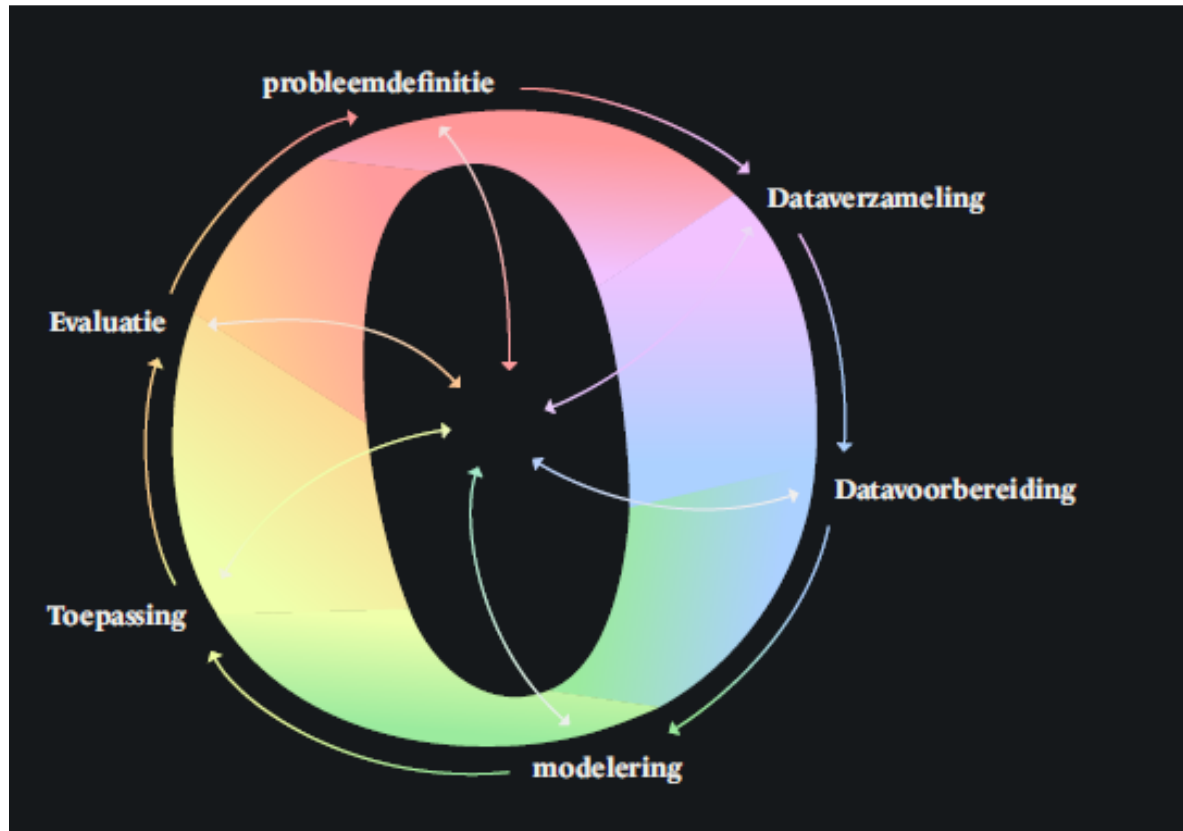


Non-discriminatie by design



Bart van der Sloot, Esther Keymolen, Merel Noorman, Mykola Pechenizkiy, Hilde Weerts, Yvette Wagensveld, Bram Visser en in samenwerking met het College voor de Rechten van de Mens

Inhoud

0. Inleiding.....	4
0.1 Aanleiding.....	4
0.2 Aard en doel van dit onderzoek	6
0.3 Behoeftte aan nieuwe standaard.....	7
0.4 Definitie van AI als gehanteerd voor dit onderzoek	14
0.5 Aanpak en leeswijzer	16
1. Probleemanalyse	19
1.1 Bouwstenen.....	19
1.2 Literatuuranalyse ethiek en STS	27
1.3 Technische aandachtspunten.....	37
1.4 Aandachtspunten binnen de beleidscontext	42
2. Bestaande standaarden en principes	56
2.1 Discriminatie/gelijke behandeling	57
2.2 Richtsnoeren en codes of conduct.....	74
2.3 Statistische standaarden en methodiek.....	96
2.4 Privacy en gegevensbescherming	101
3. Evaluatie bestaande standaarden en principes.....	107
3.1 Overzicht standaarden.....	107
3.2 Testgroepen.....	115
3.3 Aanvullend onderzoek	117
4. Ontwikkeling handreiking	151
5. Validatie standaarden	175
5.1 Workshop 1	175
5.2 Workshop 2	175
5.3 Workshop 3.....	176
6. Finalisering standaarden	177

7. Annex.....	178
7.1 Enquête 1	179
7.2 Enquête 2	196
7.3 Ingevulde enquêtes (enquête 1).....	217
7.4 Ingevulde enquêtes (enquête 2).....	540
7.5 Verslagen van testgroepen	682
7.6 Eerste design Handreiking	707
7.7 Verslagen Workshops	740

0. Inleiding

Dit eerste hoofdstuk geeft een kort overzicht van de aanleiding voor dit onderzoek (0.1), de aard en het doel van dit onderzoek (0.2), de behoefte aan een nieuwe handreiking (0.3), de voor dit onderzoek gehanteerde definitie van Artificial Intelligence (AI) en Machine Learning (ML) (0.4) en de aanpak en opzet van dit onderzoek in de diverse stappen (0.5).

0.1 Aanleiding

Algoritmische besluitvorming raakt steeds meer ingebed in organisatiestructuren; dergelijke besluitvorming kan grote (rechts)gevolgen hebben voor individu en samenleving. De ontwikkeling dat computers worden betrokken bij besluitvorming en het nemen van feitelijke beslissingen is zeker in Nederland niet nieuw. De uitvoering van een uitgebreid socialezekerheidsstelsel is onmogelijk zonder geautomatiseerde besluitvorming¹ en ook administratieve processen in de private sector zijn onwerkbaar zonder algoritmisering. Traditioneel betreffen zulke systemen implementaties van beslisbomen of productieregels (alsdan regels), waarbij beleidsdoelen zijn vertaald naar algoritmen.²

Met de komst van Big Data en de daarmee samenhangende opleving van Artificiële Intelligentie vindt er een doorontwikkeling van methoden en technieken plaats; hierdoor kan worden gewerkt op basis van data in plaats van duidelijk geformuleerde beleids- of beslisregels. De toename van rekenkracht en opslagcapaciteit maakt het mogelijk om grote hoeveelheden data te analyseren en daarin (op nieuwe manieren) patronen te ontdekken; deze patronen kunnen vervolgens worden gebruikt in besluitvormingsprocessen.

Initieel gaat het daarbij om dataverzamelingen van een bestaande praktijk waarbij zowel de invoer als de daarbij behorende beslissing bekend zijn. Door middel van statistische methoden of Machine Learning wordt vervolgens het verband bepaald tussen invoer en uitvoer. In tegenstelling tot de regelgebaseerde systemen, die bijna per definitie betrekkelijk overzichtelijk zijn,³ is het bij veel van de huidige deep-learning systemen niet nodig dat de relatie tussen invoer en uitvoer simpel en eenduidig is. In tegenstelling tot regelgebaseerde systemen waarbij de relatie tussen invoer en beslissing expliciet en extensief is beschreven is dat bij niet-regelgebaseerde (gebaseerd op statistiek of lerend) niet het geval. Dergelijke systemen kunnen ook een uitvoer geven over gevallen waarvan niet op voorhand duidelijk is wat de juiste (alsdan) beslissing is.

Dit kan het landschap van besluitvorming wezenlijk veranderen. Niet alleen is het door de aard van de technologie niet altijd duidelijk (te maken) hoe een beslissing tot stand komt of hoe een beslissing gerechtvaardigd of gemotiveerd kan worden, maar ook is het niet langer met zekerheid te zeggen of de beslissing is zoals beoogd door de ontwerper van het systeem.⁴ Deze problemen worden versterkt wanneer systemen niet alleen in de ontwerpfase leren op basis van

¹ De Vereniging van Administratief Recht hield al in 1993 een baanbrekende jaarvergadering onder de titel 'Beschikken en automatiseren' waarin dit thema centraal stond.

² E. Bayamlioglu & R. Leenes (2018). The 'rule of law' implications of data-driven decision-making: A technology-regulatory perspective. *Law, Innovation and Technology*, 10(2), 295-313.

³ Aangetekend moet worden dat ook regelsystemen de totale set regels erg (cognitief) complex kan zijn en dat mensen moeite hebben die regels 'correct' toe te passen.

⁴ M. Hildebrandt, 'Algorithmic Regulation and the Rule of Law' (2018) 376 *Philosophical Transactions of the Royal Society A*, doi:10.1098/rsta.2017.0355

een training-set, maar ook de algoritmen zich aanpassen op basis van nieuwe kennis en correcties van buiten. De algoritmen lopen op die manier een risico meer een ‘eigen leven’ te gaan leiden.⁵

De thematiek van transparantie (en accountability) van geautomatiseerde besluitvorming staat reeds lang op de agenda,⁶ met name van beslissingen in de publieke sector (besluiten in de zin van de Algemene wet bestuursrecht (Awb)).⁷ Meer recent is de aandacht voor privaatrechtelijke beslissingen. Met name door de opkomst van Data Science en de invoering van de Algemene Verordening Gegevensbescherming (AVG)⁸ staat uitleg van geautomatiseerde beslissingen volop in de aandacht.⁹ De vraag die hier centraal staat is wat precies de reikwijdte is van artikelen 12-15 en 22 en overweging 71 AVG. Daarnaast bestaan er tal van bepalingen in andere wetgeving met betrekking tot beslissingen/besluiten en de uitleg daarvan, bijvoorbeeld afdeling 3.3 Awb.

Maar de effecten van geautomatiseerde besluitvorming gaan veel verder. In wezen gaat het bij dergelijke systemen om techno-regulering,¹⁰ het inbouwen van normen in systemen die vervolgens zelfsturend en -handhavend zijn. Dat levert weliswaar systemen op waarin normhandhaving niet langer afhankelijk is van de capaciteit van traditionele (menselijke) handhaving of toezicht, maar roept daarmee ook allerlei legaliteits- en legitimiteitsvragen op.¹¹

Alhoewel vraagstukken rond legaliteit en legitimiteit met name betrekking hebben op de publieke sector rijzen er ook in de private sector tal van vragen rond de gebruikmaking van algoritmische besluitvorming. Een voorbeeld zijn de regels die geïmplementeerd zullen moeten worden in autonome voertuigen, waarbij deze zich aan de vigerende verkeersregels dienen te houden en het voor de eigenaar niet duidelijk is hoe het voertuig zich zal gedragen in bepaalde

⁵ R. Leenes, E. Palmerini, B.-J. Koops, A. Bertolini, P. Salvini & F. Lucivero (2017). Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues, *Law, Innovation and Technology*. R. Leenes & F. Lucivero (2014). *Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design*, in *Law, Innovation, and Technology*.

⁶ H. Franken et al. (1993). *Beschikken en automatiseren* (Var-reeks, 110). Alphen aan den Rijn: Samsom H.D. Tjeenk Willink. M. M. Groothuis (2005). *Beschikken en digitaliseren. Over normering van de elektronische overheid*. Den Haag: SDU Uitgevers.

⁷ Wet van 4 juni 1992, houdende algemene regels van bestuursrecht (Algemene wet bestuursrecht).

⁸ Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (algemene verordening gegevensbescherming).

⁹ S. Wachter, B. Mittelstadt & L. Floridi (2016). "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", *International Data Privacy Law*. 7 (2) 76-99. I. Mendoza & L.A. Bygrave (2017). *The Right Not to Be Subject to Automated Decisions Based on Profiling*. T. Synodinou, P. Jougoux, C. Markou & T. Prastitou (eds.) (2017). *EU Internet Law: Regulation and Enforcement*, University of Oslo Faculty of Law Research Paper No. 2017-20.

¹⁰ M. Hildebrandt, 'A Vision of Ambient Law' in R. Brownsword and K. Yeung (eds.), *Regulating Technologies* (Hart, 2008) 175-92; H., Mireille (2008) *Legal and Technological Normativity: More (and Less) than Twin Sisters*. *TECHNE*, 12, 3, 169-83. B.-J. Koops (2008) *Criteria for Normative Technology – The Acceptability of 'Code as Law' in Light of Democratic and Constitutional Values*. in: R. Brownsword & K. Yeung (eds.), *Regulating Technologies – Legal Futures, Regulatory Frames and Technological Fixes*. Oxford: Hart, 157-74. R. Brownsword (2005). *Code, control, and choice: why East is East and West is West*. *Legal studies*, 25, 1, 1-21. R. Leenes (1999), 'Hercules of Carneades, Hard cases in recht en rechtsinformatica', Enschede Twente University Press.

¹¹ R. Leenes (2011). 'Framing techno-regulation: an exploration of state and non-state regulation by technology' *5 Legisprudence* 147

verkeerssituaties. Daarmee worden de geïmplementeerde normen voor de weggebruikers zelf intransparant. Anders ligt dat voor ‘object collision avoidance’ normen, omdat een duidelijk juridisch en normatief kader ontbreekt om op basis van regels de koers te bepalen van een voertuig.¹² Wie bepaalt deze regels, welke inspraak is daarbij en welk toezicht wordt daarop uitgeoefend?

Grootschalige data-analyse, personalisatie en geautomatiseerde besluitvorming kunnen ook op systeemniveau consequenties hebben. Niet alleen kan dit een verschuiving van verklaring via voorspelling naar voorkoming (pre-emption) betekenen,¹³ waarbij keuze/beslissingsopties op voorhand verdwijnen uit het blikveld van het individu, ook de grondslagen van huidige juridische stelsels kunnen onder druk komen te staan. In de verzekeringswereld is er bijvoorbeeld een inherente druk om nieuw verkregen kennisinzichten te gebruiken om risico’s te verkleinen. Op basis van inzicht in gezondheidsrisicofactoren, gekoppeld aan gedetailleerd inzicht op individueel niveau kan een verzekeringsmaatschappij in de verleiding komen bepaalde potentiële verzekerden te weren of andere tarieven te rekenen.¹⁴ Solidariteit als grondslag voor risicoverzekering komt daarmee onder druk. Niet weten wie, wanneer, welk risico loopt maakt het eenvoudiger om het gezamenlijk dragen van risico’s als basis te nemen voor het verzekeringsstelsel wanneer individuele risico’s nauwkeurig in kaart kunnen worden gebracht.

0.2 Aard en doel van dit onderzoek

De achtergrond van dit onderzoek is dat AI en ML op verschillende manieren tot discriminatie kunnen leiden, maar ook zouden kunnen worden gebruikt om bestaande discriminatie te bestrijden of te verminderen. In de beleidsbrief “AI, publieke waarden en mensenrechten” van 8 oktober 2019¹⁵ heeft het ministerie van Binnenlandse Zaken en Koninkrijksrelaties (BZK) toegezegd te willen onderzoeken hoe publieke waarden en mensenrechten geoperationaliseerd kunnen worden tot AI-systeemprincipes, te beginnen met non-discriminatie. Dit rapport volgt uit die toezegging en onderzoekt of en zo ja, hoe AI-systeemprincipes kunnen worden ontwikkeld om het discriminatieverbod, zoals onder meer volgt uit artikel 1 Grondwet,¹⁶ de Algemene wet gelijke behandeling (Awgb),¹⁷ artikel 14 Europees Verdrag voor de Rechten van de Mens (EVRM)¹⁸ en het Twaalfde Protocol bij het EVRM,¹⁹ artikelen 21-23 van het EU-Handvest²⁰ en de diverse richtlijnen van de EU²¹ op dit punt, te borgen in een samenleving waarin gebruik van AI steeds meer wijdverbreid zal raken.

¹² Dit roept Trolley-problem-achtige vragen op. Het voertuig detecteert bijvoorbeeld 3 onverwachte objecten op mogelijke koersen. Een van de drie kan niet worden ontweken. Welke koers te kiezen?

