaart
23x Waardenkaarten met belangen
16x User storykaarten
5x AI-toepassingskaarten met uitdagingen
33x Wegingskaarten
8x Stakeholderkaarten

Een Uitwisbare stift per deelnemer (niet inbegrepen)

Voorbereiding

Voordat het ontwerpspel kan beginnen moet worden bepaald wie de voorzitter van de denktank is. De voorzitter loodst de groep door de stappen heen en zal de discussies van de denktank begeleiden. Wanneer de voorzitter is aangewezen kunnen de voorbereidende stappen doorlopen worden:

1. Neem de Spelregelkaart en loop het verhaal en de richtlijnen gezamenlijk door.
2. Neem aansluitend de AI-Kaart en lees gezamenlijk de definitie van AI door.
3. Pak vervolgens de AI-toepassingskaarten met uitdagingen erbij en kies samen met de groep één toepassing uit. Dit bepaalt de uitdaging waarover jullie je advies als denktank gaan uitbrengen. Gevorderde spelers kunnen ook zelf een toepassing en uitdaging inbrengen.
4. Pak dan de Stakeholderkaarten erbij. Selecteer gezamenlijk de vier méést relevante stakeholders bij deze uitdaging. Gebruik de ‘jokerkaart’ om zelf waar nodig een stakeholder in te brengen.
5. Neem daarna de Waardenkaarten met belangen. Verspreid deze over een helft van de tafel. Leg alle kaarten dusdanig neer dat de overkoepelende waarden zichtbaar zijn (zoals menselijke autonomie en transparantie) en de gespecificeerde belangen gesloten op tafel liggen. Deze waarden en belangen komen overeen met de vereisten van de ‘Ethics guidelines for trustworthy AI’ van de Europese Commissie uit 2019.
6. Pak vervolgens de User storykaarten. Verdeel de Invulkaarten over de spelers. Leg de overige vier storykaarten op tafel: Als… wil ik… om zo… bij…
7. Deel dan aan iedere speler een set Wegingskaarten uit (S, M, L en XL). Leg de koffiekaart op tafel.

Spelverloop

Nu kan het ontwerpspel echt beginnen! Als denktank moeten jullie samen discussiëren over de ideale AI-toepassing voor ‘Ethics Inc.’ en hierover tot consensus komen. De voorzitter leidt de discussie zo neutraal mogelijk.

Totstandkoming spel

Het ethisch ontwerpspel is ontstaan uit een samenwerking tussen:

Stichting Toekomstbeeld der Techniek (STT)
STT is in 1968 opgericht door het Koninklijk Instituut van Ingenieurs (KIVI). STT voert al meer dan 50 jaar brede toekomstverkenningen uit op het snijvlak van technologie en samenleving die domeinoverstijgend en interdisciplinair zijn.

Stichting Koninklijk Nederlands Normalisatie Instituut (NEN)
NEN verbindt partijen zodat zij komen tot afspraken die worden vastgelegd in (inter)nationale normen, standaarden en richtlijnen. Daarnaast ondersteunt NEN partijen bij de toepassing en het gebruik van normen in de praktijk, door middel van trainingen en praktijkgidsen.

Lectoraat Artificial Intelligence van de Hogeschool Utrecht (HU).
Het lectoraat Artificial Intelligence onderzoekt toepassingen van AI en data-gedreven innovatie. Het is van maatschappelijk belang de ontwikkeling van AI in goede banen te leiden en de grenzen van AI te verkennen en waar nodig vast te leggen.

Ethisch Ontwerpspel voor AI

Ethics Inc.

1. Match de stakeholders en de belangen

Welke belangen hebben de stakeholders bij deze uitdaging?

- Maak voor elke stakeholder verschillende combinaties door een waardenkaart om te draaien en een user story te formuleren: Als…(stakeholder) wil ik…(belang) om zo…(invuloefening) bij…(uitdaging). Schrijf bij de invuloefening de motivatie voor de stakeholder op met behulp van de invulkaarten en leg de kaarten op de juiste plek. Zorg ervoor dat je voor iedere waarde tenminste één combinatie maakt.
- Het is mogelijk dat een belang beter bij een andere stakeholder past. Het kan ook voorkomen dat er voor sommige combinaties geen user stories te formuleren zijn. Leg deze waardenkaarten weg.
- Doe dit totdat je voor elke stakeholder drie user stories hebt geformuleerd (of er 30 minuten voorbij zijn. Stel de klok bijvoorbeeld in op je mobiele telefoon).
- Bepaal of jullie als groep alles gezamenlijk doen of de stakeholders onderling verdelen.

2. Weeg de belangen

Hoe zwaar wegen de belangen?

- Bepaal voor elke user story hoe zwaar het belang en de motivatie wegen door gebruik te maken van de Wegingskaarten (S,M,L,XL). Toe aan een break? Zet dan de koffiekaart in en het hele team neemt een pauze.
- Elke speler legt een Wegingskaart gesloten op de user story. Als iedereen een kaart heeft gelegd worden de kaarten gelijktijdig omgedraaid. Herhaal dit voor alle user stories. De voorzitter houdt de scores bij.
- Discussie: hierbij worden de spelers met de hoogste en laagste weging uitgenodigd om hun keuze te verklaren aan de rest van de groep. Is er al consensus mogelijk? Zo niet, dan volgt een tweede ronde waarin spelers opnieuw kaarten op kunnen leggen. Mogelijk moeten user stories worden aangepast. De voorzitter begeleidt deze stappen: doel is dat er consensus ontstaat en er voor elk belang een weging is waar de groep zich als geheel in kan vinden. Er zijn geen individuele winnaars.

3. Vergelijk de waarden onderling

Hoe zwaar wegen de belangen onderling?

- Bekijk de uiteindelijke scores waar consensus over is:
  - S en M: leg deze belangen even apart. Deze komen in een latere stap weer aan bod.
  - L en XL: deze belangen staan in deze stap centraal.
- Maak combinaties tussen de verschillende belangen: weeg voor elk belang hoe zwaar deze weegt ten opzichte van het belang van een andere waarde. Welk belang weegt dan zwaarder? Maak opnieuw gebruik van de Wegingskaarten. Herhaal dit voor meerdere combinaties.
- Discussie: Tussen welke waarden ontstaan er conflicterende belangen? Tussen welke waarden zijn overeenkomstige belangen? Welke waarden kunnen verenigd worden? Ook hier moet gestreefd worden naar consensus.

4. Construeer de ideale ontwerpprincipes

Hoe ziet de ideale AI-toepassing eruit?

- Leg de Belangenkaarten die overblijven bij elkaar
- Hoe ziet de ideale AI-toepassing er nu uit, gegeven de uitdaging? Wat is er minimaal nodig voor een betrouwbare toepassing? (Minimum Viable Product).
- Pak de belangen met de S- en M-weging er weer bij: welke belangen kunnen geïntegreerd worden in de ideale toepassing?
- Kom tot een gezamenlijk advies voor de startup. Desgewenst kan de ideale AI-toepassing gevisualiseerd worden op een flipover.

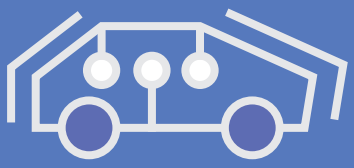
5. Einde van het spel

Welke inzichten kunnen we hieruit halen?

- Ga naar www.ethicsinc.nl en vul het formulier in. Hoe heb je het doorlopen van het proces ervaren? En welke inzichten levert dit op? De inzichten worden gebruikt voor verder onderzoek.

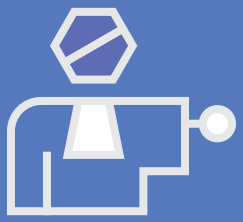
Colofon
Concept en tekst: Rudy van Belkom (STT), Stefan Leijnen (HU) en Roland Bijvank (HU). Grafisch ontwerp: YURR.studio

Zelfrijdende auto



Een autonoom voertuig brengt ons van A naar B, waarbij de bestuurder niet meer nodig is om alles in de gaten te houden. De bestuurder moet echter wel op ieder moment kunnen ingrijpen, wanneer het voertuig daarom vraagt.

Robotrechter

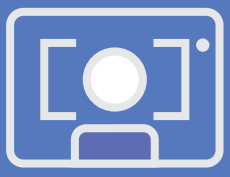


De robotrechter helpt een handje in de rechtspraak door automatisch de zaak van een verdachte te toetsen aan het wetboek en te vergelijken met miljoenen soortgelijke zaken.

DNA-analyse

Het menselijk genoom bestaat uit miljarden basenparen, die met behulp van AI-technieken kunnen worden geanalyseerd. Zo kan voor iedereen een persoonlijke DNA-kaart worden gemaakt.