¹³ I. Kerr & J. Earle, ‘Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy’ (2013) 66 Stanford Law Review Online 65.

¹⁴ T. Timmer, I. Elias, L. Kool & R. van Est (2015). Berekende risico’s. Verzekeren in de datagedreven samenleving, Den Haag, Rathenau Instituut.

¹⁵ Kamerstukken II 26 643 en 32 761, nr. 642.

¹⁶ Grondwet voor het Koninkrijk der Nederlanden van 24 augustus 1815

¹⁷ Wet van 2 maart 1994, houdende algemene regels ter bescherming tegen discriminatie op grond van godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat.

¹⁸ Verdrag tot Bescherming van de Rechten van de Mens en de Fundamentele Vrijheden Rome, 4.XI.1950.

¹⁹ Protocol No. 12 to the Convention for the Protection of Human Rights and Fundamental Freedoms. ETS No.177. Rome, 04/11/2000

²⁰ Handvest van de grondrechten van de Europese Unie 2012/C 326/02.

²¹ Richtlijn 2000/78/EG van de Raad van 27 november 2000 tot instelling van een algemeen kader voor gelijke behandeling in arbeid en beroep, Publicatieblad Nr. L 303 van 02/12/2000 blz. 0016 – 0022.

Uitgangspunt is daarbij de gedachte dat rondom anti-discriminatie en gelijke behandeling de afgelopen decennia gedetailleerde wetgeving met ge- en verboden tot stand is gekomen en een uitgebreide jurisprudentie is opgebouwd. Toch worden deze principes niet altijd meegenomen door ontwerpers en ontwikkelaars van AI-systemen (of bijvoorbeeld opdrachtgevers bij de overheid of het bedrijfsleven, zoals beleidsmedewerkers en medewerkers van innovatieafdelingen), waardoor het voorkomen van een discriminatoire uitwerking van AI, bijvoorbeeld als gevolg van bias in de data, in de interpretatie van de data, in de doelstelling van het betreffende beleid, of in het algoritme, onvoldoende in het ontwerp van AI-systemen wordt geadresseerd. Het begrip van normen en de vertaling daarvan in ontwerpkeuzes is echter van belang voor het voorkomen van discriminatie als gevolg van het gebruik van AI en hiermee voor het vertrouwen van de samenleving in het gebruik van AI door overheid, instellingen en bedrijven.

Voor een passende borging van anti-discriminatie in AI-systemen is het noodzakelijk om een vertaling te maken van het juridisch kader naar concreet toepasbare systeemprincipes, oftewel van normen naar concrete ontwerpstrategieën die kunnen worden gebruikt bij het ontwerp van AI in een concreet geval door de overheid, door instellingen of door een private organisatie (bij beleid, wetgeving of bij het ontwerpen van toepassingen waarin AI een rol speelt). Dergelijke systeemprincipes kunnen ontwerpers helpen om non-discriminatie- en gelijkebehandelingswetgeving zo goed mogelijk te borgen in het ontwerp van AI-systemen (non-discrimination-by-design). De non-discrimination-by-design principes die in dit onderzoek worden ontwikkeld zijn in hun doelstelling vergelijkbaar met het data-protection-by-design-principe uit de AVG. Non-discrimination-by-design principes zijn uitgangspunten die ontwerpers en ontwikkelaars bij het ontwerp van een AI-systeem kunnen gebruiken om discriminerende resultaten als gevolg van het gebruik van AI zoveel mogelijk te voorkomen. Diverse technologische en organisatorische maatregelen zijn daarbij van belang, zoals gebruikmaking van representatieve data, hantering van wetenschappelijk verantwoorde statistische methoden, etc. Het is, mede gelet op het veranderlijke karakter van AI ná de ontwerp- en ontwikkelfase, van belang dat de non-discrimination-by-design principes ook zoveel mogelijk hun effect kunnen hebben ná deze fasen (bijvoorbeeld in de gebruiksfase). Tools die daarbij van belang kunnen zijn, zijn onder meer logs en audit-systemen.

Een voorbeeld ten aanzien van de ontwikkeling van data-protection-by-design principes is het Kleine Blauwe Boekje van Jaap-Henk Hoepman, zelf ‘techneut’, die de juridische beginselen helder en simpel heeft vertaald naar de beleavingswereld en praktijk van data-analisten en systeemontwikkelaars.²² Het Kleine Blauwe Boekje geniet aanzien en bekendheid onder de doelgroep.

0.3 Behoeft aan nieuwe standaard

Richtlijn 2006/54/EG van het Europees Parlement en de Raad van 5 juli 2006 betreffende de toepassing van het beginsel van gelijke kansen en gelijke behandeling van mannen en vrouwen in arbeid en beroep (herschikking). Richtlijn 2004/113/EG van de Raad van 13 december 2004 houdende toepassing van het beginsel van gelijke behandeling van mannen en vrouwen bij de toegang tot en het aanbod van goederen en diensten. Zie ook: Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation {SEC(2008) 2180} {SEC(2008) 2181}.

²² <<https://www.cs.ru.nl/~jhh/publications/pds-booklet.pdf>>.

Zowel op nationaal als internationaal vlak is de consensus dat bestaande tekortkomingen in de ontwikkeling van AI en ML nu en in de toekomst kunnen leiden tot sociale stratificatie en maatschappelijke ongelijkheid en tot schrijnende situaties op individueel niveau. De urgentie om dit probleem te adresseren wordt onderstreept door de keur aan rapporten en werkgroepen van onder meer de Verenigde Naties, de OECD, de Raad van Europa en diverse instanties in Nederland. Deze zullen uitgebreid aan bod komen in hoofdstuk 2 van dit rapport. Uitgangspunt voor dit rapport is de constatering dat deze richtsnoeren nog onvoldoende effect hebben op de praktijk. Dat kan een aantal oorzaken hebben, zoals, maar niet beperkt tot:

- De documenten zijn vaak opgesteld door juristen en beleidsmakers, die een taal hanteren die niet altijd toegankelijk is voor technici en data-analisten die de systemen ontwerpen.
- Er mist een vertaalslag van de algemene juridische beginselen naar wat dat betekent voor de concrete keuzes ten aanzien van systeemontwerpen.
- Er spelen twee werelden een rol die nog niet genoeg op elkaar aansluiten, namelijk de juridische beginselen omtrent non-discriminatie en gelijke behandeling en de statistische en methodologische standaarden omtrent *fair* en *non-biased data processing*.
- De juridische principes ten aanzien van non-discriminatie zien met name op de uitkomst van beslissingen of de gronden waarop die beslissing wordt genomen, terwijl de oplossing voor discriminatie door toedoen van AI en ML-systemen met name ligt in een betere inrichting van de voorfasen van het besluit.
- Het ontbreekt vaak aan een integrale aanpak waarbij zowel de kaders als ook de organisatie en specifieke domeincontext in samenhang worden behandeld.

Om te verifiëren of er inderdaad een lacune is tussen de juridische beginselen en beleidsdocumenten enerzijds en de praktijk van systeemontwerpers anderzijds en zo ja, wat de oorzaak daarvan zou kunnen zijn, is op dit punt in de twee enquêtes die zijn verstuurd in het kader van dit onderzoek (zie voor een nadere beschrijving paragraaf 0.5) een vraag opgenomen. Na de opzet van dit onderzoek te hebben uitgelegd werd de volgende vraag voorgelegd: ‘Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI-systemen uitgewerkt. De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analisten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI-systemen gebruiken voor beleids- en besluitvorming. Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk).’

Daarbij werden de volgende antwoordcategorieën gegeven:

Ja Nee

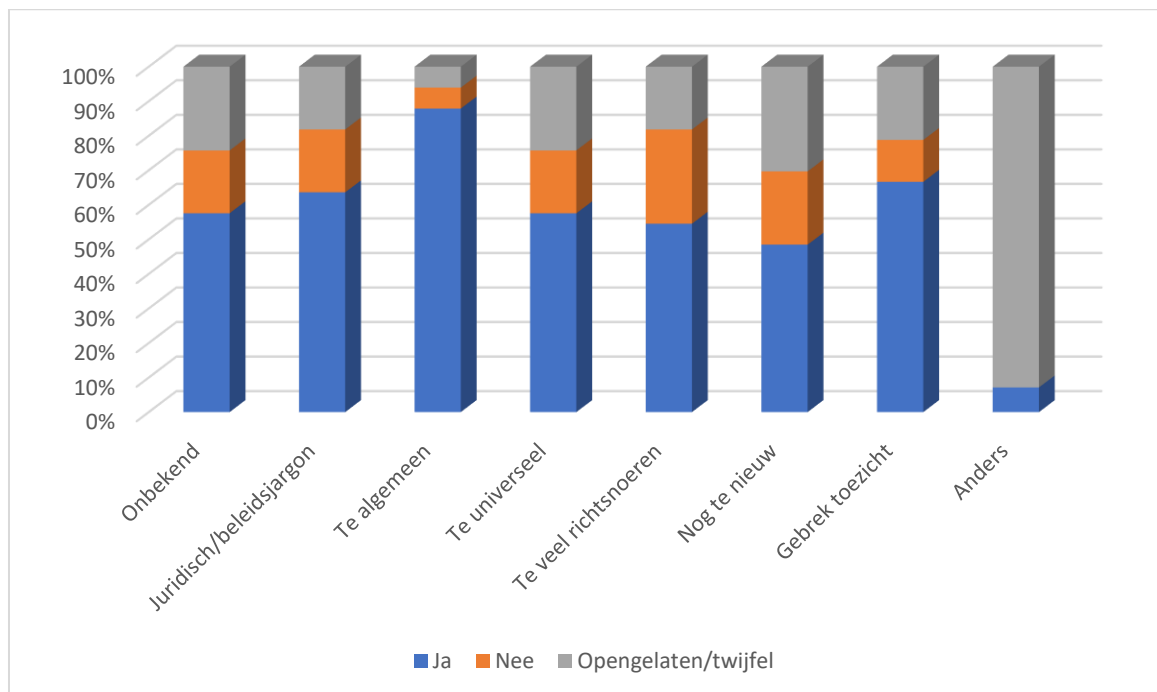
- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analisten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit:
- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit:
- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analisten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

De resultaten die zijn binnengekomen zijn als volgt (hierin zijn niet meegenomen de respondenten die deze vraag hebben opengelaten). Daarbij valt allereerst op dat bijna geen enkele respondent een andere oorzaak heeft opgegeven dan diegenen die in de enquête waren vervat; klaarblijkelijk geven de gegeven categorieën een vrij exhaustieve verklaringslijst. Ook is er geen respondent geweest die aangaf het oneens te zijn met de aanname dat er een hiaat tussen wet en praktijk bestaat of bij alle genoemde categorieën nee heeft ingevuld. Alhoewel niet expliciet is gevraagd naar de mate waarin respondenten deze lacune ervaren (bijvoorbeeld

van zeer klein tot zeer groot), blijkt uit de uitleg bij de antwoordcategorieën dat de urgentie op bepaalde punten zeer aanwezig is.



Van de categorieën valt op dat bij alle mogelijkheden vaker eens dan oneens is aangekruist. Duidelijk is dat vrijwel alle respondenten als belangrijke oorzaak voor de lacune tussen beleid en praktijk zien dat de richtsnoeren nu te algemeen zijn. Ook het feit dat ze in te universele en abstracte termen zijn opgesteld, het gebrek aan toezicht en het feit dat ze in juridisch- of beleidsjargon zijn opgesteld worden vaak genoemd als oorzaken voor de lacune. Aan de andere kant is minder dan de helft van de respondenten het eens met de stellingen dat de richtsnoeren nog te nieuw zijn om te worden meegenomen en dat er te veel richtsnoeren zijn om in acht te nemen. Tot slot denkt ongeveer de helft van de respondenten dat de oorzaak kan worden gevonden in het simpele feit dat systeembouwers en mensen die in de praktijk aan de slag gaan met de systemen de diverse bestaande richtsnoeren en beleidsdocumenten niet kennen. Ook op dat punt is derhalve nog het nodige te winnen. Hieronder wordt kort een aantal, maar zeker niet alle, relevante toelichtingen bij de gegeven antwoorden weergegeven ter illustratie, achtergrond en duiding van de gegeven antwoorden.

- **Onbekend:** Bij de antwoorden op categorie ‘Onbekend’ valt op dat een aantal respondenten met een juridische/beleidsachtergrond het beeld hebben dat ethische en juridische vragen simpelweg niet leven bij de ‘techneuten’. De onbekendheid met de bestaande standaarden zou in dat geval voort kunnen komen uit desinteresse of een gebrek aan prioriteit.
 - Ja, ik denk dat deze mensen voornamelijk bezig zijn met optimaliseren van hun product en niet met de bredere maatschappelijke implicaties ervan.
 - Law/ethics is not considered to be important in the development and technical professionals sometimes do not receive proper training in these fields.
 - Het gaat niet alleen om bekend zijn met, maar vooral dat techneuten /programmeurs niet bekend zijn met de regels die voor overheden gelden. Ze

- hebben niet het gevoel voor een overheid te werken, maar aan software te werken.
- Vaak zijn de data/computer scientists niet zo van het lezen, maar van het doen.
 - Mijn indruk is dat de techneuten wel interesse hebben, maar ook focus op hun eigen opdracht.
 - Hoewel ik binnen sommige organisaties verandering zie, werken veel ontwikkelaars taakgestuurd, zien zij het (soms expliciet) niet als hun taak om over de beleidscontext na te denken en worden ze niet actief betrokken bij de beleidscontext/besluitvorming. Ook het gebruik en beheer van AI-systemen wordt vaak belegd bij functioneel beheerders, die meestal in de organisatiestructuur geen onderdeel zijn van het beleidsteam dat de inhoudelijke afwegingen maakt.
 - Zijn primair documenten voor politici en juristen, niet voor data scientists/AI experts.
- **Jargon:** Uit de toelichting die op de antwoordcategorie 'Jargon' is gegeven blijkt aan de andere kant dat techneuten, systeemontwikkelaars en mensen die in de praktijk aan de slag gaan met AI en ML wel willen, maar vaak niet goed weten wat ze met de bestaande richtsnoeren aan moeten. Andere respondenten wijzen op het feit dat de bestaande lacune niet alleen is te verklaren door een discrepantie in taal en jargon, maar door een discrepantie in de belevingswereld als zodanig.
- Ja, niet-bindende documenten die het werk van ontwikkelaars enkel bemoeilijken motiveert hen niet om in acht te nemen.
 - Ik zou b.v. graag zien dat naast beleidsdocument ook voorbeeld code wordt gedeeld. Dan weet ik beter hoe ik dat in mijn werk kan integreren.
 - Zou kunnen maar met name juridische taal is bepalend voor het resultaat. Je hebt ITers nodig die het juridische begrijpen (zeldzaam) of je moet rule-based, vanuit het juridisch kader, gaan programmeren. Dit vergt juristen met overzicht over de materie en aardigheid in IT.
 - Discussie rond wet- en regelgeving wordt in eerste instantie binnen de eigen beroepsgroep gevoerd. Binnen een beroepsgroep zal eerst de controverse overwonnen moeten worden alvorens beginselen algemeen aanvaard kunnen worden en het mogelijk wordt deze beginselen in andere bewoordingen toegankelijk voor derden te maken.
 - Het bekende eigentaal fenomeen, dat verder gaat dan jargon! Er zijn concepten die per discipline van inhoud verschillen en zelfs binnen een discipline variëren. 'Vertalen' zou wel eens ontoereikend kunnen blijken
- **Algemeen:** Bij de vraag naar de algemeenheid van richtsnoeren wordt primair instemmend gereageerd.
- absoluut! [] De richtlijnen van EC, Raad voor Europa en MinJ&V zijn duidelijk in de zin dat duidelijk is welke waarden en uitgangspunten we moeten hanteren; maar als je moet programmeren is 'transparantie' echt te vaag.
 - Het automatiseren van open normen valt niet mee. Toch moeten voor succes de juridische en IT-technische wereld bij elkaar worden gebracht. Rule based werken kan een gemeenschappelijkheid creëren.
 - Ja, dit is een probleem. Veel organisaties willen graag maar weten niet hoe.
 - Ja, het moet concreter (wat is transparantie?). Met name aan de voorkant (risico-inschatting) gebeurt denk ik te weinig. Data scientists zijn wel degelijk bezig overigens met het voorkomen en detecteren van bias, omdat het ook de voorspellende waarde van je model ondermijnt.