Slimme deurbel

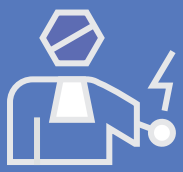


Steeds meer deurbellen zijn uitgerust met een camera. Zo kan met behulp van gezichtsherkenning worden bepaald wie er voor de deur staat en kan de deur automatisch openen.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Onterecht schuldig

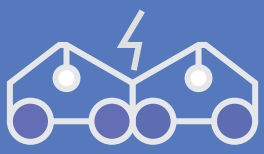
Robotrechter



De robotrechter beoordeelt een verdachte als schuldig en de rechter neemt dit oordeel een-op-een over. Later blijkt echter dat de veroordeling onterecht was. Hoe gaan we ethisch verantwoord met deze situatie om? En wat kunnen we hieruit leren bij de ontwikkeling van betere robot-rechters in de toekomst?

Een botsing

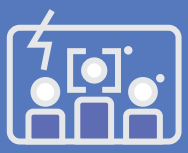
Zelfrijdende auto



Door de zelfrijdende auto zijn er bijna geen files meer en zijn er steeds minder ongelukken. Toch kunnen niet alle ongelukken voorkomen worden. Twee autonome voertuigen botsen op elkaar. Hoe gaan we ethisch verant-woord met deze situatie om? En wat kunnen we hieruit leren bij de ontwikkeling van betere autonome voertuigen in de toekomst?

Iedereen in beeld

Slimme deurbel



Zelfs als je op vakantie bent kun je zien wie er voor de deur staat. De camera van de deurbel staat altijd aan en de fabrikant ontvangt alle beelden. De gezichten van mensen die toevallig op straat voorbij lopen worden echter ook herkend. Hoe gaan we ethisch verantwoord met deze situatie om? En wat kunnen we hieruit leren bij de ontwikkeling van betere gezichtsherkenningssoftware in de toekomst?

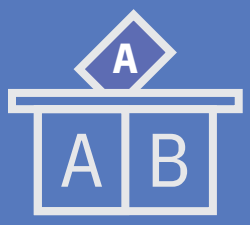
Nooit meer ziek

DNA-analyse

Mensen kunnen laten voorspellen of ze een verhoogd risico hebben op ernstige ziektes, of waar ze een grote kans op hebben om aan te komen overlijden. In sommige gevallen heeft het systeem het echter mis. Hoe gaan we ethisch verantwoord met deze situatie om? En wat kunnen we hieruit leren bij de ontwikkeling van betere DNA-analyses in de toekomst?

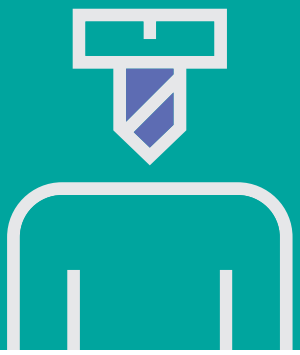
Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Intelligente stemhulp

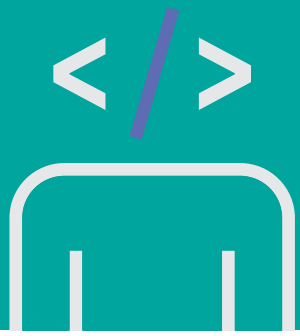


Op basis van een persoonlijk data-profiel kan een intelligente stemhulp met hoge mate van zekerheid bepalen welke politieke partij het beste bij iemand past. Dit vormt de basis voor het stemadvies.

Fabrikant



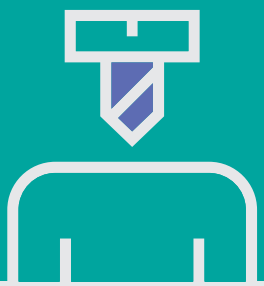
Programmeur



Aanbieder

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Fabrikant



De fabrikant maakt nieuwe producten of diensten en houdt daarbij goed rekening met ethische en juridische kaders. Dat is een lastige klus, omdat niet altijd duidelijk is hoe de producten of diensten zullen worden toegepast door de aanbieder, of in welke context de gebruiker ermee zal werken.

Opnieuw de favoriet

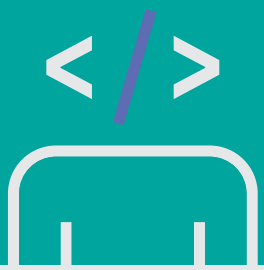
Intelligente stemhulp

Een intelligente stemhulp bespaart burgers veel tijd tijdens verkiezingen. Bij elke verkiezing moet de stemhulp opnieuw worden vastgesteld aan de hand van de verkiezingsprogramma's. Sommige politieke partijen weten echter steeds beter wat ze in hun programma's moeten zetten om als favoriet naar boven te komen. Hoe gaan we ethisch verantwoord met deze situatie om? En wat kunnen we hieruit leren bij de ontwikkeling van betere stemhulpen in de toekomst?

Aanbieder

De aanbieder brengt het AI-product van de fabrikant naar de gebruiker en past deze aan voor de context van de gebruiker. De aanbieder weet vaak beter dan de fabrikant hoe het product zal worden gebruikt, maar wordt tegelijkertijd beperkt door de mogelijkheden die de fabrikant het product meegeeft.

Programmeur

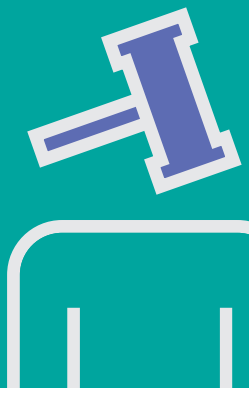


De programmeur is bekend met de mogelijkheden en beperkingen van de AI-software en weet op welke wijze data in het product wordt gebruikt. Dit stelt hem of haar in staat om problemen te signaleren en oplossingen te bedenken die anderen ontgaan.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Gebruiker

Wetgever

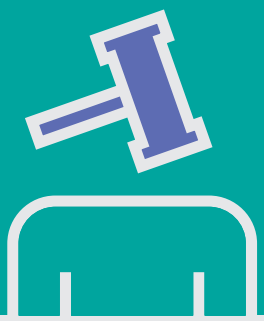


Toezichthouder

Maatschappij

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Wetgever

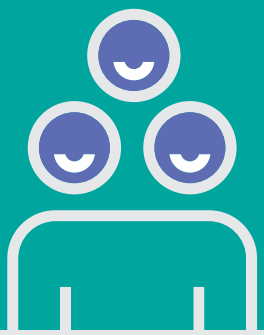


De wetgever kan de toepassing van AI stimuleren en kan door middel van wetten burgers en bedrijven beschermen tegen onwenselijke gevolgen. Daarnaast maakt de overheid als wetgever zelf in toenemende mate gebruik van AI, waarbij ze zich natuurlijk ook aan deze wetten dient te houden.

Gebruiker

De gebruiker is een persoon die interacteert met het product waar de AI in is verwerkt. Vaak gaat dit in ruil voor betaling aan de producent of aanbieder, of is er een andere partij die voor het gebruik betaalt. In sommige gevallen wordt data van de gebruiker verzameld.

Maatschappij



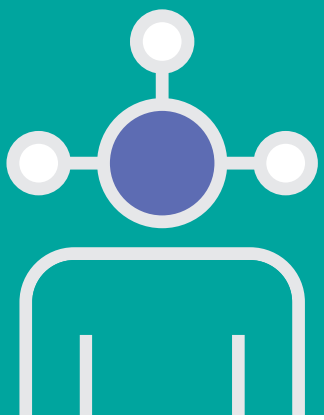
De maatschappij kan direct of indirect gevolgen ondervinden van vrijwel iedere AI-toepassing. Vaak is de indirecte impact niet meteen helder. Dit dwingt alle stakeholders diep na te denken over de maatschappelijke gevolgen voor burgers, samenleving, economie en klimaat.

Toezichthouder

Wanneer er negatieve gevolgen optreden door het gebruik van AI-toepassingen kan de toezichthouder bepalen welke stakeholders aansprakelijk gesteld moeten worden. De toezichthouder kan er daarnaast op toezien (ondersteund door toezichthoudende organen) dat wetten door alle betrokken stake-holders worden nageleefd.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Jokerkaart



Menselijke autonomie

AI-systemen moeten menselijke autonomie en beslissingen onder-steunen. AI-systemen mogen mensen niet onbewust misleiden, manipuleren of dwingen. Bij het ontwerp van AI-toepassingen moeten menselijke keuzes en mogelijkheden centraal staan.

Menselijke autonomie

AI-systemen moeten menselijke autonomie en beslissingen onder-steunen. AI-systemen mogen mensen niet onbewust misleiden, manipuleren of dwingen. Bij het ontwerp van AI-toepassingen moeten menselijke keuzes en mogelijkheden centraal staan.

Menselijke autonomie

AI-systemen moeten menselijke autonomie en beslissingen onder-steunen. AI-systemen mogen mensen niet onbewust misleiden, manipuleren of dwingen. Bij het ontwerp van AI-toepassingen moeten menselijke keuzes en mogelijkheden centraal staan.