- Vaak zijn de richtlijnen zo geformuleerd dat je er niet mee oneens kan zijn.
- I agree, because even for a legal expert these principles are sometimes hard to understand. However, in the daily life the developers/users of the AI systems are the ones who have to observe the rules. This is not or just limited possible, if the principles are too general to understand for not legal informed persons.
- **Universeel:** Ook op de vraag naar de universaliteit van de richtsnoeren wordt primair instemmend geageerd door de respondenten. Net zoals bij de vorige antwoordcategorie stellen sommigen zich echter ook de vraag of het wel anders zou kunnen.
 - De ontwikkeling en werking van AI-systemen is sterk afhankelijk van de context waarbinnen zij opereren. Algemene standaarden vangen de bijzonderheden van deze context niet altijd, al leggen zij wel een abstracte ondergrens neer.
 - Het ontwikkelen van een algoritme is heel domein-specifiek en vergt een veelheid aan keuzes (b.v. waardes van hyperparameters). Ook hier: graag voorbeelden of misschien zelfs referentieimplementaties.
 - AI technologie is zo divers, dat het moeilijk is dit te generaliseren naar principes. Dat kan wel, maar dan kom je op het niveau uit waarvoor nu wordt gekozen, en dan blijft het vaak bij het functionele (het domein waarin de toepassing wordt gebruikt). Het technische wordt vaak overgeslagen omdat dit dus niet te generaliseren valt.
- **Veelheid:** Over het algemeen zien de respondenten dat er een veelheid aan richtsnoeren is ontstaan, maar over de beoordeling daarvan bestaat discussie.
 - Ik vind het een super goed teken dat er een wildgroei aan richtsnoeren en standaarden is ☺ dat betekent dat veel mensen een probleem voelen! En dat was een aantal jaar geleden echt nog niet zo op dit gebied. We zijn er ook nog lang niet uit met elkaar wat echt werkt. Maar de echt werkbare standaarden en richtsnoeren komen ‘vanzelf’ bovendrijven. Hoewel het natuurlijk wel helpt als er centraal een beetje gecoreerd wordt; het is prettig als bijvoorbeeld BZK op een gegeven moment overzicht geeft van wat er allemaal is en bepaalde tools, richtlijnen en standaarden aanbeveelt.
 - Absoluut het grootste probleem. Dit zie je al ontstaan binnen het zorgdomein.
 - De OECD werkt aan kaders en handreikingen, de Europese Commissie werkt hier aan, het Europees Parlement heeft een voorstel gedaan voor een verordening, toezichthouders werken hier op sectoraal niveau aan, op nationaal niveau wordt door de wetgever nagedacht: er is inderdaad erg veel. En dan heb je ook nog eens verschillende toezichthouders die vinden dat zij er over gaan: voor de financiële sector AFM/DNB enerzijds, en de AP anderzijds.
 - Het is niet erg dat er veel richtlijnen, etc. zijn. Dit geeft organisaties de ruimte om keuzes te maken wat ze wel/niet willen volgen. Het probleem zit in de concreetheid en de handrekeningen die uitblijven om dit te operationaliseren.
- **Nieuwigheid:** Bij de vraag of de nieuwheid van de standaarden een rol speelt bij de lacune tussen regelgeving en praktijk lopen de antwoorden uiteen. Sommige respondenten menen dat dit inderdaad het geval is en verwijzen naar de richtsnoeren die de afgelopen maanden of jaren zijn verschenen, anderen stellen juist dat het ethische kader en de discussies over AI al jaren of zelfs decennia bekend zijn.
 - Precies. Dat betekent niet dat we maar moeten afwachten of doorgaan met stimuleren van eindeloos veel initiatieven. Probeer vanuit Rijksoverheid te zoeken naar de standaarden die echt potentie hebben en stimuleer de doorontwikkeling daarvan.
 - Een aantal richtlijnen bestaan al best heel lang.

- **Toezichtslacune:** Bij de vraag naar de toezichtslacune geven veel respondenten aan dat er vrijwel geen enkel toezicht bestaat. Dit wordt als probleem ervaren; de oorzaak van de toezichtslacune wordt gezocht zowel in het gebrek aan expertise en het gebrek aan prioriteit bij toezichthouders als het feit dat de bestaande richtsnoeren niet echt te handhaven zijn omdat de daarin vervatte principes te vaag en algemeen zouden zijn.
 - Ja, zeker. Maar ook: het ontbreekt ons aan bestaande/enforceable/auditable standaarden.
 - Een brug te ver, niet te handhaven, te vaag, niet afdwingbaar, geen reële wetgeving etc..
 - Het is zo dat de richtlijnen en standaarden nog geen dwingend karakter hebben. Echter zijn ze nu zo “open” dat dit ook vrij lastig zal zijn om te doen.
 - Toezicht en handhaving richt zich op bindende kaders (zoals de AVG en Awb). Ik kan mij goed inbeelden dat, dat voor niet bindende standaarden (waarvan er bovendien meerdere zijn) in het geheel niet het geval is, althans niet van overheidswege.
 - Toezicht is nog niet toegerust om tot adequaat en effectief toezicht te komen. Daarvoor ontbreekt de expertise. Met name de expertise om te zien wat in de praktijk daadwerkelijk gebeurt vraagt om een combi van technische, juridische en beleidsmatige achtergrondkennis. Dat is op dit moment moeilijk te organiseren, onder meer doordat technische specialisten eerder te vinden zullen zijn bij de partijen die de toepassingen aan het maken zijn.
 - Er zijn nog geen afdwingbare standaarden of wet- en regelgeving. Je kan veel AI vraagstukken onderbrengen in bestaande wet- en regelgeving (Awb, AVG, etc.) maar dat is op een hoog abstractieniveau. Aldus is er geen specifieke handhaving op de zorgvuldige toepassing van AI (audits van modellen e.d.).
 - Toezichthouders beginnen zich nu pas te verhouden tot AI. Pas als duidelijk is waarop toezicht en handhaving zich zal richten zullen ondertoezichtgestelden zich tot AI principes gaan verhouden.
 - Ik zou het goed vinden als we in Nederland een soort risicoprofielen op gaan stellen voor algoritmes (ik geloof dat Harvard met zoiets bezig is). En ik zou het dan logisch vinden dat we toezicht en handhaving inrichten langs die risicoprofielen: een algoritme met het laagste risicoprofiel kan door een ontwikkelaar alleen in de gaten gehouden worden, iets hoger bijvoorbeeld door een ontwikkelaar en beleidsverantwoordelijke; en een algoritme met het hoogste profiel wordt door een nationale toezichthouder gecontroleerd.
- **Tips:** Tot slot is de respondenten gevraagd naar tips en adviezen voor het ontwikkelen van een nieuwe handreiking. Uit de antwoorden blijkt een nadruk op en noodzaak van een interdisciplinair team; gezien de vorige antwoorden wordt ook de wens uitgesproken dat deze handreiking specifiek zal zijn. Wel is er scepsis of een handreiking überhaupt de bestaande lacune zou kunnen dichten.
 - Mijn ervaring is dat veel beleidsmakers (oa HLEG) van mening zijn dat zij geen guidance hoeven te geven over de vertaling van wetgeving naar algoritmen. Dit is echter verre van triviaal! Er zijn veel wiskundige definities van fairness en deze zijn in het algemeen strijdig. Dit speelt met name voor ‘discriminatie’ via proxy-variabelen. Zo kan het bijv. dat de ontwikkelaar van COMPAS kan zeggen dat het systeem deugt terwijl de waakhond ProPublica een fundamenteel probleem zag. Vanuit die optiek ben ik benieuwd hoe de handreiking tot stand gaat komen en wat het detailniveau gaat zijn. Naar mijn mening zijn er al meer dan genoeg rapporten en raamwerken met hoog-over principes (uitlegbaarheid,

fairness, transparantie etc.). Zaak is dus om een handreiking te ontwikkelen waarmee een data scientist uit de voeten kan. Mijn concrete vraag zou zijn: hoe kan ik als ontwikkelaar aantonen dat een model voldoet aan de wetgeving. In zekere zin vereist dit dus rule-based kaders als aanvulling op highly principle-based wet- en regelgeving.

- Een GDPR achtige AI wetgeving van de EC is wenselijk. Verder is het belangrijk per domein te kijken welke principes belangrijk zijn. In de zorg is privacy bijvoorbeeld belangrijker dan in de financiële sector. Het zou helpen om consolidatie te brengen; dus evt alle stukken door gaan en dan de gemeenschappelijkheden in 1 document presenteren.
- Een handleiding op zichzelf is nuttig, maar idealiter moet dit worden ingebed in een breder governance proces voor AI toepassing (model risk management).
- Het opbreken in stappen/delen van het proces is logisch op papier, maar in de praktijk is het veel rommeliger. Stappen vloeien in elkaar over, en er hoeft geen lineaire ontwikkeling te zijn. Neem het socio-technische aspect mee. Het systeem bestaat niet in een vacuüm. Als er als aan het project als iets schort kun je je AI nog zo non-discriminatoir maken, maar dan schiet je er als nog niets mee op.

0.4 Definitie van AI als gehanteerd voor dit onderzoek

De term Artificiële Intelligentie heeft sinds de jaren 50 - toen de term voor het eerste geïntroduceerd werd - vele verschillende betekenissen en definities gekregen. Waar de term voorheen met name refereerde aan een opkomend onderzoeksveld gericht op het begrijpen en nabootsen van (menselijk) denken en gedrag, verwijst hij vandaag de dag vooral naar een aantal slimme technologieën die steeds vaker onderdeel zijn van alledaagse computersystemen. Zo omschrijven Russell en Norvig in hun veelgebruikte handboek AI als een breed onderzoeksveld dat veel verschillende methodes, perspectieven en aanpakken omvat en dus ook door diverse betrokkenen op verschillende manieren is gedefinieerd.²³ Zij maken een onderscheid tussen definities die de nadruk leggen op enerzijds het bestuderen of nabootsen van *denken* of *redeneren* met behulp van computers, en anderzijds definities die juist de nadruk leggen op *gedrag*. Een tweede onderscheid dat zij maken is tussen definities die succes van AI-technieken afmeten aan hoe gelijkend de prestaties van het systeem zijn aan menselijke prestaties of aan hoe *rationeel* de resultaten zijn (als een ideaal om naar te streven). Russell en Norvig zien zelf AI als computersystemen die in een omgeving het beste (meest rationele) gedrag laten zien gegeven de kennis die zij hebben.

De term is tegenwoordig ook steeds meer verbonden met nieuwe slimme technologieën die hun weg gevonden hebben in alledaagse computersystemen, van slimme deurbellen, smart phones en horloges, tot online platforms, fraudebestrijdingssystemen, auto's en fabrieksrobots. De *Higher Level Expert Group on AI* (HLEGAI) van de Europese Commissie definieert AI-systemen als systemen die in een fysieke of digitale wereld een complex doel proberen te bereiken door hun omgeving waar te nemen, de verzamelde structuren of data te interpreteren; te redeneren over kennis afgeleid van deze data en zo te bepalen hoe het doel het beste bereikt kan worden.²⁴ Verder, stelt de HLEGAI dat AI-systemen ook zodanig kunnen worden

²³ S. J. Russell & P. Norvig (2016). Artificial intelligence: a modern approach. (third edition), Prentice Hall.

²⁴ The European Commission's High-Level Expert Group on Artificial Intelligence (2018). A Definition of AI: Main Capabilities and Scientific Disciplines. 12 December, p. 6.

ontworpen dat zij hun gedrag aanpassen aan de omgeving door te analyseren hoe eerdere handelingen van het systeem de omgeving beïnvloeden.

De HLEGAI maakt ook een onderscheid tussen AI als systeem of technologie en AI als een wetenschappelijke discipline. Als wetenschappelijke discipline kent AI, volgens de HLEGAI, verschillende vakgebieden, aanpakken en technieken. De groep stelt dat de meeste technieken kunnen worden onderverdeeld in twee groepen: *redeneren* en *leren*. Onder *redeneren* vallen technieken om kennis te representeren, plannen en roosteren (*scheduling*), zoeken en optimaliseren. Met behulp van deze technieken kan, bijvoorbeeld, kennis afgeleid uit sensordata worden gemodelleerd, naar de beste oplossing voor een planning vraagstuk worden gezocht of een optimale oplossing voor een complex probleem worden geïdentificeerd. Het gaat hier veelal om regel-gebaseerde systemen (gebaseerd op bijvoorbeeld formele logica), maar ook om technieken die gebruik maken van statistische methodes. Regel-gebaseerde systemen bestaan uit expliciet geformuleerde regels - vaak in niet-numerieke symbolen uitgedrukt - die specificeren wat het systeem moet doen.

Onder *leren* vallen de technieken die het mogelijk maken om problemen op te lossen die niet precies kunnen worden gespecificeerd of kunnen worden uitgedrukt in expliciete regels. Het gaat dan bijvoorbeeld over technieken zoals neurale netwerken, *Deep Learning (DL)*, en beslisbomen. Het soort problemen waarop deze technieken vaker wordt toegepast zijn onder andere het begrijpen van taal, beelden interpreteren en het voorspellen van gedrag. Zo zijn er lerende technieken om foto's van gezichten te classificeren in gegeven categorieën, zoals 'man' of 'vrouw' of 'blij' of 'boos'. Andere lerende technieken kunnen voorspellingen doen op basis van data over de waarschijnlijkheid van bepaalde gebeurtenissen, zoals de waarschijnlijkheid dat een bepaald persoon een lening zal aflossen. De technieken die onder het kopje *leren* vallen, kunnen gebruik maken van regel-gebaseerde technieken (zoals bijvoorbeeld beslisbomen), maar veelal maken zij gebruik van neurale netwerken of probabilistische of statistische methodes (Naive bayes, support-vector machines, cluster methoden, lineaire regressie). Veel van de verschillende aanpakken, methodes en technieken onder de twee kopjes die voortgekomen zijn uit het onderzoeksveld AI vinden hun toepassing in AI-systemen voor gebruik in de praktijk.

Met name AI-systemen die gebruik maken van lerende technieken, ook wel Machine Learning (ML) technieken genoemd, zijn de laatste jaren in opkomst als gevolg van doorbraken op dit gebied. In het bredere discours gebruikt men tegenwoordig zelfs de term AI als inwisselbaar met ML. Hoewel er diverse ML-technieken zijn, heeft vooral Deep Learning de laatste jaren de aandacht getrokken en wordt deze techniek inmiddels op grotere schaal toegepast in allerlei applicaties, zoals bijvoorbeeld gezichtsherkenning.²⁵ Het gaat hierbij om artificiële neurale netwerken – algoritmes die op een sterk versimpelde manier menselijke neurale netwerken nabootsen. Deze netwerken worden getraind aan de hand van grote hoeveelheden data - zoals digitale beelden - om bepaalde patronen in de data te herkennen. Deep Learning systemen hebben niet een volledig gespecificeerd stappenplan nodig om tot een uitkomst te komen, maar kunnen, gegeven een hoeveelheid data, zelf een model maken dat de juiste output bij bepaalde input geeft. Aan de hand van nieuwe data kunnen zij dat model weer aanpassen. Het voordeel van deze techniek is dat het niet precies bekend hoeft te zijn wat bijvoorbeeld een kat van een hond onderscheidt om de digitale beelden aan die twee verschillende categorieën toe te kennen.

²⁵ I. Goodfellow, Y. Bengio & A. Courville (2016). Deep learning. MIT press.

Het nadeel is dat dan ook niet precies duidelijk is op basis van welke correlaties het systeem het onderscheidt maakt en of die wel relevant, veilig of legaal zijn.