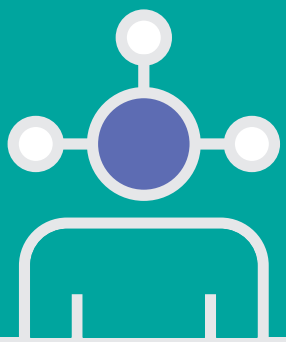
Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Grondrechten

Menselijke autonomie

AI-systemen kunnen menselijke grondrechten zowel bevorderen als belemmeren. Zo kan het bijvoorbeeld ons recht op privacy eerbiedigen (door het beschermen van onze persoonsgegevens), maar ook schenden (door het misbruiken van onze persoonsgegevens). Bij de ontwikkeling van AI-systemen moeten mogelijke risico's beperkt of gerechtvaardigd kunnen worden om zo de rechten en vrijheden van mensen in een democratische samenleving te respecteren.

Jokerkaart



Beschrijf zelf een stakeholder die je mist in het spel.

Menselijk toezicht

Menselijke autonomie

Een AI-systeem mag de menselijke autonomie niet ondermijnen.
Daarom is menselijk toezicht van groot belang. Dit kan worden verwezenlijkt door menselijke interventie te waarborgen in zowel de besluitcyclus van het systeem, als in het ontwerpproces van het systeem. Het is aan de mens om te bepalen wanneer en hoe het systeem in iedere specifieke situatie wordt gebruikt. Dit kan dus ook betekenen dat een keuze van een AI-systeem in een bepaalde situatie niet wordt opgevolgd.

Menselijke controle

Menselijke autonomie

Gebruikers van AI-systemen moeten in staat zijn om het systeem in voldoende mate zelf te kunnen controleren of indien nodig beslissingen van AI-systemen aan te kunnen vechten. Mensen mogen niet onderworpen worden aan systemen die uitsluitend geautomatiseerde besluiten nemen, waarbij gebruikers aanzienlijk getroffen kunnen worden of waarbij het besluit van het systeem rechtsgevolgen kan hebben voor de gebruiker.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Technische robuustheid en veiligheid

AI-systemen moeten in staat zijn om onacceptabele schade te voorkomen en onbedoelde en onverwachte schade zoveel mogelijk te beperken. Hierbij moet de fysieke en mentale gezondheid van de mens te allen tijde gewaarborgd blijven.

Technische robuustheid en veiligheid

AI-systemen moeten in staat zijn om onacceptabele schade te voorkomen en onbedoelde en onverwachte schade zoveel mogelijk te beperken. Hierbij moet de fysieke en mentale gezondheid van de mens te allen tijde gewaarborgd blijven.

Technische robuustheid en veiligheid

AI-systemen moeten in staat zijn om onacceptabele schade te voorkomen en onbedoelde en onverwachte schade zoveel mogelijk te beperken. Hierbij moet de fysieke en mentale gezondheid van de mens te allen tijde gewaarborgd blijven.

Technische robuustheid en veiligheid

AI-systemen moeten in staat zijn om onacceptabele schade te voorkomen en onbedoelde en onverwachte schade zoveel mogelijk te beperken. Hierbij moet de fysieke en mentale gezondheid van de mens te allen tijde gewaarborgd blijven.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Uitwijkplan en algemene veiligheid

Technische robuustheid en veiligheid

AI-systemen moeten in staat zijn om bij problemen een uitwijkplan in werking te stellen. Dit kan betekenen dat het systeem om menselijke tussenkomst vraagt voordat het verder gaat. Er moeten processen worden ingesteld om potentiële risico's in verschillende toepassingsdomeinen te verduidelijken en te beoordelen. Bij hoge risico's moeten er proactief veiligheidsmaatregelen worden ontwikkeld en getest. Onbedoelde gevolgen en fouten moeten worden geminimaliseerd.

Weerbaarheid en beveiliging

Technische robuustheid en veiligheid

AI-systemen moeten weerbaar zijn tegen vijandige aanvallen, zodat ze niet door kwaadwillende individuen of organisaties misbruikt kunnen worden. Er moet daarom rekening worden gehouden met mogelijke onbedoelde toepassingen van AI en potentieel misbruik van een AI-systeem door kwaadwillende actoren. Er moeten stappen worden ondernomen om kwetsbaarheden in het systeem op te sporen en mogelijk misbruik van het systeem te voorkomen en te beperken.

Betrouwbaarheid en reproduceerbaarheid

Technische robuustheid en veiligheid

AI-systemen moeten betrouwbaar zijn en optimaal kunnen werken met verschillende soorten input en in verschillende situaties. Op deze wijze kan het systeem worden gecontroleerd en kan onbedoelde schade voorkomen worden. AI-systemen moeten daarbij reproduceerbaar zijn en moeten hetzelfde gedrag vertonen wanneer een experiment onder gelijke omstandigheden wordt herhaald. Dat maakt het proces van het testen en reproduceren van gedrag eenvoudiger.

Nauwkeurigheid

Technische robuustheid en veiligheid

AI-systemen moeten in staat zijn om op basis van gegevens en modellen correcte voorspellingen en aanbevelingen te doen en beslissingen te nemen. Door een gedegen ontwikkelings- en evaluatieproces kunnen onbedoelde risico's vanwege onjuiste voorspellingen worden beperkt en gecorrigeerd. Het AI-systeem moet hierbij aangeven hoe groot de kans op dergelijke fouten is.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Privacy en datagovernance

AI-systemen moeten persoonsgegevens afschermen en privacyschade voorkomen. Hiervoor is het noodzakelijk om de kwaliteit en integriteit van de gebruikte gegevens te waarborgen.
Om het recht op privacy te beschermen kunnen onder andere toegangsprotocollen worden ingesteld.

Privacy en datagovernance

AI-systemen moeten persoonsgegevens afschermen en privacyschade voorkomen. Hiervoor is het noodzakelijk om de kwaliteit en integriteit van de gebruikte gegevens te waarborgen.
Om het recht op privacy te beschermen kunnen onder andere toegangsprotocollen worden ingesteld.

Privacy en datagovernance

AI-systemen moeten persoonsgegevens afschermen en privacyschade voorkomen. Hiervoor is het noodzakelijk om de kwaliteit en integriteit van de gebruikte gegevens te waarborgen.
Om het recht op privacy te beschermen kunnen onder andere toegangsprotocollen worden ingesteld.

Transparantie

Alle elementen die relevant zijn voor een AI-systeem moeten transparant zijn. Zowel de gegevens en het systeem als de bedrijfsmodellen. Dit betekent dat de processen, capaciteiten, doelen en beslissingen van AI-systemen inzichtelijk moeten zijn voor belanghebbenden.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Kwaliteit en integriteit van gegevens

Privacy en data-governance

De kwaliteit van de gebruikte gegevens-sets is cruciaal voor de prestaties van AI-systemen. Bij de verzameling van gegevens moeten vertekeningen, onnauwkeurigheden, fouten en vergis-singen voorkomen worden vóórdat het AI-systeem met dergelijke gegevenssets getraind wordt. De gebruikte gegevens-sets moeten daarom bij elke stap in het proces worden getest en gedocumenteerd. Er moet hierbij worden getracht zo min mogelijk gevoelige gegevens te gebruiken.

Privacy en gegevens-bescherming

Privacy en data-governance

AI-systemen moeten de privacy en gegevensbescherming van gebruikers gedurende de volledige levenscyclus van het systeem garanderen. Dit omvat niet alleen de informatie die oorspronkelijk door de gebruiker is aangeleverd, maar ook de informatie die in de loop van de interactie met het systeem over de gebruiker is ge-genereerd. Het AI-systeem moet ervoor zorgen dat de verzamelde gegevens niet worden ingezet om gebruikers onwettig of onrechtvaardig te discrimineren.

Traceerbaarheid

Transparantie

De beslissingen van AI-systemen moeten herleidbaar zijn. Dit betekent dat de gegevenssets en de processen waaruit de beslissing van het AI-systeem voortkomt zo goed mogelijk gedocumenteerd moeten worden. Dit geldt zowel voor de verzameling en indeling van de gegevens, als voor de gebruikte algoritmen. Dit maakt de controleerbaarheid van AI-systemen mogelijk, waardoor toekomstige fouten voorkomen kunnen worden.

Toegang tot gegevens

Privacy en data-governance

Om persoonsgegevens te kunnen beschermen moet de toegang tot gegevens beheerd kunnen worden. Hiervoor moeten gegevensprotocollen worden ingesteld. In deze protocollen moet worden beschreven wie onder welke omstandigheden gegevens kan inzien. Alleen gekwalificeerd personeel met de juiste bevoegdheid en noodzaak om gegevens van personen in te zien, mag toegang tot dergelijke gegevens verkrijgen.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Transparantie

Alle elementen die relevant zijn voor een AI-systeem moeten transparant zijn. Zowel de gegevens en het systeem als de bedrijfsmodellen. Dit betekent dat de processen, capaciteiten, doelen en beslissingen van AI-systemen inzichtelijk moeten zijn voor belanghebbenden.

Transparantie

Alle elementen die relevant zijn voor een AI-systeem moeten transparant zijn. Zowel de gegevens en het systeem als de bedrijfsmodellen. Dit betekent dat de processen, capaciteiten, doelen en beslissingen van AI-systemen inzichtelijk moeten zijn voor belanghebbenden.

Diversiteit en rechtvaardigheid

Bij de inzet van AI-systemen moeten individuen en groeperingen vrij zijn van onrechtvaardige discriminatie en stigmatisering. Gedurende de gehele levenscyclus van het AI-systeem moeten inclusie en diversiteit gewaarborgd worden. Alle betrokkenen moeten hierbij gelijk behandeld worden.

Diversiteit en rechtvaardigheid

Bij de inzet van AI-systemen moeten individuen en groeperingen vrij zijn van onrechtvaardige discriminatie en stigmatisering. Gedurende de gehele levenscyclus van het AI-systeem moeten inclusie en diversiteit gewaarborgd worden. Alle betrokkenen moeten hierbij gelijk behandeld worden.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Communicatie

Transparantie

Gebruikers van AI-systemen hebben het recht om te weten dat ze met een AI-systeem te maken hebben. Dit betekent dat AI-systemen zich niet als mensen mogen voordoen en herkenbaar moeten zijn. De capaciteiten en beperkingen van een AI-systeem moeten daarbij tijdig aan gebruikers gecommuniceerd worden. Denk hierbij aan de nauwkeurigheid van het systeem. Gebruikers moeten op elk moment op de hoogte gesteld kunnen worden van de redenen achter de resultaten van het AI-systeem.

Verklaarbaarheid

Transparantie

De beslissingen van een AI-systeem moeten door mensen begrepen kunnen worden. Hiervoor moet een geschikte verklaring van het besluitvormingsproces van het AI-systeem tijdig beschikbaar zijn. Dit geldt zowel voor de technische processen van een AI-systeem, als de daaraan gerelateerde menselijke beslissingen. Deze verklaring moet afgestemd zijn op de mate van deskundigheid van de betrokken belanghebbende (bijv. leek, regelgever of onderzoeker).

Toegankelijkheid en universeel ontwerp

Diversiteit en rechtvaardigheid

Mensen moeten ongeacht hun leeftijd, geslacht, vermogens of eigenschappen gebruik kunnen maken van AI-toepassingen. Het is in het bijzonder van belang dat deze technologie toegankelijk is voor mensen met een beperking. Het ontwerp van AI-systemen moet zijn gericht op het breedst mogelijke scala aan gebruikers, volgens relevante toegankelijkheidsnormen. Daardoor wordt voor alle mensen gelijke toegang tot AI-systemen mogelijk.

Voorkomen van onrechtvaardige vertekening

Diversiteit en rechtvaardigheid

Bij de ontwikkeling en toepassing van AI-systemen moet vertekening worden tegengegaan. Dit kan onder andere door onvolledige gegevenssets niet te gebruiken. Op deze wijze kunnen onbedoelde vooroordelen en discriminatie tegen bepaalde individuen of groeperingen worden voorkomen. Daarnaast kan de diversiteit van meningen worden gewaarborgd door personeel met diverse achtergronden en uit verschillende culturen en disciplines in te zetten bij de ontwikkeling van AI-systemen.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Diversiteit en rechtvaardigheid

Bij de inzet van AI-systemen moeten individuen en groeperingen vrij zijn van onrechtvaardige discriminatie en stigmatisering. Gedurende de gehele levenscyclus van het AI-systeem moeten inclusie en diversiteit gewaarborgd worden. Alle betrokkenen moeten hierbij gelijk behandeld worden.

Maatschappelijk en milieu-welzijn

AI-systemen moeten worden gebruikt in het voordeel van alle mensen, inclusief toekomstige generaties. Daarnaast moet AI de samenleving in bredere zin bedienen, dus ook de natuur en het milieu. Duurzaamheid en ecologische verantwoordelijkheid van AI-systemen moeten worden aangemoedigd.

Maatschappelijk en milieu-welzijn

AI-systemen moeten worden gebruikt in het voordeel van alle mensen, inclusief toekomstige generaties. Daarnaast moet AI de samenleving in bredere zin bedienen, dus ook de natuur en het milieu. Duurzaamheid en ecologische verantwoordelijkheid van AI-systemen moeten worden aangemoedigd.

Maatschappelijk en milieu-welzijn

AI-systemen moeten worden gebruikt in het voordeel van alle mensen, inclusief toekomstige generaties. Daarnaast moet AI de samenleving in bredere zin bedienen, dus ook de natuur en het milieu. Duurzaamheid en ecologische verantwoordelijkheid van AI-systemen moeten worden aangemoedigd.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Duurzame en milieu-vriendelijke AI

Maatschap-pelijk en milieuwelzijn

Het ontwikkelings-, installatie- en gebruiksproces van AI-systemen moeten zo milieuvriendelijk mogelijk gebeuren. Maatregelen om de milieu-vriendelijkheid van de volledige ontwikkelings- en toeleveringsketen van het AI-systeem te waarborgen moeten worden gestimuleerd.
Door kritisch onderzoek te doen naar het gebruik van hulpbronnen en de energieconsumptie tijdens de training kan bijvoorbeeld gekozen worden voor minder schadelijke opties.

Participatie van belanghebbenden

Diversiteit en recht-vaardigheid

Bij de ontwikkeling van AI-systemen moeten belanghebbenden gedurende de gehele levenscyclus van het systeem geraadpleegd worden. Zo kunnen bijvoorbeeld gebruikers ook na de ingebruikname van het systeem om feedback gevraagd worden. Daarbij moeten mechanismen ingesteld worden om de participatie en betrokkenheid van ontwikkelaars te bevorderen bij het volledige ontwikkelingsproces van AI-systemen. Dit draagt bij aan een betrouwbare ontwikkeling van AI.

Samenleving en democratie

Maatschap-pelijk en milieuwelzijn

Bij de inzet van AI-systemen moet rekening worden gehouden met het effect op instellingen, de democratie en de samenleving als geheel.
Het gebruik van AI-systemen moet zorgvuldig worden afgewogen, met name in situaties die verband houden met het democratische proces.
Denk hierbij aan politieke besluit-vorming en verkiezingen. Hierbij moeten de bredere maatschappelijke gevolgen van het gebruik van het AI-systeem worden onderzocht.

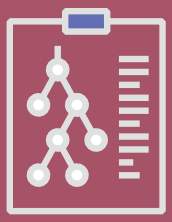
Sociale gevolgen

Maatschap-pelijk en milieuwelzijn

AI-systemen kunnen bijdragen aan de verslechtering van menselijke vaardigheden, zoals arbeidsvaardigheden en sociale vaardigheden.
Wanneer mensen steeds meer interactie hebben met sociale AI-systemen kan dit gevolgen hebben voor onze sociale relaties en mentale gezondheid.
De sociale gevolgen van AI-systemen moeten daarom zorgvuldig worden gemonitord en afgewogen. Denk hierbij ook aan de economische gevolgen, zoals het verlies van banen.

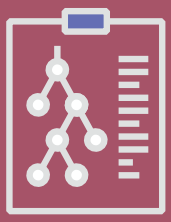
Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Verantwoording



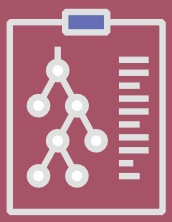
Beslissingen die door AI-systemen genomen worden moeten, afhankelijk van de context en de ernst van de gevolgen, door gebruikers aangevochten kunnen worden. Hiervoor is het noodzakelijk dat de entiteiten die verantwoordelijk zijn voor de beslissing identificeerbaar zijn.

Verantwoording



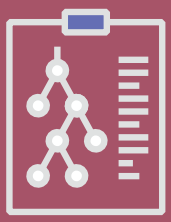
Beslissingen die door AI-systemen genomen worden moeten, afhankelijk van de context en de ernst van de gevolgen, door gebruikers aangevochten kunnen worden. Hiervoor is het noodzakelijk dat de entiteiten die verantwoordelijk zijn voor de beslissing identificeerbaar zijn.

Verantwoording



Beslissingen die door AI-systemen genomen worden moeten, afhankelijk van de context en de ernst van de gevolgen, door gebruikers aangevochten kunnen worden. Hiervoor is het noodzakelijk dat de entiteiten die verantwoordelijk zijn voor de beslissing identificeerbaar zijn.

Verantwoording



Beslissingen die door AI-systemen genomen worden moeten, afhankelijk van de context en de ernst van de gevolgen, door gebruikers aangevochten kunnen worden. Hiervoor is het noodzakelijk dat de entiteiten die verantwoordelijk zijn voor de beslissing identificeerbaar zijn.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

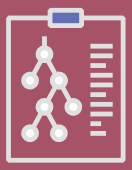
Verslaglegging van negatieve gevolgen

Verantwoording

De vaststelling, beoordeling en verslaglegging van de potentiële negatieve effecten van AI-systemen moeten worden gewaarborgd.
Ngo's, vakverenigingen, klokkenluiders en andere entiteiten moeten worden beschermd wanneer zij gerechtvaardigde zorgen uiten over een AI-systeem.
Het gebruik van effectbeoordelingen tijdens de ontwikkeling,
de installatie en het gebruik van AI-systemen, kan helpen bij het minimaliseren van negatieve gevolgen.