Nu DL-gebaseerde systemen sterk in opkomst zijn neemt de roep om meer inzicht in en uitlegbaarheid van DL-systemen ook toe. Hiermee groeit ook de aandacht voor systemen die verschillende technieken, bijvoorbeeld regel-gebaseerde technieken, combineren met neurale netwerken om zo systemen transparanter te maken. Tegenwoordig is het dan ook niet voldoende om alleen naar DL- of ML-systemen te kijken, maar moet men naar de verschillende soorten AI-systemen en -technieken kijken. Bovendien zijn veel AI-technieken vaak onderdeel van grotere hybride systemen die meerdere soorten technieken combineren.

Het is dan ook lastig om een duidelijke scheidslijn te trekken tussen AI en niet-AI, maar met het oog op het voorkomen van discriminatie door AI-gebaseerde systemen zijn er wel een aantal kenmerken uit de definitie van HLEGAI die helpen om de voor dit rapport gekozen focus nader te omschrijven. Ten eerste leiden met name de nieuwe vormen van onzekerheid die lerende systemen met zich meebrengen tot zorgen over (onbedoelde) discriminatie. Als een taak vooraf niet volledig en expliciet gespecificeerd kan worden en er technieken worden gebruikt die zelf patronen en correlaties afleiden uit data, kunnen er biases ontstaan die niet eenvoudig of direct te herkennen zijn. Ook is het moeilijker om te voorspellen hoe een systeem gebaseerd op lerende technieken zal interacteren met een nieuwe omgeving of nieuwe data. Ten tweede kunnen systemen die een complex doel proberen te bereiken door het combineren van veel verschillende soorten data al snel moeilijk te interpreteren en controleren worden. Ten derde en tot slot hebben veel AI-systemen een grote mate van zelfstandigheid (autonomie) in het uitvoeren van hun taken, zonder dat daar direct toezicht of directe controle van mensen voor nodig is. Ook als gevolg van deze autonomie kunnen zich nieuwe onzekerheden en onvoorspelbaarheden voordoen die discriminatie in de hand kunnen werken.

In dit rapport wordt daarom de nadruk gelegd op systemen die:

- een zekere mate van autonomie hebben;
- een bepaalde mate van complexiteit hebben die het moeilijk maakt om inzicht te hebben in hoe het systeem tot een bepaalde uitkomst komt;
- en/of zich (of modellen) kunnen aanpassen aan de hand van een analyse van eerdere acties en de omgeving/nieuwe data.

Dat kunnen systemen zijn die gebaseerd zijn op lerende technieken, maar het kan even goed gaan om een regel-gebaseerde of statistische technieken.

0.5 Aanpak en leeswijzer

Om dit onderzoek te verrichten is gekozen voor een brede en interdisciplinaire benadering. Het gaat er om een link te leggen tussen de realiteit van juristen en beleidsmakers enerzijds en die van data-analisten en systeembouwers anderzijds. Bij vraagstukken omtrent discriminatie en gelijke behandeling in AI-systemen spelen onder meer juridische, statistische, ethische, sociale en technische vraagstukken. Het schrijfteam dat betrokken is bij dit onderzoek kent daarom een brede diversiteit aan achtergronden. Het College voor de Rechten van de Mens, een instituut dat is gespecialiseerd in discriminatiewetgeving en daar oordelen over velt; Julia Janssen is creative designer is verantwoordelijk voor de vormgeving van de handreiking; Esther Keymolen is ethica gespecialiseerd in vertrouwen in en de normatieve randvoorwaarden van moderne technologieën; Merel Noorman heeft een STS achtergrond en is expert op het gebied

van AI en ML-systemen; Mykola Pechenizkiy is gespecialiseerd in data mining en heeft een achtergrond in knowledge discovery systems; Hilde Weerts is Artificial Intelligence Engineer; en Bart van der Sloot is gespecialiseerd op het gebied van mensenrechten en beleidskeuzes ten aanzien van de data-gedreven samenleving. Het schrijfteam heeft ook een interne klankbordgroep samengesteld bestaande uit Mark Bovens (WRR), Francien Dechesne (LU), Ronald Leenes (TiU), Egge van der Poel (TIAS) en Johan Wolswinkel (TiU).

Om tot een handreiking tot komen zijn de volgende stappen gezet, die ieder worden beschreven in een apart hoofdstuk.

0.5.1 Stap 1: Probleemanalyse (hoofdstuk 1)

Een probleemanalyse ten aanzien van het bestaan van discriminatie en bias in AI en ML-systemen wordt primair gedaan aan de hand van literatuuronderzoek, aangevuld met de resultaten van de twee enquêtes die zijn verspreid onder juristen, beleidsmakers, advocaten en burgerrechtenorganisaties enerzijds en onder data-analisten, systeembouwers en organisaties die in de praktijk aan de slag gaan met dergelijke systemen anderzijds. De twee enquêtes komen in grote lijnen overeen, maar stellen op punten ook andere vragen en maken soms gebruik van andere begrippen en andere uitleg om beter aan te sluiten bij de belevingswereld van de twee groepen. De lijst van personen en organisaties waaraan de enquête is verzonden is tot stand gekomen op basis van een select aantal namen opgegeven door de leden van het schrijfteam, de interne klankbordgroep en de klankbordgroep zoals samengesteld door de opdrachtgever (BZK), bestaande uit vertegenwoordigers van diverse ministeries, uitvoeringsinstanties en private organisaties. Het gaat hierbij om onder meer diverse professoren aan universiteiten, experts die systemen ontwerpen en bouwen en private en publieke organisaties die met dergelijke systemen aan de slag zijn of daar toezicht op houden. De enquêtes en de reacties daarop zijn opgenomen in de bijlagen bij dit rapport.

0.5.2 Stap 2: In kaart brengen bestaande standaarden (hoofdstuk 2)

De tweede stap bestaat uit een beschrijving van bestaande standaarden en principes in literatuur, beleidsdocumenten en jurisprudentie ten aanzien van non-discriminatie en gelijke behandeling. Daarbij wordt met name ingegaan op de ontwikkelingen van de laatste 5 jaar. Daarbij wordt zowel de juridische standaarden, technische principes en statistische methodieken bestudeerd en wordt gekeken naar overlap, verbanden en conflicten. Op basis hiervan is een overzicht ontwikkeld van de bestaande documenten, standaarden en principes die reeds zien op AI en discriminatie specifiek of automatische besluitvorming en fairness in het algemeen. Tevens worden hier de resultaten uit de twee eerder genoemde enquêtes meegenomen.

0.5.3 Stap 3: Evaluatie bestaande standaarden (hoofdstuk 3)

In de derde stap is beoordeeld in hoeverre de bestaande standaarden (stap 2) de gesignaleerde problemen (stap 1) adequaat adresseren. Dit is gedaan door een bespreking van een samenvatting van de standaarden en een overzicht van de huidige en mogelijk toekomstige problemen in drie testgroepen. Een groep bestaat primair uit juristen, beleidsmakers en burgerrechtenorganisaties; de tweede groep bestaat primair uit data-analisten, technici en statistici; tot slot bestaat de derde groep uit een gemêleerd gezelschap, om zo een interdisciplinaire discussie op gang te brengen. Een belangrijkaandachtspunt in deze stap is de toepassing van juridische standaarden op de technische realiteit, mogelijke lacunes in de

standaarden en de vraag welke standaarden als meest adequaat uitgangspunt gelden voor de ontwikkeling van een handreiking voor non-discriminatie by design systeemprincipes. Tevens worden hier de resultaten uit de twee eerdergenoemde enquêtes meegenomen en wordt vervolgonderzoek gedaan ten aanzien van de in de testgroepen gesignaleerde aandachtsgebieden.

0.5.4 Stap 4: Ontwikkeling handreiking (hoofdstuk 4)

In de vierde stap is de eerste aanzet gemaakt voor de ontwikkeling van een handreiking. Naast een algemeen deel wordt op drie gebieden een vertaalslag gemaakt van de specifieke jurisprudentie ten aanzien van gelijke behandeling en anti-discriminatie, namelijk het gebruik van AI in de strafrechtketen, in de gezondheidszorg en op de arbeidsmarkt.

0.5.5 Stap 5: Validatie standaarden (hoofdstuk 5)

Vervolgens worden de ontwikkelde standaarden getest. Dat gebeurt in drie workshops. Een zal zijn toegespitst op het gebruik van de handreiking ten aanzien van het gebruik van AI binnen de strafrechtketen en een ander binnen de medische context. De derde workshop zal een meer algemene workshop zijn waarin experts met diverse achtergronden en expertisegebieden met elkaar in discussie gaan. Tijdens deze workshops staat de vraag centraal of alle relevante principes en standaarden zijn verwerkt in de handreiking en of er verder risico's zijn die niet zijn afgedekt. Ook zal worden gekeken naar waar knelpunten liggen en waar er zich onduidelijkheden, miscommunicaties en conflicten kunnen voordoen. Per domein wordt gekeken of de principes genoeg ruimte laten en tegelijkertijd voldoende houvast bieden, gezien de specifieke vereisten van het domein.

0.5.6 Stap 6: Finaliseren standaarden (hoofdstuk 6)

Tot slot worden de standaarden zoals ontwikkeld in stap 4 gefinaliseerd met behulp van de bevindingen in stap 5. Tijdens deze laatste stap wordt een visualisatieslag gemaakt. De algemene juridische, statistische en ethische principes worden op een toegankelijke manier gepresenteerd door middel van icoontjes met een begeleidende tekst. De uiteindelijke handreiking zal los van dit rapport komen te staan en zal een begeleidende tekst hebben per principe en toepassingsdomein.

1. Probleemanalyse

Dit hoofdstuk geeft een kort overzicht van de belangrijkste problemen ten aanzien van discriminatie, fairness en daarmee samenhangende beginselen zoals transparantie en controleerbaarheid van AI-systemen en daarop gebaseerde besluitvorming. Eerst zal een korte, algemene beschouwende tekst worden gegeven (paragraaf 1.1) met daarin bouwstenen voor, maar ook de begrenzingen van dit rapport en de ontwikkeling van de handreiking. Dit heeft een hoger abstractniveau en gaat in op het karakter van taal, data en classificering als zodanig. Dan volgt een literatuuranalyse (paragraaf 1.2), met daarin aandacht voor het bestaan van bias, transparantie en fairness. Deze analyse is primair gestoeld op een analyse van literatuur uit de ethiek en de STS-community. Dan volgt een analyse van de literatuur en de daarin beschreven punten uit de informatica en data mining wereld (paragraaf 1.3). Tot slot wordt een overzicht gegeven van de belangrijkste aanvullende, vaak bredere en maatschappelijke thema's, die zijn geïdentificeerd in diverse beleidsdocumenten en rapporten (paragraaf 1.4).

1.1 Bouwstenen

Wat een analyse van de problemen die zijn gemoeid met gevaren omtrent discriminatie, stigmatisatie en ongelijke behandeling zo ingewikkeld maakt is dat zij samenhangen met en gebaseerd zijn op een aantal hele basale concepten, ideeën en benaderingen die eigen zijn aan de taal, samenleving en maatschappijinrichting. Die zullen niet veranderen en kunnen dat voor een belangrijk deel ook niet. Toch is het belangrijk om ze kort te bespreken, omdat een deel van de (onvermijdelijke) problemen met AI daardoor verklaard en begrepen kan worden. Een deel van de discriminatoire elementen en gevolgen van artificiële intelligentie hangt simpelweg samen met de bouwstenen van onze samenleving en zal daarom niet kunnen worden weggenomen door een handreiking.

1.1.1 Taal en categorisering

Een van de oudste filosofische discussies betreft de vraag omtrent categorieën. Terwijl Plato er in essentie van uitging dat concepten, categorieën en ideeën als zodanig bestaan in een ideële zijns wereld, waar mensen slechts indirect toegang tot hebben en die zij slechts via de concrete objecten die een particuliere gestalte zijn van die abstracte ideeën kunnen kennen,²⁶ redeneerde Aristoteles andersom: begrippen, concepten en categorieën bestaan niet van nature, zij zijn door mensen gemaakt en volgen de feitelijke realiteit.²⁷ Het idee of concept van een boom is in deze visie in feite het destillaat van verschillende waargenomen objecten, wier overeenkomstige karakteristieken worden samengebundeld in één overkoepelend concept. In feite draait de discussie tussen de twee om de vraag naar een inductief of deductief epistemologisch model. Alhoewel er nog steeds aanhangers van de Platonische visie zijn, gaan veruit de meeste mensen tegenwoordig uit van het idee dat concepten, categorieën en ideeën een menselijk construct zijn.²⁸ Ook dit rapport zal dat doen. In die zin kunnen categorieën worden begrepen als concepten die de wereld inzichtelijk maken. In plaats van een brij aan feiten, staan concepten

²⁶ J. Adam (Ed.) (1902). *The republic of Plato* (Vol. 2). University Press.

²⁷ J. L. Ackrill (1963). *Aristotle: Categories and De Interpretatione*, Oxford: Clarendon Press. Annas, J., 1974. "Individuals in Aristotle's Categories: Two Queries," *Phronesis*, 19: 146–152.

²⁸ Wel is er nog steeds discussie over de vraag wat er als eerste is: verschil of overeenkomst. J. Williams, (2013). *Gilles Deleuze's Difference and repetition*. Edinburgh University Press.

en categorieën de mens toe orde te scheppen en verbanden, overeenkomsten en verschillen te zien.²⁹

Taal is in feite zelf een samenbundeling van categorieën en begrippen, die mensen niet alleen in staat stelt hun eigen gedachten te ordenen, maar die vervolgens ook te categoriseren. Op hun beurt zijn de bestaande categorieën en begrippen weer bepalend voor hoe mensen de wereld percipiëren. Taal is daarmee een symbolische orde die gebaseerd is op en gerelateerd is aan de realiteit, maar daar ook los van staat. Taal kent een beperkte set aan woorden en woordcombinaties om de realiteit te beschrijven. Daarbij treedt een versimpeling en ‘idealisering’ (taal is vaak gebaseerd op ideaaltypen) op. Taal is bijvoorbeeld niet alleen ongeschikt om complexe emoties te beschrijven, er zijn duizend manieren waarop iemand zich boos kan voelen, terwijl er maar een term voor bestaat, maar categoriseert ongelijke objecten, elementen en verschijnselen in één groep. Terwijl elk particulier grassprietje toch weer anders is dan alle anderen worden alle grassprietjes onder dezelfde noemer geschaard. Dat is in relatie tot grassprietjes een redelijk gevolgen vrije keuze, maar uiteraard wordt dit gevoeliger naarmate de concepten en begrippen directer aan mensen en de maatschappij zijn gerelateerd.³⁰

De ontwikkeling van taal, begrippen en categorieën is daarmee ook sterk contextafhankelijk. Naarmate het onderscheid tussen of identiteit van objecten, fenomenen of organismen belangrijker is worden er meer begrippen en concepten ontwikkeld die de relevante variatie kunnen vangen (het meest basale maar klassieke voorbeeld hiervan is dat in bepaalde culturen meerdere woorden voor sneeuw bestaan, terwijl in andere culturen, waar het niet vaak sneeuwt maar wel vaak regent, er een dozijn aan woorden bestaat dat de diverse verschillen tussen regen en de intensiteit daarvan kan weergeven).³¹ Dat betekent dat de taal die dieren spreken verschilt van die van mensen, dat de culturele impact op de ontwikkeling van menselijke taal groot is en dat het gebruik van taal tussen verschillende groepen (bijvoorbeeld jargon) van elkaar verschilt. Taal geeft daarmee altijd een destillaat van de werkelijkheid in begrippen categorieën, maar welke categorieën en begrippen er worden ontwikkeld is afhankelijk van de context waarin die vorm krijgen en worden gebruikt.