Controleer-baarheid

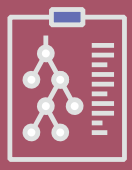
Verantwoording



De algoritmen, gegevens en ontwerp-processen van AI-systemen moeten gecontroleerd kunnen worden. Dat betekent niet dat informatie over bedrijfsmodellen en intellectueel eigendom altijd openbaar beschikbaar moet zijn. Interne en externe controleurs moeten een evaluatieverslag kunnen opstellen die indien nodig openbaar gemaakt kan worden. Bij toe-passingen die van invloed zijn op grond-rechten moetenAI-systemen onafhankelijk kunnen worden gecontroleerd.

Beroep

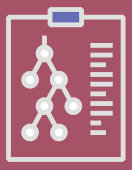
Verantwoording



Wanneer er bij het gebruik van een AI-systeem iets verkeerd gaat en er negatieve gevolgen zijn, dan moeten de betrokkenen in staat zijn om in beroep te gaan tegen de beslissing van een bestuursorgaan. Er moeten hiervoor voldoende mogelijkheden zijn om een bezwaarschrift in te kunnen dienen bij de rechtbank om de beslissing aan te vechten. Er moet hierbij bijzondere aandacht worden besteed aan kwetsbare personen of groepen.

Afwegingen

Verantwoording



Bij de ontwikkeling van AI-systemen kunnen spanningen ontstaan tussen verschillende belangen en waarden, die tot onvermijdelijke afwegingen kunnen leiden. Alle beslissingen over te maken afwegingen moeten worden onderbouwd en goed worden gedocumenteerd. In situaties waarin geen ethisch acceptabele compromissen kunnen worden gevonden, mag de verdere ontwikkeling en het gebruik van het AI-systeem niet in die vorm worden voortgezet.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Onbelangrijk

Onbelangrijk

Onbelangrijk

Onbelangrijk

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Onbelangrijk

Onbelangrijk

Onbelangrijk

Onbelangrijk

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Niet erg belangrijk

Niet erg belangrijk

Niet erg belangrijk

Niet erg belangrijk

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Niet erg belangrijk

Niet erg belangrijk

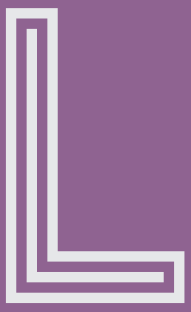
Niet erg belangrijk

Niet erg belangrijk

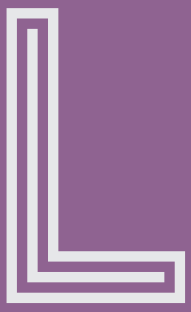
Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Belangrijk



Belangrijk



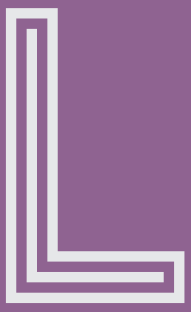
Belangrijk

Belangrijk

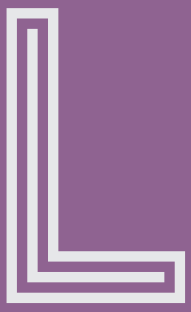
Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Belangrijk



Belangrijk



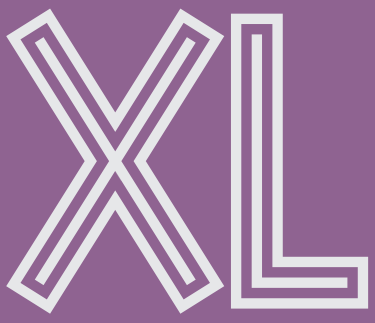
Belangrijk

Belangrijk

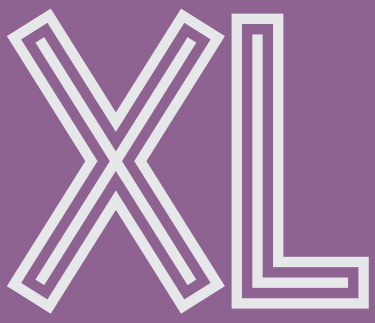
Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

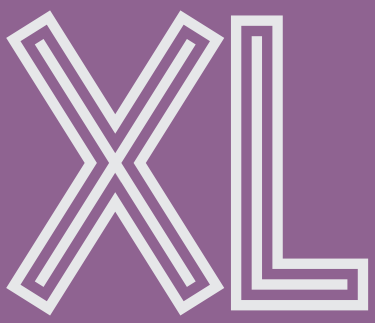
Heel erg belangrijk



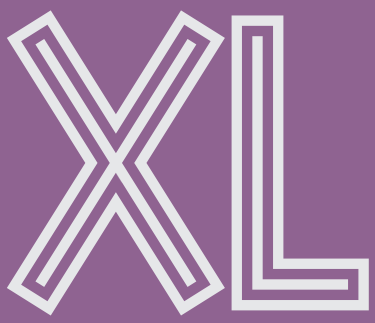
Heel erg belangrijk



Heel erg belangrijk



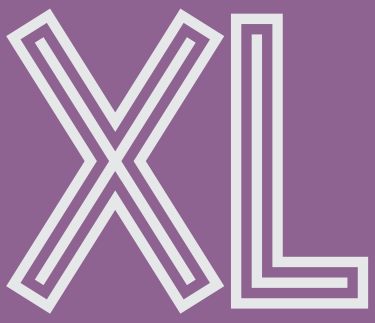
Heel erg belangrijk



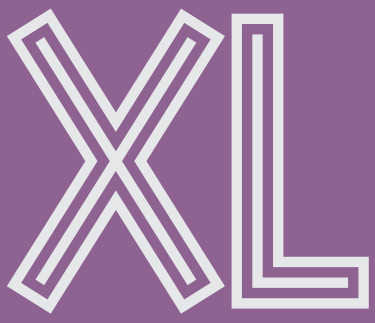
Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

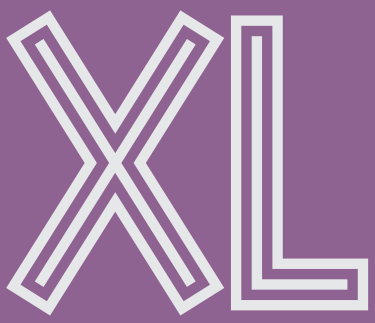
Heel erg belangrijk



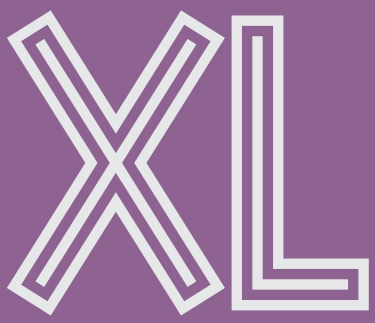
Heel erg belangrijk



Heel erg belangrijk



Heel erg belangrijk



Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Als…

om zo…

wil ik…

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

bij:

Invulkaart

_____

_____

_____

_____

_____

_____

Invulkaart

_____

_____

_____

_____

_____

_____

Invulkaart

_____

_____

_____

_____

_____

_____

_____

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Invulkaart

Invulkaart

Invulkaart

Invulkaart

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Invulkaart

Invulkaart

Invulkaart

Invulkaart

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Invulkaart

_____

_____

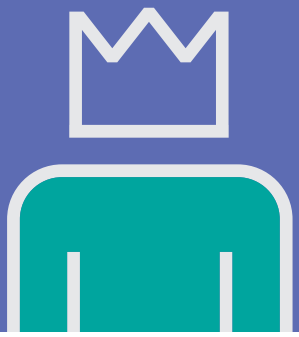
_____

_____

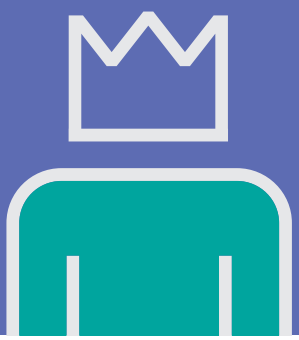
_____

_____

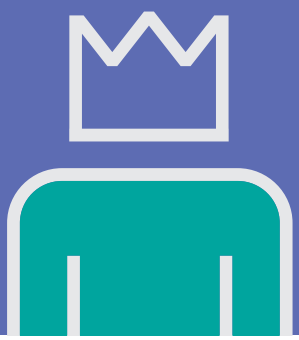
1.Match de stakeholders en de belangen



2.Weeg de belangen



3.Vergelijk de waarden onderling



_____

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Tip voor de voorzitter

Stap 1

Zo zou een user story eruit kunnen zien bij de zelfrijdende auto:

Als fabrikant wil ik de traceer-baarheid vergroten om zo de systeem-fout te kunnen opsporen en verhelpen bij een botsing

Ethics Inc.