1.1.2 Data en classificering

Een van de meest beschreven problemen ten aanzien van AI is het (onjuiste) geloof dat zowel data als techniek neutraal zouden zijn. Als gevolg daarvan worden de uitkomsten of beslissingen van op AI gebaseerde systemen voor ‘waar’ of ‘absoluut’ aangenomen. Een greep uit de antwoorden van de respondenten op de voor dit onderzoek uitgestuurde enquêtes bevestigt dat beeld, bijvoorbeeld ten aanzien van de vraag naar de gevaren van AI, waarop verschillende antwoorden deze richting uit wijzen: ‘Dat mensen gaan geloven in de ‘waarheid’ zoals die gepresenteerd wordt door de systemen’; ‘Te veel vertrouwen in de correctheid van de geautomatiseerde beslissingen’; ‘Het geloven van de uitkomst.’

Een van de problemen die hierbij speelt is dat het woord data, afgeleid van *datum*, oftewel gegeven, misleidend is. Data zijn niet gegeven, ze zijn het resultaat van een menselijk productieproces, zo stelt ook Maxim Februari. ‘Data hebben de zweem van neutraliteit, van

²⁹ Het menselijke brein is ook van nature gericht op het zien van verschillen en afwijkingen van de ‘normale’ toestand.

³⁰ https://repository.ubn.ru.nl/bitstream/handle/2066/148187/mmubn000001_027108686.pdf?sequence=1

³¹ <https://www.newscientist.com/article/mg21628962-800-are-there-really-50-eskimo-words-for-snow/?ignored=irrelevant>

objectiviteit. Het lijkt een datum, een gegeven, een natuurwerkelijkheid. Het tegendeel is echter waar. Data zijn gemaakt door mensen, ze zijn een product en het resultaat van subjectieve keuzes. Dat is belangrijk omdat er nu vaak nogal naïef van wordt uitgegaan dat data de waarheid representeren en data-analyse de waarheid oplevert, zonder dat daar verder nog een evaluatie overheen moet gaan. Ook worden de resultaten van data-analyses niet zelden gewoon overgenomen en gebruikt als directe input voor beleidsbeslissingen. Er is een gebrek aan controles en standaarden op dit punt.’³²

Net als taal vormen data een representatie van de werkelijkheid. Wie data verzamelt, waarover, op welke manier de gegevens worden verzameld, hoe ze worden verzameld en hoe ze worden gecategoriseerd is steeds een subjectieve keuze die sterk bepalend is voor hoe de data eruit komen te zien. Daarbij geldt dat net als bij taal er uiteindelijk tot een classificering moet worden gekomen om de data in op te slaan en te analyseren. Ook hier verschilt de classificering per context en gebruik, al geldt ook op dit punt dat data(analyse) in de perceptie soms een aparte status krijgen. Net zoals bij communicatie en vertaling tussen verschillende taalgebieden bestaan er communicatieproblemen tussen verschillende datasets, die anders tot stand zijn gekomen en een ander doel dienen, waarbij data op een andere wijze zijn verzameld en andere categorieën worden gebruikt of deze anders zijn afgebakend. Toch is soms het geloof dat AI zonder problemen data uit verschillende databases zou kunnen vergelijken en analyseren en zelfs met zogenoemde ‘ongestructureerde’ data aan de slag zou kunnen (data zijn uiteraard nooit ongestructureerd).³³

Ten aanzien van mogelijke gevaren van discriminatie, stigmatisatie en ongelijke behandeling is een eerste belangrijk punt dat omdat zowel data als technologie (zoals AI) een menselijk product zijn, de subjectieve perceptie, aannames en vooroordelen van de mens(en) die deze producten in gang zet(ten) wordt gereflecteerd in de data en de techniek. Bij de risico’s ten aanzien van AI wordt door meerdere respondenten van de enquêtes dit punt onderstreept, zoals onder meer: ‘Onbedoelde vooroordelen die altijd al onderdeel van menselijke keuzes waren vastleggen in een systeem’; ‘Er worden vooroordelen aangeleerd’; ‘Selectie bias in traindata’ en ‘Data corruption (in training data of use data)’. Het probleem van deze ‘bias’ die reeds in de mens zit is tweërlei. Enerzijds is de mens zich slechts van een (zeer) beperkt gedeelte van de aannames en vooroordelen bewust. Anderzijds is een deel van deze vooroordelen en aannames onvermijdelijk; het is onmogelijk om echt neutraal de wereld te beschouwen.

AI kan, zij het slechts gedeeltelijk, dan ook helpen bij een ‘rationaliseringsslag’, het objectiever en neutraler maken van besluitvormingsprocessen die nu ook door aannames en vooronderstellingen worden gestuurd. Die mogelijke kans zit op een aantal punten. Ten eerste kunnen de onbewuste bias en aannames die nu in besluitvormingsprocessen zijn geslopen worden gesignaleerd en eruit worden gehaald of er kan voor worden gecompenseerd. Ten tweede kunnen er meer data worden verzameld, waardoor een genuanceerder en correcter beeld kan worden verkregen. Ten derde en daarmee samenhangend kunnen de gekozen categorieën fijnmaziger zijn, waardoor steeds beter is te verwijzen naar en in een analyse te selecteren is op de meest relevante factor of datapunt.

³² B. van der Sloot & S. van Schendel, 'De Modernisering van het Nederlands Procesrecht in het licht van Big Data: Procedurele waarborgen en een goede toegang tot het recht als randvoorwaarden voor een data-gedreven samenleving', WODC 2019.

³³ T. K. Das, & P. M. Kumar (2013). Big data analytics: A framework for unstructured data analysis. International Journal of Engineering Science & Technology, 5(1), 153.a

Deze punten worden bevestigd door de respondenten van de enquêtes, die onder meer schrijven ten aanzien van de kansen van AI: ‘Door de hoeveelheid aan data die verwerkt kan worden zullen de aanbevelingen/beslissingen die daaruit volgen doorgaans zeer accuraat zijn.’; ‘Beter geïnformeerde beslissingen’.; ‘besluiten objectiever maken als die objectief moeten en kunnen zijn (voorbeeld parkeercontrole: een parkeerbeambte zou het merk van de auto of een woordenwisseling met de bestuurder kunnen meewegen in het besluit een boete uit te schrijven)’; ‘Verhelpen discriminatie’.

Ook bij de vraag of AI kan helpen bestaande ongelijke behandeling en discriminatie tegen te gaan wordt vaak bevestigend geantwoord: ‘The capacity of AI to process more data.’; ‘je kan het gebruiken om bijv. bias aan te tonen. (Zie Amazon hiring algorithm) Met die resultaten zou je bijv. weer nieuw en beter beleid kunnen voeren. Ook zou het positief zijn als besluiten beter getraceerd kunnen worden’; ‘Bestaande problemen: te beperkte dataset om juiste inferenties te plegen - AI gedreven oplossingen: cureren en bijeenbrengen van gestandaardiseerde datasets; opknippen bijeengebrachte datasets in kleinere toegesneden verzamelingen’; ‘AI zou in potentie objectiever moeten kunnen zijn dan mensen. Hiervoor is dan wel de dataselectie van groot belang’; ‘computers hebben geen last van vooroordelen.’

1.1.3 Macht en standaardisering

Dat computers, AI en data op zich geen last hebben van vooroordelen is precies ook het risico. Ze zijn, in ieder geval in aanvang, geheel het product van de menselijke geest en het menselijk handelen. Dat betekent dat alle bias, aannames en vooroordelen door toedoen van de mens zijn ontstaan en doordat de mens altijd dergelijke bias heeft is het onmogelijk om ze helemaal te vermijden. Hoogstens kan door een divers team en door kritische keuzes de bestaande bias zoveel mogelijk worden beperkt. Dit heeft een aantal zaken tot gevolg. Ten eerste wijzen veel respondenten er op dat door het geloof dat data, kunstmatige intelligentie en de uitkomsten van analyses waar, neutraal of correct zijn, de aannames en vooroordelen die de mens erin heeft gestopt dreigen te worden verabsoluteerd en ‘witgewassen’. Ten tweede is door een deel van de intransparantie van dergelijke processen niet meer goed te achterhalen in hoeverre bepaalde uitkomsten te herleiden zijn tot of te verklaren zijn door deze vooraf bestaande aannames of vooroordelen. Ten derde is een probleem dat door de feedbackloop deze aannames en vooroordelen kunnen worden bevestigd en versterkt.

Omdat hoe data worden verzameld en in welke categorieën data worden geclassificeerd uiteindelijk een subjectieve keuze is, heeft degene die hierover controle heeft uiteindelijk macht. Zelfs ogenschijnlijke basale keuzes, zoals het verzamelen van gegevens over geslacht, kunnen ongemakkelijke situaties opleveren, bijvoorbeeld voor mensen die zich niet in de ene of de andere categorie thuis voelen. Nog schrijnender wordt het als er gevolgen zijn verbonden aan het al dan niet in de ene of de andere categorie geplaatst worden. Daarbij valt niet alleen te denken aan de mogelijkheid om te trouwen met iemand van gelijk geslacht, dat in veel landen nog verboden is, maar ook bijvoorbeeld aan sporters waarvan wordt betwijfeld of zij wel vrouw zijn, gezien hun spierontwikkeling of testosteronniveau.

Het maken van categorieën en het verzamelen van data hebben niet alleen gevolgen voor de datasets en de eventuele keuzes die op basis daarvan worden gemaakt, ze beïnvloeden en veranderen de wereld. Zoals bekend verandert de aanwezigheid van een onderzoeker het gedrag van onderzoekssubjecten; ditzelfde gebeurt bij de aanwezigheid van een niet-menselijke observator, zoals een camera. Zelfs als onderzoekssubjecten zich niet bewust zijn van het feit

dat zij worden bekeken kan het enkele feit van het meten en registreren van data de werkelijkheid beïnvloeden.³⁴ Datzelfde geldt door het enkele bestaan van categorieën; mensen gaan de wereld zien door die categorieën en daardoor begrijpen.

Daarbij komt dat actoren, zoals staten, bedrijven maar ook burgers de data en categorieën die een weerspiegeling zijn van de realiteit vaak gebruiken als een indelingsmatrix voor de realiteit. Man/vrouw wordt niet alleen een handig onderscheid, voor veel burgers is dit een wezenlijk onderscheid, identificeren zij zich sterk met het ene of het andere geslacht en vinden zij mensen die niet in een van beide categorieën vallen of daarin zijn opgeschoven op zijn minst 'raar'. James C. Scott laat in zijn boek *Seeing like a State* ook zien hoe staten historisch gezien steeds meer macht en controle hebben verkregen, niet alleen door informatie over burgers, het landschap en de samenleving te verzamelen en daarop te acteren, maar door de samenleving vorm te geven aan de hand van de data en de gemaakte categorieën. Hij schets in feite een standaardiseringsproces, waarbij een standaardtaal, standaardmeter en standaardkilo langzamerhand de lokale meeteenheden vervangen:

'Most early measures were human in scale. One sees this logic at work in such surviving expressions as a "stone's throw" or "within earshot" for distances and a "cartload," a "basketful," or a "handful" for volume. Given that the size of a cart or basket might vary from place to place and that a stone's throw might not be precisely uniform from person to person, these units of measurement varied geographically and temporally. Even measures that were apparently fixed might be deceptive. The *pinte* in eighteenth-century Paris, for example, was equivalent to .93 liters, whereas in Seine-en-Montagne it was 1.99 liters and in Precy-sous-Thil, an astounding 3.33 liters. The *aune*, a measure of length used for cloth, varied depending on the material (the unit for silk, for instance, was smaller than that for linen), and across France there were at least seventeen different aunes'.³⁵ Uiteraard is een standaardisering op dit punt handig en wellicht onvermijdelijk in het kader van de voortschrijdende internationalisering, maar het heeft ook betekent dat meeteenheden die aansluiten bij de lokale omstandigheden zijn afgeschaft.

Op basis van deze standaardmaten en eenheden werd vervolgens steeds meer de realiteit vormgegeven, zo stelt Scott. Verkaveling en herverkaveling geschiedde op basis van standaardmeters en rechte lijnen op de kaart, eerder dan natuurlijke grenzen en afbakeningen van stukken land; bossen werden gekapt omdat het onduidelijk was welke vegetatie zich daarin bevond en herbeplant met rechte lijnen bomen, veelal monocultuur, ten behoeve van bosbouw; en ook binnen de veeteelt is de diversiteit in het type, geslacht en ras steeds meer achteruitgegaan ten faveure van het fokken van een superieur ras met de gewenste kenmerken (bovendien is naar schatting van het aantal zoogdieren op land nog slechts 3% wild en ongeveer 97% gedomesticeerd, dat wil zeggen huisdier, vee of proefdier).³⁶ Ook burgers, zo schets Scott, en de maatschappij zijn langzamerhand aan standaardisering onderworpen. 'The development of the personal surname (literally, a name added to another name, and not to be confused with a permanent patronym) went hand in hand with the development of written, official documents such as tithe records, manorial dues rolls, marriage registers, censuses, tax record, and land

³⁴ <<https://plato.stanford.edu/entries/science-theory-observation/>>.

³⁵ J. C. Scott (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press, p. 25.

³⁶ F. de Waal (2019). *Mama's laatste omhelzing: over emoties bij dieren en wat ze ons zeggen over onszelf*. Atlas Contact.

records.³⁷ De data over de wereld kon worden gebruikt om de wereld in kaart te brengen, te organiseren en te herstructureren. Net zoals de natuur, stelt Scott, is het voor staten het handigst als de realiteit zich zoveel mogelijk vormt naar de data en de gekozen categorieën.³⁸

Eenzelfde gevaar wordt bij AI gesignaleerd. Alhoewel het begrip ‘personalisatie’ vaak wordt gekoppeld aan data-gedreven systemen is dit in feite een misleidende term. Er worden categorieën en statistische correlaties ontdekt en vervolgens worden personen, objecten of fenomenen in die categorieën geplaatst en worden daar conclusies aan verbonden. Omdat geen enkele categorie perfect is af te bakenen en er vrijwel geen absolute correlatie bestaat, is een groep en een besluit daarom altijd zowel onder- als overinclusief. Dat betekent dat categorieën altijd een mal zijn. Net zoals het voor staten volgens sommige commentatoren handig kan zijn als de realiteit zich voegt naar de data, zo ook bestaat het idee dat het voor bedrijven die bijvoorbeeld adverteren handiger is als personen inderdaad de groepskenmerken krijgen of aannemen die zij statistisch (volgens dat bedrijf) zouden moeten hebben. Dat maakt adverteren immers efficiënter en effectiever. Alhoewel mensen niet in die mate te sturen zijn is het wel duidelijk dat alleen al het aanbod van advertenties en de diversifiëring daarvan per groep een impact kan hebben op de belevingswereld van mensen, laat staan als het gaat om bijvoorbeeld het nieuws (zie hierover meer in paragraaf 1.4).

1.1.4 Discriminatie en regulering

Een bepaald type categorisering heeft in het juridische domein een status apart gekregen en dat is discriminatie en ongelijke behandeling. Artikel 1 van de Nederlandse Grondwet stelt: ‘Allen die zich in Nederland bevinden, worden in gelijke gevallen gelijk behandeld. Discriminatie wegens godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht of op welke grond dan ook, is niet toegestaan.’ Artikel 1 van de Algemene Wet gelijke Behandeling stelt:

1. ‘In deze wet en de daarop berustende bepalingen wordt verstaan onder:
 - onderscheid: direct en indirect onderscheid, alsmede de opdracht daartoe;
 - direct onderscheid: indien een persoon op een andere wijze wordt behandeld dan een ander in een vergelijkbare situatie wordt, is of zou worden behandeld, op grond van godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat;
 - indirect onderscheid: indien een ogenschijnlijk neutrale bepaling, maatstaf of handelwijze personen met een bepaalde godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat in vergelijking met andere personen bijzonder treft.
2. Onder onderscheid op grond van geslacht wordt mede verstaan onderscheid op grond van geslachtskenmerken, genderidentiteit en genderexpressie.
3. Onder direct onderscheid op grond van geslacht wordt mede verstaan onderscheid op grond van zwangerschap, bevalling en moederschap.’