4

Tip voor de voorzitter

Stap 3

Mocht de discussie niet van de grond komen dan kun je de discussie aanjagen door dilemma's in te brengen, zoals het trolleyprobleem bij de zelfrijdende auto, een onterechte levenslange gevangenisstraf bij de robotrechter, het 'shoppen' van DNA voor je baby bij de DNA-analyse, het preventief oppakken van criminelen bij de slimme deurbel of misbruik en manipulatie bij de intelligente stemhulp.

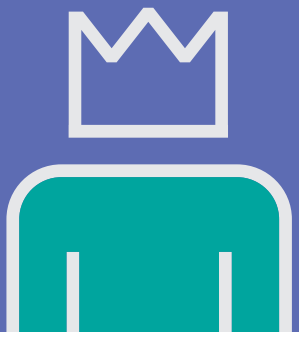
Tip voor de voorzitter

Stap 2

Mocht de discussie niet lopen dan kun je de discussie aanjagen door een 'actualiteit' in te brengen: denk aan een fictief nieuwsbericht waarin een rechter een bepaalde uitspraak heeft gedaan of dat er onethische praktijken aan het licht zijn gekomen bij een organisatie.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

4.Construeer de ideale ontwerp-principes



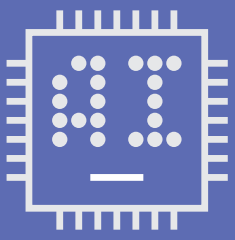
Het verhaal

Ethics Inc.

Spelregels

De startup ‘Ethics Inc.’ heeft jullie benaderd om een betrouwbare AI-toepassing te ontwikkelen die voldoet aan Europese richtlijnen. Samen vormen jullie een denktank om de startup te adviseren over hoe zij deze toepassing het best kunnen ontwikkelen.

Artificiële Intelligentie (AI)



Artificiële Intelligentie (AI) verwijst naar het gedrag van machines dat door observatie en analyse van data leidt tot uitkomsten en acties die passend zijn voor het bereiken van specifieke doelen.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

De richtlijnen

- Je strijdt niet tegen elkaar, maar werkt met elkaar samen
- Iedereen heeft een gelijke stem in de denktank
- Werk toe naar consensus, pas dan kun je verder in het spel
- Iedere mening telt. Ook als je minder ervaring hebt met AI
- Probeer stereotypen over stakeholders te vermijden
- Ook de kleinere belangen tellen mee
- Je kunt een slechte daad niet wegstrepen door een goede daad
- Streef naar het best mogelijke resultaat
- Volg de instructies van de voorzitter op

Tip voor de voorzitter

Stap 4

Zijn er bestaande standaarden waar de AI-toepassing aan moet voldoen? Over ethische afwegingen voor de ideale AI-toepassing zijn vereisten, richtlijnen en best practices opgenomen in internationale standaarden voor AI. Je vindt ze via www.nen.nl/ai. Andersom kunnen nieuwe inzichten weer in de standaarden worden verwerkt. Dat gaat via de norm-commissie Artificial intelligence & big data bij NEN.

Ethics Inc.

Print dit document dubbelzijdig, spiegelen aan de lange zijde.

Actiepunt

Reflectiepunt

De Ethische Data Assistent

DATAGERELATEERDE OVERWEGINGEN

OPSLAG

GEBRUIK

VERZAMELEN

1 ALGORITMEN

Kunt u deze uitleg communiceren met het publiek?

Is er iemand in het team die kan uitleggen hoe het gebruikte algoritme werkt? Is het hoogstwaarschijnlijk dat iemand kan uitleggen wat het doet?

Gebruikt u een algoritme, een vorm van *machine learning* of neurale netwerken in uw project? Zo nee, ga door naar 'Bron'

2 BRON

Waar komen de data vandaan?

Heeft u de kwaliteit van de dataset(s) gecontroleerd?

Hebben de data een houdbaarheidsdatum?

3 ANONIMISEREN

Is het nodig om de dataset(s) te anonimiseren, pseudonimiseren of te generaliseren?

Wie heeft de sleutel om de pseudoniemisering terug te draaien?

4 VISUALISEREN

Hoe worden de uitkomsten van het project weergegeven? Worden deze gevisualiseerd?

Wat zou een andere manier van visualiseren zijn?

5 TOEGANG

Wie heeft toegang tot de dataset(s)?

Hoe wordt de toegang gemonitord?

6 OPEN TOEGANG EN HET HERGEBRUIK VAN GEGEVENS

Zijn delen van de data geschikt om te worden hergebruikt?

Zo ja, onder welke voorwaarden?

Wat zijn de mogelijkheden en valkuilen van het geautomatiseerd of openbaar maken van gegevens?

7 VERANTWOORDELIJKHEDEN

Wie is er verantwoordelijk voor het maken van deze strategieën?

Zijn er communicatiestrategieën voor het geval er iets mis gaat?

Wat is de communicatiestrategie voor dit project?
Zitten alle samenwerkingspartners hierover op één lijn?

8 COMMUNICATIE

Is dit project geschikt voor samenwerking met (commerciële) partners? Zo ja, welke partijen zouden dat kunnen zijn?

Zijn de verantwoordelijkheden van de personen/personen helder?

Wie is/zijn eindverantwoordelijk voor het project?

Welke wetten, voorschriften of richtlijnen zijn van toepassing op uw project?

9 TRANSPARANTIE

Bestaat het risico op publieke verontwaardiging?

Hoe transparant bent u over uw project naar burgers toe?

Kunnen burgers bezwaar maken tegen uitkomsten van het project?

Is een opt-out mogelijk voor burgers?
Zo ja, op welk moment kunnen burgers ervoor kiezen om deelname te beëindigen?

10 PRIVACY

Heeft u de privacyfunctionaris en/of functionaris gegevensbescherming betrokken bij het project?

Heeft u een PIA (Privacy Impact Assessment) of DPIA (Data Protection Impact Assessment) geïnventariseerd?

Zijn er persoonsgegevens gebruikt in het project?
Zo nee, ga verder bij 'Vooringenomenheid'

Krijgt u via de dataset inzicht in de persoonlijke levenssfeer van burgers?

11 VOORINGENOMENHEID

Welke resultaten verwacht u persoonlijk?

Heeft u een onbestemd gevoel over het project? Zo ja, wat zijn uw zorgen? Bespreek dit met uw teamleden

Gebruikt u de uitkomsten van het model als leidend of aanvullend in uw beslissingsmodel?

Bestaat het gevaar dat bepaalde mensen of groepen gediscrimineerd zouden kunnen worden door uw project?

Zijn alle verschillende groepen burgers vertegenwoordigd in de dataset(s)? Wie missen er? Zijn hier zichtbaar?

Zit er een feedback loop in het model die negatieve consequenties kan hebben?

Verzamelt u de goede informatie voor uw doel?

Veranderen uw antwoorden als u de mogelijke langetermijneffecten in acht neemt? Waarom?

Function creep: kunt u zich een toekomstig scenario voorstellen waarin de uitkomsten van dit project voor een ander doeleinde gebruikt worden?

ALGEMENE OVERWEGINGEN

Doeltreffendheid

Autonomie

Inbegrepen

Openheid

Inclusiviteit

Rechtvaardigheid

Legitimiteit

RESULTAAT

STAP DRIE

Welke waarden streeft u persoonlijk en vanuit uw organisatie na? Worden deze waarden vertegenwoordigd in de antwoorden die u zojuist heeft geformuleerd? Bespreek dit. Loop daarna de spiraal verder af en bespreek de waarden die niet aan bod zijn gekomen maar wel relevant zijn voor uw project. Denk hierbij ook vooral verder dan de waarden die hier aangereikt worden.

Zijn er (mogelijke) problemen met uw project?

Wat zijn de voordelen van dit project?

Wie kan hierdoor worden beïnvloed?

Wat voor data gebruikt u?

Wat houdt het project in en wat is het doel?

Deelnemers

Projectnaam, datum, plaats

STAP TWEE

Loop door de vragen heen en beantwoord ze, bijvoorbeeld door het toevoegen van post-its. Maak een actiepunt voor elke vraag die niet direct beantwoord kan worden.

STAP ÉÉN

Wijs één persoon aan om de antwoorden te noteren. Deze poster wordt ondersteund door de DEDA handleiding. De handleiding biedt aanvullende informatie over de vragen. Daarnaast bespreekt het de ethische stromingen waarop DEDA is gebaseerd.

START

Concept:
Mirko Schäfer en Aline Franzke

De 'Data Ethische Beslissingen handleiding' en deze poster zijn ontwikkeld door de Utrecht Data School en de Universiteit Utrecht.