Allereerst valt daarbij op dat de categorieën die zijn genoemd limitatief zijn opgesomd. Wel moet daarbij worden bedacht dat artikel 14 van het Europees Verdrag voor de Rechten van de Mens een restcategorie bevat. ‘Het genot van de rechten en vrijheden die in dit Verdrag zijn vermeld, moet worden verzekerd zonder enig onderscheid op welke grond ook, zoals geslacht,

³⁷ J. C. Scott (1998). Seeing like a state: How certain schemes to improve the human condition have failed. Yale University Press, p. 67.

³⁸ J. Baudrillard (1994). Simulacra and simulation. University of Michigan press.

ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst, het behoren tot een nationale minderheid, vermogen, geboorte of andere status'. Desalniettemin geldt bij anti-discriminatiewetgeving en -jurisprudentie een nadruk op het maken van onderscheid op basis van een beperkt aantal categorieën. De relevante wetgeving en jurisprudentie en de standaarden die daaruit volgen voor besluitvorming gebaseerd op AI-systemen zal uitgebreid aan bod komen in hoofdstuk 2 van dit rapport. Voor de probleemanalyse die in dit hoofdstuk wordt geboden is een klein aantal punten van belang.

Ten eerste is wetgeving per definitie gericht op categorieën; een wet onderscheidt zich van een besluit of een beschikking in het feit dat het individu- en casus-overstijgend moet zijn. Wetgeving gebaseerd op individuele gevallen zet de deur open naar willekeur. Zowel de rechterlijke als de uitvoerende macht worden wel geacht de omstandigheden van het geval mee te nemen in de toepassing van de wet- en regelgeving. Toch heeft wet en beleid per definitie effect op groepen, in ieder geval bedoel en ook onbedoeld. Dat betekent enerzijds dat onderscheid maken tussen en diversifiëring van effect op groepen de bedoeling is van wetgeving. Anderzijds betekent het dat wet- en regelgeving niet casusspecifiek kan zijn. Dat brengt ook met zich dat een richtsnoer of handreiking niet de mate van specificiteit kan hebben die data-analisten en systeembouwers wellicht wensen. Wet- en regelgeving geeft nooit aan wat een persoon of organisatie in een concreet geval moet doen, het geeft algemene uitgangspunten, voorschriften en verboden die in de concrete omstandigheden van het geval moeten worden geïnterpreteerd en waarop altijd uitzonderingen mogelijk zijn.

Niet-juristen hopen of verwachten soms van juristen dat zij 'gewoon' aangeven wat de niet-juristen moeten doen, maar dat is een utopische wens. Een van de respondenten geeft bijvoorbeeld mee: 'Wees erop bedacht dat onderscheid maken vaak juist wel moet om een juiste toepassing van wet- en regelgeving te realiseren. Essentieel is dat gelijkheid soms wel en soms niet rechtvaardig is. Daar zit een waardeoordeel achter, per sector of wet variërend. Het heeft de voorkeur dat een brede consensus over welke factoren mee mogen tellen en welke niet in een systeem wordt ingebouwd. De gedachte dat we dit even in een handreiking voor de systeembouwers kunnen gieten is techno-optimisme.'

Ten tweede is van belang dat in feite alle data, categorieën en correlaties indirect naar de door de wet genoemde bijzondere categorieën verwijzen – dit doen ze per definitie, de vraag is alleen hoe groot de afgeleide correlatie naar bijvoorbeeld ras of levensovertuiging is. Een bekend voorbeeld is dat een gekozen wijk of postcodegebied een indirecte voorspellende waarde heeft voor ras of etniciteit; zo kan bijvoorbeeld 75% van de bevolking van een wijk direct of indirect afstammen van niet-westerse migranten. Maar zo zal een postcodegebied ook altijd iets indirects zeggen over bijvoorbeeld geslacht, al is het maar omdat in een wijk 50,5% vrouwen en 49,5% mannen wonen. Indirect zegt alles iets over al het andere.

Ten derde is ook de meer principiële vraag van belang of de bestaande categorieën uit de anti-discriminatie- en gelijkebehandelingswetgeving en -jurisprudentie (ras, geloof, gender, etc.) nog de juiste zijn. Als bijvoorbeeld, een gevaar dat vaak aan AI wordt gekoppeld, de armen armer worden en de rijke rijker, ook wel het Mattheus effect genoemd, en dergelijke effecten zowel door data-gedreven processen binnen de publieke als de private sector worden versterkt, is dat dan minder erg dan als er een dergelijk nadelig effect is voor mannen dan voor vrouwen? En wat als het niet gaat om armen en rijken, maar mensen met een rode auto? Een respondent geeft bijvoorbeeld mee: 'Het lijkt me goed om ook te kijken naar vormen van discriminatie die nieuw zijn en die minder snel in het oog zullen lopen, juist omdat ze zich onttrekken aan al

bestaande patronen (discriminatie op huidskleur, etc.). Voorbeeld kan zijn discrimination by device: het apparaat maakt uit wat je krijgt aangeboden. Bekend voorbeeld is de iPhone en de iPad die tot andere voorgeschotelde keuzes leiden dan een ‘gewone’ Android telefoon of tablet. Uitgangspunt is dat een iPhone een proxy is voor bestedingsgedrag en dergelijke maar dit ontstijgt traditionele sociaaldemografische categorisering (en introduceert nieuwe).’

Ten vierde zijn principes omtrent discriminatie en gelijke behandeling extern aan AI en data-analyse. Data-analyse en AI gaat juist om het maken van onderscheid. Als het zo zou zijn dat personen met een bepaalde etniciteit statistisch gezien criminelere zouden zijn dan anderen, dan is dat een relevant punt om mee te nemen. Dat verbieden uit het oogpunt van het voorkomen van discriminatie heeft daarmee ten gevolg dat systemen en voorspellingen minder accuraat worden. Er is dus altijd een trade-off tussen discriminatie en effectiviteit. Daarbij moet worden bedacht dat omdat alle indirecte data ook identifiers zijn voor directe discriminatiegronden, zoals een postcodegebied, de kleur van een auto en het type boeken dat iemand leent bij de bibliotheek. Als alle data(verbanden) uit een systeem worden gefilterd die indirect naar één van de discriminatiegronden verwijzen dan wordt het voorspellend systeem waardeloos.³⁹ Eén respondent schrijft bijvoorbeeld: ‘De juridische teksten zijn veel te generiek en het is niet duidelijk hoe je deze precies moet interpreteren. Veel juristen zijn zich bijv. helemaal niet bewust dat het primaire doel van veel modellen is om te discrimineren (in de zin van een onderscheid maken tussen objecten). De vervolgstap is dat gedefinieerd moet worden op welke dimensies niet gediscrimineerd mag worden en hoe we discriminatie willen meten (bijv. we kijken hoeveel vrouwen een hypotheek krijgen vs. % voor mannen `of we kijken hoeveel vrouwen ten onrechte geen hypotheek krijgen tov mannen). In mijn ervaring lopen we in deze stap vast: onze wetgeving biedt nu te weinig houvast.’

Ten vijfde is daarbij een extra lastigheid dat zonder directe discriminerende gegevens, het voor data-analisten en systeembouwers vrijwel ondoenlijk is om te weten of een dataset direct of indirect discriminerend is. Immers, zelfs zonder dat in de dataset zelf deze data zijn vervat, kan een dataset of de uitkomsten van op AI-gebaseerde besluitvorming discrimineren op één van de in de wet genoemde gronden. Vele data-analisten pleiten er dan ook voor om dit soort gegevens juist te verzamelen. Een van de respondenten schrijft op dit punt dan ook: ‘Enigszins paradoxaal lijkt de huidige controle ter bescherming van de privacy de ‘eerlijkheid’ in algoritmen tegen te werken. Voor het controleren of corrigeren van discriminatie tussen bepaalde groepen is het nodig om te weten wie tot welke groep behoort. Dit betekent al snel dat er met ‘bijzondere persoonsgegevens’ gewerkt moet worden, en dat levert meer werk en weerstand op door de privacywetgeving. Het is daardoor minder aantrekkelijk geworden voor instanties om te controleren op, en te corrigeren voor, ongewenst discriminatie.’ Terwijl één van de pijlers van het gegevensbeschermingsrecht de zogenoemde dataminimalisatie is, kan het zowel voor de transparantie, als voor mogelijke controles op fairness van AI-systemen als in het kader van de voorkoming van discriminatie wenselijk zijn om juist meer gegevens en gevoeligere gegevens te verzamelen en deze langer te bewaren dan strikt noodzakelijk zou zijn voor het doen van voorspellingen.⁴⁰

³⁹ Ook op dit punt bestaat er een overeenkomst met de privacyliteratuur. ‘Data can be either useful or perfectly anonymous but never both.’ <https://www.uclalawreview.org/pdf/57-6-3.pdf>

⁴⁰ Zie al eerder: B. van der Sloot, ‘From Data Minimization to Data Minimummization’, In: B. Custers, T. Calders, B. Schermer & T. Zarsky (eds.), ‘Discrimination and Privacy in the Information Society’, Springer, Heidelberg 2013, p. 273-287.

Tot slot zijn nog drie punten relevant, die later in dit hoofdstuk uitgebreider aan bod zullen komen, en dat betreft de zogenoemde historische bias. Zowel organisaties binnen de publieke als binnen de private sector hebben nu al veel data in hun bezit.

- Daarbij is een allereerste probleem dat er een bias in die dataset kan zitten. Een klassiek voorbeeld is dat de politie meer heeft gesurveilleerd in bepaalde ‘probleemwijken’, waar bovengemiddeld veel mensen met een migratieachtergrond wonen, waardoor er bovenmatig veel data over die personen in een dataset zitten. Een ander voorbeeld is dat medicijnen traditioneel zijn getest op mannen van 20-45 met een westerse achtergrond. Om echt een frisse start te maken zou die dataset helemaal moeten worden verversd en aangepast. Het is vrijwel onmogelijk om correcte/onbevooroordeelde voorspellingen te doen zonder nieuwe data over bijvoorbeeld de effecten van medicijnen op vrouwen.
- Ten tweede zijn databases nu vaak aangelegd met gebruikmaking van bepaalde categorieën, terwijl in die keuze alleen al een bias en keuze is gelegd. Waarom zou de overheid bijvoorbeeld überhaupt registreren of iemand man of vrouw is; waarom is dat relevant? Daarbij komt dat deze categorieën nu worden gebruikt als indirecte indicatoren. Vrouwen worden bijvoorbeeld opgeroepen voor preventief onderzoek naar baarmoederhalskanker, terwijl niet alle vrouwen van nature een baarmoeder hebben, die bij sommigen om medische redenen is verwijderd en vrouwen die ooit een man waren ook geen baarmoeder hebben. In de meeste toepassingen van de categorieën ‘man’ en ‘vrouw’ zijn dat eigenlijk niet de juiste of meest precieze factoren die nodig zijn voor de analyse of besluitvorming, het zijn slecht in mindere of meerdere mate geschikte indirecte identifiërs voor de ‘eigenlijke’ categorie. Hetzelfde geldt voor de statistische correlatie tussen etniciteit en criminaliteit, waarbij etniciteit een indirecte identifiër kan zijn voor verklarende oorzaken voor criminaliteit, zoals armoede, een gebrek aan toekomstperspectief of vervreemding van de maatschappij.
- Ten derde en tot slot is de wereld zelf gebiased. Er is nu veel te doen om ‘systemisch racisme’ en de structurele ongelijkheid tussen man en vrouw. Dat betekent dat zelfs als het mogelijk zou zijn om op een volstrekt neutrale wijze data te verzamelen, deze op een neutrale manier te categoriseren en het AI-systeem op een volstrekt neutrale manier vorm te geven (quod non), een AI-systeem de ongelijkheid in de samenleving alsnog zou reproduceren. Een extreem voorbeeld is de chatbot van Microsoft die moest leren van hoe online discussies normaal verliepen en op basis van die lessen zelf zou meedoen met de discussies, maar die al na een dag offline moest worden gehaald omdat het een racistische chatbot bleek te zijn geworden die had ‘geleerd’ om seksistische, racistische en antisemitische taal uit te slaan.⁴¹

1.2 Literatuuranalyse ethiek en STS

Een meer algemeen probleem dat wordtesignaleerd in de bestudeerde literatuur is dat een juridisch begrip van discriminatie niet altijd aansluit bij de manier waarop ontwikkelaars van AI-systemen proberen ongelijke behandeling of uitsluiting van groepen tegen te gaan.

Wachter et al wijzen erop dat er een verschil zit tussen de bewust abstract gehouden definities van discriminatie en fairness in Europese wetgeving en de geformaliseerde statische begrippen in computer science. Volgens de auteurs kunnen deze niet op een eenduidige manier worden vertaald. Zo maken zij een onderscheid tussen discriminatie door mensen en discriminatie door systemen. Mensen discrimineren door negatieve houdingen en onbewuste bias. Dit kunnen

⁴¹ <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

signalen zijn dat discriminatie heeft plaatsgevonden. Dat soort signalen bestaan niet bij algoritmische systemen. Discriminatie door systemen is meer abstract, subtiel, niet intuïtief, ontastbaar en moeilijk te detecteren. Bestaande juridische mechanisme zijn daar niet op toegesneden. Er bestaat volgens hen ook een verschil tussen statistische interpretaties van fairness en het meer op intuïtie gestoelde concept zoals het is ingebed in de Europese juridische context.

Volgens Wachter et al zijn de definities in het recht bewust abstract gehouden om de verschillende lidstaten voldoende flexibiliteit te bieden om ze toe te passen binnen specifieke contexten. Ze noemen dit “contextual equality”. Er is dan ook niet een eenduidige richtlijn te stellen die aangeeft welke statistische instrumenten gebruikt moeten worden om discriminatie uit te sluiten: “The admissibility and relevance of statistical tests, the make-up of disadvantaged and comparator groups and the potential justifications for indirect discrimination and disparity across groups are traditionally decided on a case-by-case basis.” Ze vervolgen “Judicial interpretive flexibility’ is not a ‘bug’ of EU non-discrimination law; rather, it is both intentional and advantageous []. Contextual equality does not lend itself to automation”.⁴²

Er zijn dus verschillende definities in computer science en juridische contexten. Toch zijn het technisch, juridisch en sociale domein niet strikt van elkaar te scheiden. Bias kan in data voorkomen, maar de discriminatie die voortkomt uit het gebruik van een algoritme kan niet gereduceerd worden tot de bias in de data of het algoritme. Om te begrijpen waarom een algoritme tot discriminatie leidt, dient er verder te worden gekeken, bijvoorbeeld naar de samenstelling van de mensen die de data verzamelen, naar de keuzes die gemaakt worden bij het testen en trainen van de data, naar waarom bias in historische data voorkomt, naar de manier waarop de technologie gebruikt wordt en meer. Deze paragraaf zal op drie punten kort de problemen en knelpunten bespreken die in de literatuur worden gesignaleerd: bias (paragraaf 1.2.1), transparantie (paragraaf 1.2.2) en fairness (paragraaf 1.2.3).

1.2.1 Bias

Discriminatie wordt in de ML en AI-literatuur voornamelijk gekoppeld aan het bestaan van ‘bias’. Bias is een breed begrip dat kan refereren aan tal van specifieke problemen die zich in de *lifecycle* van een AI-gebaseerd systeem kunnen voordoen. Het is zelfs zo dat bias zich noodzakelijkerwijs voordoet bij machine learning systemen. Dit heet ook wel de “no-free-lunch” theorie.⁴³ Bij de ontwikkeling van deze systemen moeten beslissingen gemaakt worden over welk model het beste past bij de gegeven data. Elke keuze zal een bias introduceren. Het model is namelijk een abstractie van de gegeven data en elke abstractie zal bepaalde kenmerken van de data benadrukken ten koste van anderen. Om het systeem een model te laten leren zal bijvoorbeeld een keuze gemaakt moeten worden tussen het optimaal reduceren van valse negatieve of van valse positieven. Dergelijke afwegingen zullen altijd gemaakt moeten worden in het ontwerp van een ML-systeem; afwegingen tussen snelheid, of accuraatheid, tussen optimaliseren van een model op een trainingset of ruimte laten voor generalisatie over test data, etc. Deze afwegingen hebben altijd een prijs;⁴⁴ een keuze voor het ene gaat ten koste van het

⁴² S. Wachter, B. Mittelstadt & C. Russell (2020). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Available at SSRN.

⁴³ D. H. Wolpert & W. G. Macready (1997). No Free Lunch Theorems for Optimization, 1 IEEE Transactions Evolutionary Computation 67.

⁴⁴ M. Hildebrandt (2019). Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. Theoretical Inquiries in Law, 20(1), 83-121.

andere. De ene bias is echter de andere niet, en sommige zijn meer problematisch dan andere, met name als het gaat om discriminatie.

Suresh en Gutttag⁴⁵ onderscheiden bijvoorbeeld zes verschillende soorten problematische biases, die onder andere verschillen in het moment en de plek waarop zij gedurende de data life cycle ontstaan:⁴⁶

- (1) *Historische bias*: de discriminatiepraktijken die in de wereld aanwezig zijn komen zo ook in de verzamelde data terecht;
- (2) *Representatie bias*: ontstaat wanneer delen van de inputruimte onder- of overgerepresenteerd zijn;⁴⁷
- (3) *Meting bias*: kan ontstaan als de keuze van labels bevooroordeeld zijn;
- (4) *Aggregatie bias*: kan ontstaan bij het bouwen van een model, bijvoorbeeld wanneer verschillende populaties op een foutieve wijze worden gecombineerd;
- (5) *Evaluatie bias*: kan voorkomen bij het testen van het model, bijvoorbeeld wanneer de performance metrics niet geschikt zijn voor de wijze waarop het model wordt toegepast;
- (6) *Toepassing bias*: ook bij het gebruik van het model kunnen bias ontstaan, bijvoorbeeld als de resultaten niet op een goede manier worden geïnterpreteerd.

Deze zes soorten bias, met uitzondering van de laatste, zijn vooral technisch van aard en direct van toepassing op het systeem. Naast de technische biases die ontstaan in de ML-pijplijn is er ook sprake van biases die ontstaan door de menselijke en institutionele keuzes die worden gemaakt. Een aantal niet-limitatieve voorbeelden:

- (7) *Sociale bias*: Data scientists, computerwetenschappers en ontwikkelaars van modellen en applicaties hebben zelf biases.⁴⁸ Deze professionals worden voornamelijk getraind en geselecteerd op hun technische skills.⁴⁹ Veel minder aandacht is er voor de sociale, juridische en ethische kennis die nodig is om ook de maatschappelijke impact van hun werk te overzien.⁵⁰ Deze sociale bias op individueel niveau kan niet los worden gezien van de sociale bias op organisatorisch niveau.⁵¹ Het gebrek aan diversiteit in AI teams maakt bijvoorbeeld dat er minder gecorrigeerd wordt in de ontwikkeling en toepassing van AI.⁵²
- (8) *Techno bias*: Er bestaat een techno-optimisme bij de bedrijven en organisaties. Dit maakt dat niet altijd goed wordt onderzocht of de ingezette technologie wel een

⁴⁵ H. Suresh & J.V. Gutttag (2020) "A framework for understanding unintended consequences of machine learning" v.3, arXiv:1901.10002.v3 [cs.LG]

⁴⁶ Zie ook: D. Boyd, K. Levy & A. E. Marwick (2014). The networked nature of algorithmic discrimination. In Data and discrimination: Collected essays (pp. 43–57). Washington, DC: Open Technology Institute.

⁴⁷ Zie ook: L. Sweeney (2013) 'Discrimination in Online Ad Delivery', ACM Queue, 11(3), p. 10.

⁴⁸ D. Boyd & K. Crawford (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662–679. M. Fuller (2017). Big data, ethics and religion: New questions from a new science. Religions, 8(5), 88.

⁴⁹ V. Granville (2014). Developing analytic talent: Becoming a data scientist. John Wiley & Sons. A.T. Chatfield et al. (2014). Data scientists as game changers in big data environments.

⁵⁰ <https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html>; bezocht op 10 april 2020.

⁵¹ J. C. Williams (2014). "Hacking tech's diversity problem." Harvard Business Review 92.10. 94-100.

⁵² K. Crawford (2016). "Artificial Intelligences's White Guy Problem." The New York Times, June 25 <https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html>; bezocht 10 april 2020; S. U. Noble, (2013). 'Google search: Hyper-visibility as a means of rendering black women and girls invisible', InVisible Culture, 19.

adequate oplossing is voor het probleem. Soms is het zelfs niet duidelijk of er wel een probleem is.

- (9) *Bias in uitkomst*: Zelfs als het zou lukken om een niet ge-biasde dataset, algoritme en data-analyse te creëren, dan kan dit alsnog leiden tot onwenselijke uitkomsten.⁵³
- (10) *Bias in perceptie*: Er bestaat bij een deel van de beleidsmakers en techneuten de perceptie dat data-analyses neutraal zijn en data (een *datum*, een gegeven) objectief, waardoor er minder wordt getwijfeld aan en gecontroleerd op de uitkomsten (zoals besproken in paragraaf 1.1).

De verschillende soorten bias laten zien dat discriminatie het gevolg kan zijn van zowel blinde vlekken of onbewuste aannames, als van bewuste besluiten. Onbewuste bias kan zich bijvoorbeeld voordoen bij het ontwerp van een AI-systeem waarbij de ontwikkelaars uit gaan van een gemiddelde gebruiker met een bepaald kennisniveau en bepaalde capaciteiten. Het kan ertoe leiden dat mensen met een visuele beperking geen gebruik kunnen maken van het systeem. Maar, zoals eerder beschreven in sectie (X.X), data-gedreven lerende systemen kunnen ook bias bevatten als het systeem in staat wordt gesteld om correlaties te vinden die als proxy dienen voor bijvoorbeeld etniciteit, seksuele voorkeur of religie. Sweeney liet in een studie van Google Adsense zien hoe namen die vaker voorkomen bij zwarte Amerikanen eerder negatief geladen gerichte (*targeted*) advertenties genereerde dan bij namen die vaker voorkomen bij witte Amerikanen. Soortgelijke discriminatie in gerichte advertenties komt ook bij andere online platforms voor. Om advertenties meer gericht af te stemmen op specifieke internetgebruikers maken platforms, zoals Google en Facebook, gebruik van algoritmes om de voorkeuren en interesses van gebruikers af te leiden uit verschillende soorten data. Met behulp van deze algoritmes worden gebruikers ingedeeld in bepaalde ‘affiniteitsgroepen’ die als proxy konden dienen voor kenmerken zoals etniciteit, seksuele voorkeur of sekse.⁵⁴ Zo kan iemand die veel sites over feminisme bezoekt, toegekend worden aan de affiniteitsgroep ‘vrouw’. Het gebruik van ‘affiniteitsgroepen’ voor het gericht aanbieden van advertenties heeft tot gevolg gehad dat bepaalde groepen geen advertenties te zien kregen voor bepaalde financiële diensten, woningen of vacatures. Facebook heeft bijvoorbeeld meerdere keren beloofd om adverteerders niet meer de mogelijkheid te geven om advertenties aan te bieden op basis van de afgeleide etnische affiniteit van facebookgebruikers. Speicher et al. laten ook zien hoe online platforms met behulp van hun algoritmes, adverteerders in staat stellen om bewust te discrimineren op basis van kenmerken door specifieke internetgebruikers te selecteren voor advertenties. Proxies, zoals ook postcode of het type websites dat iemand bezoekt kunnen bewust of onbewust dienen om onderscheid te maken op basis van beschermde gegevens.

Er kunnen dus ook groepen ontstaan als gevolg van het gebruik van AI-systemen die mensen profileren of indelen op basis van hun kenmerken, voorkeuren of gedrag zonder dat zij zich daar bewust van zijn. Dat kan resulteren in discriminatie of, meer specifiek, discrimination by association.⁵⁵ Het is daarom van belang dat ook wordt gekeken naar de effecten van bepaalde keuzes in het ontwerp, de datasetselectie en model creatie op de populaties en de praktijk.

1.2.2 Transparantie

⁵³ A. Chouldechova (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big data* 5.2: 153-163

⁵⁴ L. Sweeney (2013). *Discrimination in online ad delivery*, ARXIV PREPR. ARXIV13016822.

⁵⁵ S. Wachter (2020). Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal*, 35(2).

Transparantie, begrepen als het toegankelijk maken van de werking en uitkomst van AI-modellen en toepassingen, is een belangrijke voorwaarde om bias en discriminatie tegen te gaan. Toch kan transparantie juist ook een oorzaak van discriminatie kan zijn.

Een voorbeeld hiervan kan gevonden worden in het werk van Axelrod, bekend om zijn onderzoek naar bereidheid tot samenwerking waarin hij vaststelde op basis van modellen dat bij herhaalde interacties actoren ertoe geneigd zijn samen te werken.⁵⁶ In een vervolgonderzoek⁵⁷ werd een nieuw kenmerk aan de actoren toegevoegd: een soort van ‘stam’-teken. Actoren konden nu kiezen om al dan niet samen te werken met mensen van hun eigen stam of van een andere stam. Het onderzoek toonde overtuigend aan dat een vorm van etnocentrisme ontstond en er veel minder werd samengewerkt met actoren van een andere stam. Hier wordt dus duidelijk dat juist transparantie, als in het kenbaar maken van een specifiek attribuut, kan leiden tot ongelijke behandeling. Dit negatieve effect van transparantie is een element van wat anti-discriminatie wetgeving juist probeert tegen te gaan.

Uitlegbaarheid (explainability) is een specifieke vorm van transparantie omdat het inzet op het begrijpelijk maken van ruwe informatie voor mensen (hoewel transparantie en uitlegbaarheid in de literatuur ook vaak als synoniem van elkaar worden gebruikt). Uitlegbaarheid kan zowel een ethisch als een operationele noodzaak hebben. De ethische noodzaak voor uitlegbaarheid is om ervoor te zorgen dat de aantasting van fundamentele rechten en andere maatschappelijke belangen geadresseerd kunnen worden. De operationele noodzaak betreft het verbeteren van de robuustheid van AI-systemen.⁵⁸

Uitlegbaarheid in brede zin gaat echter gepaard met een aantal uitdagingen. Ten eerste, bij *deep learning* – wanneer bij *machine learning* gebruik wordt gemaakt van neurale netwerken – is er vaak geen zicht meer op wat zich tussen input en output afspeelt.⁵⁹ Er bestaan initiatieven om *glassbox* ML modellen te ontwikkelen die inherent uitlegbaar zijn en er worden voor *black box* modellen post-hoc verklaringen ontwikkeld om dit tegen te gaan.⁶⁰ Instrumenten voor operationele uitlegbaarheid kunnen bovendien ook bijdragen aan onder andere het detecteren van domeinen van input data waar het algoritme zwak presteert. Deze instrumenten hebben echter ook beperkingen. Zo blijken professionals die gebruikmaken van deze instrumenten hun eigen vermogen om modellen en uitkomsten accuraat te interpreteren te overschatten.⁶¹

Een tweede probleem is dat uitlegbaarheid nog niets zegt over wat een betekenisvolle uitleg is. Met het oog op de robuustheid van het AI-systeem en als onderdeel van het ontwikkelen en verder verbeteren van het systeem, bieden de hierboven genoemde instrumenten – tot op een bepaalde hoogte – een uitleg aan technische experts. Dit houdt echter niet in dat deze technische

⁵⁶ R. Axelrod (2006). The evolution of cooperation. Basic books.

⁵⁷ R. Hammond & R. Axelrod (2006). The evolution of ethnocentrism. Journal of Conflict Resolution, 50(6):926–936.

⁵⁸ V. Beaudouin et al. (2020). Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach, Operational AI Ethics: HAL Id: hal-02506409, <https://hal.telecom-paris.fr/hal-02506409>.

⁵⁹ M. Coeckelbergh (2020) AI Ethics. Massachusetts: MIT Press, p.72. Pasquale, F., The Blackbox Society. The Secret Algorithms That Control Money and Information. 2015, Harvard: Harvard University Press.

⁶⁰ H. Suresh & J.V. Gutttag (2020). “A framework for understanding unintended consequences of machine learning” v.3, arXiv:1901.10002.v3 [cs.LG]; L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning, IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 2018, pp. 80-89.

⁶¹ H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach & J. W. Vaughan (2019). Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning. Working paper.

uitleg ook betekenisvol is voor mensen met minder of andere kennis, bijvoorbeeld zij die op de werkvloer gebruik maken van een AI-toepassing, of wanneer men op zoek is naar uitleg die juridische of ethische aspecten van het AI-systeem adresseert.

Weller oppert dan ook een opvatting van transparantie die een duidelijk verband legt tussen het doel en het publiek van de uitleg.⁶² Hij onderscheidt verschillende relaties, relevant in de AI-context, waarin uitleg de maatschappij (breed opgevat) ten goede komt:

- Voor een ontwikkelaar: om te begrijpen hoe een systeem werkt, om het te debuggen of te verbeteren, om te zien wat goed of slecht werkt, en om een idee te krijgen waarom.
- Voor een gebruiker: om een idee te geven van wat het systeem doet en waarom, om een idee te geven van wat het systeem zou kunnen doen in onvoorziene omstandigheden en om een gevoel van vertrouwen in de technologie op te bouwen.
- Om de samenleving inzicht te geven en zich comfortabel te voelen met de sterke punten en beperkingen van het systeem, waardoor een redelijke angst voor het onbekende kan worden overwonnen.
- Om een gebruiker te laten begrijpen waarom een bepaalde voorspelling of beslissing is bereikt, om te controleren of het systeem naar behoren werkte en om een zinvolle uitdaging mogelijk te maken (bijv. kredietgoedkeuring of strafrechtelijke veroordeling).
- Een expert (misschien een regulator) de mogelijkheid te bieden om een voorspelling of beslissingstraject in detail te controleren, vooral als er iets misgaat (bijvoorbeeld een ongeval met een zelfrijdende auto). Dit vereist mogelijk het opslaan van belangrijke gegevensstromen en het traceren van elke logische stap en zal de toekenning van aansprakelijkheid en wettelijke aansprakelijkheid vergemakkelijken.
- Om monitoring en testen voor veiligheidsnormen te versimpelen.

De manier waarop een uitleg wordt gegeven moet aangepast worden aan de context van de ontvanger om betekenisvol te kunnen zijn. Zo wordt er voor de niet-technische context door sociale wetenschappers geopperd om bij het ontwikkelen van XAI (*explainable AI*) meer aansluiting te zoeken bij de wijze waarop uitleg in de dagelijkse praktijk wordt gegeven, namelijk in de vorm van een conversatie. Zo kunnen er directe vragen gesteld worden, kan er nagegaan worden of iemand de uitleg begrijpt en er op specifieke onderdelen om verheldering gevraagd kan worden. Een betekenisvolle verklaring van AI zou dan (1) contrastief, (2) selectief, (3) niet gebaseerd op waarschijnlijkheden/statistische gegevens en (4) sociaal/interactief moeten zijn.⁶³

Ook Felzmann et al. verdedigen een relationele opvatting van transparantie, waarbij uitleg moet worden geconceptualiseerd als communicatie die plaatsvindt tussen technologie provider en gebruiker.⁶⁴ Wat zij daar nog aan toevoegen is een voorstel om “trustworthiness assessments” uit te voeren, gebaseerd op contextuele factoren om de waarde van die communicatie weer te geven. Technische veiligheid, operationele betrouwbaarheid en de coherentie tussen wat het

⁶² A. Weller (2017). Challenges for transparency. arXiv preprint arXiv:1708.01870.

⁶³ T. Miller (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences, arXiv:1706.07269v3.

⁶⁴ H. Felzmann, E. F. Villaronga, C. Lutz & A. Tamò-Larrieux (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1), 2053951719860542.

systeem beoogt te doen en wat het feitelijk doet kunnen daarbij dienen als bewijs van trustworthiness.