Utrecht University 2019

UTRECHT DATA SCHOOL

Universiteit Utrecht

DEDA- Versie 3.0
augustus 2019
Voor updates klik op:

www.dataschool.nl/dieda

Terwijl DEDA wordt gebruikt leren we om het te verbeteren. Deze verbeteringen zullen in toekomstige versies worden geïmplementeerd. Wanneer u DEDA gebruikt en op- of aanmerkingen heeft, aarzel dan niet om deze met ons te delen. U kunt daarvoor altijd mailen naar:

Info@dataschool.nl

BRIEVEN

FOTO ANS BRYS

KAZERNE

Weten de provincie en de gemeente ook eens hoe het is

Door de provincie Zeeland en gemeente Vlissingen wordt alles uit de kast gehaald om aan te tonen hoe schadelijk en schandelijk het besluit van de regering is om de belofte een kazerne naar Vlissingen te verplaatsen niet na te komen. Enorme schadevergoedingen in geld en goederen zijn het minste waarmee Den Haag over de brug moet komen. Maar in feite krijgen provincie en gemeente een koekje van eigen deeg - nu weten ze hoe dat voelt. Gemeenten en provincies nemen dagelijks beslissingen waarbij eerder gemaakte afspraken niet worden nagekomen. De macht van deze instituties is enorm groot. Dat heeft vooral te maken met de zwakte van het bestuursrecht en met een Raad van State die niet de burger tegen de overheid, maar de overheid tegen de burger beschermt. Dat eerste komt vooral tot uitdrukking doordat de regels op eenvoudige wijze door gemeente en provincie kunnen worden aangepast om te mogen doen wat tot voor kort niet mocht. Zo zijn in Zeeland bungalowparken gebouwd op plekken waar dit eerst niet mocht. En zonder twijfel is hierbij ook gehandeld tegen beloften en afspraken in. Dat maakt burgers - terecht - net zo opstandig als het provinciebestuur van Zeeland. Burgers kunnen echter niet op hoge poten naar Den Haag. De burger loopt in eerste instantie vast bij de gemeente die zonder veel omhaal van woorden alle bezwaren ongegrond verklaart. Datzelfde gebeurt vervolgens bij de Raad van State. Provincie en gemeente moeten niet zeuren. Zo gaat dat nu eenmaal, wen er maar aan.
Harry Luykx Soest

CORONA

Experts tonen juist rust

De voormalig inspecteur-generaal bij het ministerie van Volksgezondheid, Herre Kingma, vindt dat Nederland een landsdokter nodig heeft, blijkt uit zijn ingezonden brief en uit een interview (2/3 en 4/3). Hij ziet vele gezichten, experts van RIVM en GGD, die namens de overheid het woord voeren over de aanpak van het coronavirus. Naar zijn mening kan de communicatie beter. Mij vallen juist de rust en professionaliteit op die RIVM-baas, prof. dr. Jaap van Dissel, tentoonspreidt wanneer hij op radio en tv uitleg geeft over het coronavirus en alles wat daarmee te maken heeft. Eenzelfde rust en professionaliteit gaan uit van de vele arts-microbiologen die in praatprogramma's hun mening geven. In het oog springend is hoe goed zij het beleid dat zij voorstaan op elkaar en met de GGD afstemmen. Iedere medisch-epidemiologische crisis heeft zijn eigen kenmerken en karakter. Om daar voorlichting over te geven hebben we niet zozeer een landsdokter nodig, als wel experts die juist van díé crisis verstand hebben. Dat is precies wat nu gebeurt. Laten we achter onze artsen staan die met grote kundigheid en toewijding, gecoördineerd door het RIVM, met elkaar deze epidemie te lijf gaan.
Theo van Woerkom Den Haag

CORONA (2)

Skiërs in quarantaine

Ik begrijp de halfzachte houding rond de negenhonderd studenten van de Groningse vereniging Vindicat, op vakantie in Noord-Italië, niet (*Wat als die negenhonderd in Italië skiënde studenten straks terugkomen?*, 5/3). Gewoon een maandje op een klein Waddeneiland in quarantaine zetten. Die lui lopen zelf weinig risico, en de rest van het land is dan veilig. Of willen we straks tientallen dode Groningers hebben?
Rob Breebaart Nijmegen

MIGRATIE

EU al jaren chantabel

De vluchtelingendeal tussen de EU en Turkije uit 2016, die premier Mark Rutte als toenmalig EU-voorzitter graag op zijn conto schrijft, breekt ons nu op. Tegen betaling van 6 miljard euro zou de Turkse president Erdogan migranten opvangen en tegenhouden. Maar nu heeft Erdogan de poorten naar Europa opnieuw opengezet om nogmaals geld op te strijken. De EU had in 2016 ervoor moeten kiezen om doeltreffende controles aan de Europese buitengrenzen in te voeren, zoals EU-landen met elkaar afspraken in het Verdrag van Schengen. Maar premier Rutte en bondskanselier Merkel hebben de dictator Erdogan namens de EU geprobeerd af te kopen met miljardensteun en de EU daarmee chantabel gemaakt. Rutte was gewaarschuwd, onder anderen door Geert Wilders.
Paul Schermers Apeldoorn

> Laten we achter onze kundige en toegewijde artsen staan

FLAUBERT

Dat was geen autobus

In het artikel *Hier lonkt een samenleving van onaanraakbaren* (3/3) van Henri Beunders staat deze zin: „Flaubert kon de slaap niet meer vatten als hij de geur van arbeiders in de voorbijrijdende autobus rook." Als vaststaand aannemend dat de eerste auto Benz Patent Motorwagen pas reed in 1885 en dat autobussen hier een afgeleide van zijn, voorts dat Flaubert gestorven is in 1880, vloeit hieruit voort dat die schrijver nooit een *autobus* heeft voorbij zien rijden. Hij heeft wel *omnibussen* met arbeiders kunnen zien (en ruiken), door paardentractie voortbewogen en niet door een verbrandingsmotor.
J. van Seggelen

CORRECTIES/AANVULLINGEN

Illegale migranten

In het stuk *EU en Turkije: hard tegen hard* (3/3, p. 4) staat dat er in Turkije vorig jaar 400.000 illegalen binnenkwamen. Dat was niet juist. Het gaat om mensen die illegaal de grens zijn overgestoken en aangehouden werden. Van illegalen is pas sprake als immigranten een asielprocedure hebben doorlopen en afgewezen zijn.

VN-gezant Libië

In *VN-gezant Libië stopt ermee* (4/3, p. 12) staat ten onrechte dat de vertrekkende Ghassan Salamé een Mauritaanse diplomaat is. Hij heeft de Libanese nationaliteit.

Commentaar

PRIVACY

Schimmige algoritmes bij UWV en SyRI zijn topje van de ijsberg

Bij het UWV krijgen aanvragers van een uitkering een 'risicoscore' toegekend door een computersysteem. Met die risicoscore in de hand nemen medewerkers vervolgens beslissingen over het al dan niet toekennen van een uitkering. Maar op basis van welke criteria en data neemt het systeem eigenlijk de beslissingen die tot de code leiden? Is dat überhaupt wel na te gaan?

Het klinkt als vragen die de uitkeringsinstantie uitvoerig beantwoord zou moeten hebben vóór de invoering van zo'n systeem. Maar daarover twijfelt het UWV blijkbaar. De organisatie maakte donderdag bekend dat het de algoritmes tegen het licht gaat houden om te onderzoeken of die niet discrimineren.

Het onderzoek van het UWV is op zich toe te juichen, maar het toont tegelijk aan dat de invoering van algoritmische beoordelingssystemen door overheden vaak te lichtzinnig gebeurt. De uitkeringsinstantie is namelijk verre van de eerste instelling waarbij grote twijfels ontstaan over de legitimiteit van computerbeslissingen. Dat voedt de vrees dat op veel meer plekken schimmige algoritmes ingrijpende beslissingen over burgers nemen.

Vorige week maakte de Belastingdienst vanwege vergelijkbare bezwaren een einde aan een eigen risicodatabase die het karakter van een zwarte lijst had gekregen. Dat deed het pas nadat de Autoriteit Persoonsgegevens had aangekondigd dat ze een breed onderzoek start naar algoritmische systemen bij de fiscus.

> Als de computer 'nee' zegt, moeten mensen tot de conclusie kunnen komen dat het eigenlijk 'ja' moet zijn

Deze gevallen zijn mede aan het licht gekomen dankzij de recente rechtszaak tegen een ander datasysteem voor fraudebestrijding: SyRI. Dat werd gebruikt door gemeenten om fraudeurs op te sporen. De manier waarop SyRI werd ingezet kon volgens de rechter niet door de beugel omdat het risico op discriminatie en willekeur te groot was.

Alle overheden en bedrijven die algoritmische systemen gebruiken om mensen automatisch in categorieën in te delen, zouden goed naar de SyRI-uitspraak moeten kijken. Het systeem diende volgens de rechtbank weliswaar een legitiem doel: de bestrijding van fraude is „cruciaal" voor het draagvlak voor het stelsel van sociale zekerheid in Nederland. Maar de rechters vonden dat oncontroleerbaar is hoe de algoritmes van SyRI onder de motorkap werken.

De kern van de uitspraak is dat er in dat geval te weinig waarborgen zijn voor burgers: zij kunnen niet uitzoeken of hun gegevens wel juist zijn verwerkt en weten niet waartegen ze precies bezwaar moeten maken.

Deze problemen spelen zeker niet alleen bij fraudedetectie. Een domein waar extra voorzichtigheid moet gelden is *predictive policing*. Dat zijn digitale systemen die voorspellingen doen over criminaliteit en die ook in Nederland al worden gebruikt. In het bedrijfsleven worden algoritmische systemen gebruikt voor uiteenlopende zaken: van het beoordelen van hypotheekaanvragen tot sollicitatiekandidaten.

Voor al die gevallen geldt dat burgers zich moeten kunnen verdedigen tegen beslissingen die in hun nadeel uitvallen. Of het nou gaat om een toeslag, een uitkering, een hypotheek of straks misschien zelfs een baan: als de computer 'nee' zegt, moeten mensen tot de conclusie kunnen komen dat het eigenlijk 'ja' moet zijn.

Als dat niet mogelijk is, omdat de werking van het algoritme onbekend is bijvoorbeeld, dan is dat niets minder dan een aantasting van grondrechten. Op die manier maken algoritmes mensen rechteloos en machteloos.

Algemene Rekenkamer

Lange Voorhout 8
Postbus 20015
2500 EA Den Haag
I [redacted]
T [redacted]
[redacted]
W www.rekenkamer.nl

PERSOONLIJK
Minister van Binnenlandse Zaken en
Koninkrijksrelaties, viceminister-president
Postbus 20011
2500 EA DEN HAAG

DATUM 6 november 2020
BETREFT conceptrapport 'Aandacht voor algoritmes'

Geachte mevrouw Ollongren

De Algemene Rekenkamer heeft onderzoek gedaan naar algoritmes binnen de rijksoverheid. In het kader van bestuurlijk wederhoor ontvangt u hierbij het conceptrapport 'Aandacht voor algoritmes' (zie bijlage).
Gaarne vernemen wij vóór 17 december 2020 uw reactie op dit conceptrapport. Wij merken op dat in een eerdere fase met uw ministerie afstemming heeft plaatsgevonden over de feitelijke informatie in het rapport.
In het bijzonder vragen wij u aan te geven welke maatregelen zullen worden getroffen naar aanleiding van onze conclusies en aanbevelingen.

Wij hebben het conceptrapport tevens voor commentaar toegezonden aan, de staatssecretaris van het ministerie van Binnenlandse Zaken en Koninkrijkrelaties.

Wij zullen de kernpunten uit uw reactie, tezamen met ons nawoord, opnemen in het definitieve rapport. Dit rapport zal naar verwachting op 26 januari 2021 worden gepubliceerd.

Wellicht ten overvloede wijzen wij u erop dat dit conceptrapport vertrouwelijk is en vragen u erop toe te zien dat deze vertrouwelijkheid in acht wordt genomen.

UW KENMERK
ONS KENMERK 20008002 R
BIJLAGEN 1

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief enkel digitaal toe.

Algemene Rekenkamer

drs. A.P. (Arno) Visser,
president

drs. C. (Cornelis) van der Werf,
secretaris

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het
Ministerie van Defensie
Postbus 20701
2500 ES DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte mevrouw Van Craaikamp,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Economische Zaken en Klimaat
Postbus 20401
2500 EK DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte mevrouw Ongering,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het
Ministerie van Financiën
Postbus 20201
2500 EE DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte heer Van den Dungen,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Infrastructuur en Waterstaat
Postbus 20901
2500 EX DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte heer Dronkers,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Justitie en Veiligheid
Postbus 20301
2500 EH DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte heer Schoof,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw drs. [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Landbouw, Natuur en Voedselkwaliteit
Postbus 20401
2500 EK DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte heer Goet,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Onderwijs, Cultuur en Wetenschap
Postbus 16375
2500 BJ DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte mevrouw Hammersma,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Sociale Zaken en Werkgelegenheid
Postbus 90801
25009 LV DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte mevrouw Mulder,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Volksgezondheid, Welzijn en Sport
Postbus 20350
2500 EJ DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte heer Gerritsen,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted].
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het
Ministerie van Buitenlandse Zaken
Postbus 20061
2500 EB DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte heer Huijts,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het
Ministerie van Defensie
Postbus 20701
2500 ES DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte mevrouw Van Craaikamp,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

VERTROUWELIJK
Secretaris-generaal van het Ministerie van
Economische Zaken en Klimaat
Postbus 20401
2500 EK DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte mevrouw Ongering,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Secretaris-generaal van het
Ministerie van Financiën
Postbus 20201
2500 EE DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte heer Van den Dungen,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Infrastructuur en Waterstaat
Postbus 20901
2500 EX DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte heer Dronkers,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het
Ministerie van Justitie en Veiligheid
Postbus 20301
2500 EH DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte heer Schoof,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Plaatsvervangend Secretaris-generaal van het
Ministerie van Landbouw, Natuur en
Voedselkwaliteit
Postbus 20401
2500 EK DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte mevrouw Heijblom,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Onderwijs, Cultuur en Wetenschap
Postbus 16375
2500 BJ DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte mevrouw Hammersma,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Sociale Zaken en Werkgelegenheid
Postbus 90801
2509 LV DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte mevrouw Mulder,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Volksgezondheid, Welzijn en Sport
Postbus 20350
2500 EJ DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte heer Gerritsen,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Algemene Rekenkamer

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het
Ministerie van Algemene Zaken
Postbus 20001
2500EA DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte heer Buitendijk,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted] Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Binnenlandse Zaken en Koninkrijksrelaties
Postbus 20011
2500 EA DEN HAAG

Postbus 20015
2500 EA Den Haag
070-342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

DATUM 13 juli 2021
BETREFT aanbieding projectvoorstel *vervolgonderzoek algoritmes*

Geachte heer Schurink,

In vervolg op onze brief van 16 juni jongstleden met kenmerk 21004025 R/S waarin wij ons onderzoek naar de inzet van algoritmes bij de rijksoverheid hebben aangekondigd, informeer ik u over het feit dat het onderzoek recent is gestart en zullen uitvoeren bij alle ministeries.

Onze contactpersoon op uw ministerie is op de hoogte gesteld van het feit dat wij onderzoek gaan verrichten. We zullen het onderzoek in een gesprek met betrokkenen op uw ministerie nog mondeling aankondigen en verder toelichten. Wij zijn voornemens om dit onderzoek in juni 2022 te publiceren.

Het doel van dit onderzoek is om inzichtelijk te maken in hoeverre een selectie algoritmes binnen de onderdelen van de rijksoverheid verantwoord worden ingezet. Het onderzoek valt hiermee onder onze taak zoals omschreven in artikel 7.16 van de Comptabiliteitswet 2016. Onze bevoegdheden staan vermeld in artikel 7.18 van deze wet. Voor meer informatie verwijs ik u naar de samenvatting van het projectvoorstel, die als bijlage bij deze brief is gevoegd.
Tevens treft u aan de factsheet *De Algemene Rekenkamer komt onderzoek doen: wat kunt u verwachten?*

UW KENMERK
ONS KENMERK 21005030 R/S
BIJLAGEN 2

Met vragen kunt u terecht bij de projectleider van het onderzoek, mevrouw [redacted]. Zij is telefonisch bereikbaar op [redacted] en per e-mail op [redacted]@rekenkamer.nl.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Algemene Zaken
Postbus 20001
2500 EA DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte heer Buitendijk,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted]
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Buitenlandse Zaken
Postbus 20061
2500 EB DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte heer Huijts,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted]
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)

Secretaris van de Algemene Rekenkamer

Algemene Rekenkamer

Postbus 20015
2500 EA Den Haag
070 342 43 44
voorlichting@rekenkamer.nl
www.rekenkamer.nl

PERSOONLIJK
Secretaris-generaal van het Ministerie van
Binnenlandse Zaken en Koninkrijksrelaties
Postbus 20011
2500 EA DEN HAAG

datum 16 juni 2021
betreft aankondiging vervolgonderzoek algoritmes

Geachte Schurink,

Graag informeer ik u over een onderzoek dat wij gaan verrichten naar de inzet van algoritmes bij de rijksoverheid.

Verantwoordelijk (waarnemend) directeur is mevrouw [redacted]
Zij zal het projectvoorstel van het onderzoek na vaststelling doen toekomen aan de verantwoordelijk directeur-generaal binnen uw departement.

Wij zullen het onderzoek op korte termijn aankondigen op onze website.

Vanzelfsprekend begrijpen wij dat de huidige bijzondere omstandigheden invloed hebben op de werkzaamheden van u en uw medewerkers en de prioriteiten die daarbij gesteld moeten worden. In deze context houden wij daar zo goed mogelijk rekening mee.

Het kantoor van de Algemene Rekenkamer is beperkt opengesteld. Er wordt voornamelijk vanaf de thuiswerkplekken gewerkt. Om deze reden sturen wij u deze brief digitaal toe.

Ik dank u bij voorbaat voor uw medewerking.

Voor de Algemene Rekenkamer

drs. C. (Cornelis) van der Werf

uw kenmerk
ons kenmerk 21004025 R/S
bijlage(n)