Dit alles is niet vanzelfsprekend. Het vergt van AI-ontwikkelaars verregaande contextuele kennis over hoe en door wie hun toepassingen zullen worden gebruikt en wat de verwachtingen zijn van die gebruikers. Ook vraagt het om een extra investering in bijvoorbeeld Interface en User Experience design. Een uitleg die zowel compleet als ook betekenisvol is lijkt bovendien zeer moeilijk te geven. Complete verklaringen zijn zeer moeilijk te begrijpen en begrijpelijke verklaringen kunnen te simplistisch zijn. Een trade-off tussen de twee lijkt dus onafwendbaar. Wat echter een acceptabele verklaring is en waar die zich op de curve tussen compleetheid en betekenisvolheid bevindt, laat zich niet eenduidig vaststellen.⁶⁵

Een recent uitgekomen rapport⁶⁶ heeft een raamwerk ontwikkeld wat ingezet kan worden om toch tot een betekenisvolle uitleg te komen waarbij zowel gekeken wordt naar technische, juridische als ook economische parameters. Er worden daarbij drie stappen onderscheiden:

- Ten eerste worden de **contextuele feiten** vastgelegd. Daarbij wordt gekeken naar wie het *publiek* is (wat is het expertiseniveau en hoeveel tijd is er voor een verklaring?), wat de *impact* is van de AI toepassing (over het algemeen geldt, hoe groter de impact, hoe meer baat er is van een goede uitleg), naar de *reguleringsfactoren* (welke regime is van toepassing, welke fundamentele rechten worden geraakt?) en naar de *operationele factoren* (waarvoor is de uitleg nodig? Voor veiligheid, certificeren, vertrouwen,...?)
- Ten tweede worden de **technische oplossingen** in kaart gebracht (dit betreft onder meer: het aanpassen van de ‘objective or predictor’ functie, opteren voor ‘fuzzy rules, close to natural language’, ‘output approaches’, ‘input approaches’, ‘genetic fuzzy logic’).
- Ten derde wordt in kaart gebracht wat de beste keuze is voor de **vorm van de uitleg**. Hierbij wordt een onderscheid gemaakt tussen *globale* en *lokale* uitleg. Een mogelijke vorm van een globale uitleg kan het ontwikkelen van een gebruikershandboek zijn dat het algoritme in zijn totaliteit uitlegt. Hierbij kan gevarieerd worden in de mate waarin er toegang gegeven wordt tot de broncode (afhankelijk van onder meer bedrijfsgeheimen); informatie over de trainingsdata, het algoritme, bekende biases, kwetsbaarheden en beperkingen van het algoritme. Voor het vormgeven van een lokale uitleg kan er gewerkt worden met onder meer ‘counterfactual dashboards’ waarbij het mogelijk is voor eindgebruikers om te experimenteren met “wat als” situaties; er kan voorts ingezet worden op overzichten met de belangrijkste factoren die bijdragen aan een beslissing; gelaagde uitleg instrumenten die het mogelijk maken voor de gebruiker om toegang te verkrijgen tot een steeds gedetailleerdere uitleg en toegang tot individuele beslissingslogs.

Ten slotte, een derde obstakel voor de uitlegbaarheid en transparantie van AI-systemen is dat er kosten mee gemoeid zijn, die zich op zeven punten kunnen materialiseren:

- Ontwerp -en integratiekosten (gezien de contextafhankelijkheid zullen ‘one-size-fits all’ modellen vaak niet voldoen)

⁶⁵ L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter & L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 2018, pp. 80-89.

⁶⁶ V. Beaudouin et al. (2020). Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach, Operational AI Ethics: HAL Id: hal-02506409, <https://hal.telecom-paris.fr/hal-02506409>.

- Uitlegbaarheid kan ten kosten gaan van accuraatheid waardoor een toepassing minder relevant kan worden
- Het creëren en opslaan van logs brengt operationele kosten met zich mee en kan ook spanningen opleveren met vereisten in data protectie wetgeving
- Het openbaar maken van broncode en/of andere details met betrekking tot de functionaliteit van het algoritme kan in strijd zijn met intellectueel eigendomsrecht en bedrijfsbelangen, waardoor het vaak niet opportuun is voor bedrijven om openheid van zaken te geven.⁶⁷
- Gedetailleerde uitleg van een systeem kan leiden tot “gaming the system”. Mensen gaan het systeem zo bespelen dat het de voor hen gewenste resultaten oplevert.
- Uitleg brengt impliciete regels en precedentes voort. Dit maakt dat er in de toekomst minder ruimte is voor flexibiliteit in besluitvoering.
- Het eisen van uitlegbaarheid kan de “time to market” laten oplopen waardoor innovatie wordt vertraagd.

1.2.3 Fairness

Bias in AI-systemen kan resulteren in ongelijkheid in de behandeling van bepaalde groepen of individuen. In de ML-literatuur wordt in dit verband gesproken over *unfair treatment*. De laatste jaren is er veel aandacht voor fair AI (ook wel ‘discrimination-aware data mining’ of ‘fairness-aware machine learning’ genoemd) en hoe anti-discriminatie principes en wetgeving in de modellen kan worden opgenomen.⁶⁸ Vanuit een technisch oogpunt, kan het doel van fair AI worden omschreven als: “*to translate those [legal, anti-discrimination] regulations mathematically into non-discrimination constraints, and develop predictive modeling algorithms that would be able to take into account those constraints, and at the same time be as accurate as possible*”.⁶⁹ Discriminatie in de ML literatuur wordt dan ook dikwijls begrepen als een onjuiste correlatie tussen beschermde variabelen (zoals afkomst, geslacht, geloof) en een maatstaf als uitkomst, mate van valse positieven, of een soortgelijke metric.⁷⁰

Om ervoor te zorgen dat algoritmes eerlijke (*fair*) uitkomsten opleveren, hebben technologen verschillende formele definities van *fairness* geopperd om de verdelingen van uitkomsten over groepen of individuen te kunnen beoordelen. *Fairness* kan dus op verschillende niveaus worden gedefinieerd, bijvoorbeeld als groeps gelijkheid (*group parity*) of als individuele gelijkheid (*individual parity*).⁷¹ Bij de eerste wordt gekeken of leden van verschillende groepen gelijke uitkomsten hebben en bij de tweede gaat het om een maximale nauwkeurige score voor elk individu. De verschillende *fairness* definities zijn echter niet altijd verenigbaar.⁷² Twee belangrijke uitgangspunten hierbij zijn: “(1) people that are similar in terms of non-protected

⁶⁷ Zie ook: F. Pasquale (2015). *The Blackbox Society. The Secret Algorithms That Control Money and Information*. Harvard: Harvard University Press.

⁶⁸ Voor een conceptuele analyse van het verschil tussen “fair AI” en “discrimination aware” zie: R. Binns (2018). *Fairness in machine learning: Lessons from political philosophy*. In *Conference on Fairness, Accountability and Transparency* (pp. 149-159).

⁶⁹ I. Žliobaitė. *Measuring discrimination in algorithmic decision making*. *Data Min Knowl Disc* 31, 1060–1089 (2017); p.1061 <https://doi.org/10.1007/s10618-017-0506-1>.

⁷⁰ M. Hard et al. (2016). “Equality of opportunity in supervised learning.” In *Advanced in neural information processing systems*, p. 3315-3323.

⁷¹ A. Chouldechova (2017). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.” *Big data* 5.2: 153-163

⁷² S. A. Friedler, C. Scheidegger & S. Venkatasubramanian (2016). *On the (im) possibility of fairness*. arXiv preprint arXiv:1609.07236; p.14.

characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by their non-protected characteristics.”⁷³ Het kan ook zijn dat de technische methodes om een model *fair* te maken averechts werken. Als een model wordt getraind om een recidive score toe te kennen aan verdachten en het model mag sekse niet meenemen in de score, kan dat nadelig uitpakken voor vrouwen als blijkt dat vrouwen over het algemeen minder recidiveren. Zij krijgen dan een disproportioneel hoger risico score.

De diverse definities van *fairness* in de ML-literatuur duiden op een breder probleem dat niet alleen technisch van aard is. Wat wordt gezien als een gelijke behandeling kan verschillen, niet alleen afhankelijk van de context, maar ook van de overtuiging die een persoon heeft over wat een eerlijke verdeling is. Het idee dat fairness-unfairness of gelijke-ongelijke behandeling vertaald kan worden in een wiskundige formele definitie en los gezien kan worden van een sociale, juridische en politieke context is dan ook onjuist. Er is een aantal valkuilen van een te nauwe focus op de technologie, zoals:

1. *The Portability Trap*: Het niet begrijpen hoe herbestemde algoritmische oplossingen die voor één sociale context zijn ontworpen, misplaatst zijn in een andere context;
2. *The Formalism Trap*: Het proberen te vangen van sociale concepten zoals eerlijkheid in wiskundige modellen, terwijl die concepten procedureel, contextueel en betwistbaar zijn;
3. *The Ripple Effect Trap*: Het niet begrijpen hoe de integratie van technologie in een bestaand sociaal systeem het gedrag en de ingebedde waarden van het bestaande systeem verandert.⁷⁴

Fairness in ML literatuur overlapt maar is niet gelijk aan hoe fairness in de geesteswetenschappen en het recht wordt begrepen. Zo is fairness in de ML-literatuur vooral gericht op de uitkomst, maar heeft fairness in de politieke filosofie en in het recht ook een procedurele dimensie. Het gaat hier bijvoorbeeld om hoe beslisproces is ingericht. Hebben alle betrokkenen wel voldoende inspraak en worden zij gehoord? Zijn er mogelijkheden om beslissingen te bevragen en bezwaar te maken? Xiang en Raji⁷⁵ schetsen een aantal lessen voor zowel de ML gemeenschap als het juridisch domein (met name op basis van anti-discriminatie wetgeving in de VS) over de redenen waarom deze fairness definities onderling schuren. Zo wijzen de auteurs de ML gemeenschap er onder meer op dat hun focus op uitkomsten als het gaat om het tegengaan van discriminatie en oneerlijke behandeling te nauw is. Discriminatie in bijvoorbeeld het arbeidsdomein kan juridisch ook gedefinieerd worden op basis van motief of bewijsbaar de intentie hebben om iemand uit te sluiten. Het is dus te simpel om alleen naar uitkomst te kijken. Een andere plaats waar fairness opvattingen tussen de ML gemeenschap en het recht schuren is als het gaat om wat in de ML gemeenschap “affirmative action” wordt genoemd. In een affirmative action casus, wordt er expliciet rekening gehouden met demografische gegevens om oneerlijke uitkomsten tegen te gaan. Het gebruik van dit soort gegevens is juridisch echter niet altijd evident,⁷⁶ terwijl het niet-gebruik ervan om unfaire uitkomsten tegen te gaan in de ML gemeenschap juist als naïef wordt gezien. Een mogelijke

⁷³ C. Dwork, M. Hardt, T. Pitassi, O. Reingold & R.S. Zemel (2012). Fairness through awareness. In: Proceedings of innovations in theoretical computer science, pp 214–226.

⁷⁴ A. D. Selbst et al (2019). Fairness and abstraction in sociotechnical systems. In Proc. of the Conference on FAT (pp. 59-68);

⁷⁵ A. Xiang & I. D. Raji (2019). On the legal compatibility of fairness definitions. Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada p1-6.

⁷⁶ Zie ook: M. Veale & R. Binns (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 2053951717743530.

manier om hiermee om te gaan is om een derde, betrouwbare partij in te schakelen die de sensitieve data wel bezit en kan controleren of modellen discriminerende effecten hebben. Recent onderzoek van Lahoti et al. geeft aan dat er waarschijnlijk ook technische strategieën mogelijk zijn om zonder demografische gegevens modellen te trainen en controleren; bijvoorbeeld door middel van Adversarially Reweighted Learning (ARL). ‘In particular, we hypothesize that non-protected features and task labels are valuable for identifying fairness issues, and can be used to co-train an adversarial reweighting approach for improving fairness. Our results show that ARL improves Rawlsian Max-Min fairness, with significant AUC [Area Under the Curve] improvements for worst-case protected groups in multiple datasets, outperforming state-of-the-art alternatives.’⁷⁷

Vanuit de ML gemeenschap zijn er andersom ook relevante inzichten voor het juridisch domein. Zo zou een heldere juridische basis voor het gebruik van bijzondere categorieën data om discriminatie meer effectief te kunnen meten en systemen te kunnen auditen kunnen bijdragen aan fairness. Ook zou er samengewerkt moeten worden tussen juristen en AI-specialisten om causaliteit in de context van AI beter te begrijpen. Dat dit nodig is blijkt bijvoorbeeld uit recentelijk voorgestelde wijziging onder de Fair Housing Act uit de Verenigde Staten die de lat hoger wil leggen voor klagers om aan te tonen dat er sprake is van (niet-intentionele) discriminatie als er door het algoritme geen beschermde informatie wordt gebruikt.⁷⁸ Vanuit een statistisch perspectief wil het echter niet zeggen dat wanneer er geen gebruik wordt gemaakt van beschermde variabelen of proxies (afkomst, religie, huidskleur,...) een verband ontbreekt.

De resultaten van ML algoritmes zijn heel gevoelig voor hoe een probleem is geformuleerd, welke dataset wordt gebruikt en de sociale context waarin het algoritme wordt gebruikt. Er wordt dan ook gewaarschuwd voor een te groot vertrouwen in de mogelijkheid om anti-discriminatiwetgeving in de AI context te veel te laten afhangen van de aanwezigheid van beschermde informatie of proxies.⁷⁹ ML modellen baseren zich immers op *interacties* tussen verschillende kenmerken om onverwachte patronen te ontdekken in datasets en dit kan disproportioneel minderheden raken.⁸⁰ In plaats van te focussen op specifieke vormen van achterstelling of discriminatie, is het belangrijk oog te hebben voor de praktijk waarin de systemische vormen van achterstelling en discriminatie vaak gebaseerd zijn op een combinatie van verschillende aspecten (vrouw, zwart, en gehandicapt) ook wel *intersectionaliteit* genoemd.⁸¹

Er zijn echter ook afwijkende meningen. Zo wijzen critici van de fairness-literatuur er op dat fairness een te breed en diffuus concept is, waardoor er te weinig oog is voor het probleem dat zelfs ‘fair ML’ in handen van machtige bedrijven nog steeds kan leiden tot het bestendigen en verder uitbreiden van reeds bestaande ongelijkheid en discriminatie. Zo stelt Kalluri recentelijk in Nature voor dat in plaats van door een ‘fairness’ bril er door een ‘power’ bril gekeken moeten

⁷⁷ P. Lahoti et al. (2020). Fairness without Demographics through Adversarially Reweighted Learning. arXiv preprint arXiv:2006.13114.

⁷⁸ Politico. Hud defies calls to withdraw controversial fair housing proposal. Bezocht op 15 juli 2020. <https://www.politico.com/news/2020/07/14/hud-fair-housing-proposal-discrimination-362190>

⁷⁹ A. D. Selbst (2019). “A new HUD rule would effectively encourage discrimination by algorithm”, in Slate, <https://slate.com/technology/2019/08/hud-disparate-impact-discrimination-algorithm.html>.

⁸⁰ J. Buolamwini & T. Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15.

⁸¹ A. L. Hoffmann (2019). Where fairness fails data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7), 900–915. Voor meer informatie over intersectionaliteit zie: C. D’Ignazio & L. F. Klein (2020). Data Feminism. MIT Press: Cambridge, Massachusetts.

worden naar AI, waarbij de input van minderheden en achtergestelde groepen die geraakt worden door AI-systemen centraal moet komen te staan. Meer in de richting van procedurele fairness dus. 'Researchers in AI overwhelmingly focus on providing highly accurate information to decision makers. Remarkably little research focuses on serving data subjects. What's needed are ways for these people to investigate AI, to contest it, to influence it or to even dismantle it.'⁸² Met andere woorden, als fairness in AI sowieso niet haalbaar is, is het noodzakelijk om te investeren in het contesteerbaar maken van AI.

Ook anderen hameren erop dat fairness in AI eerst en vooral een politieke vraag is en niet een primair technische; derhalve moet dit punt ook voornamelijk op politiek niveau en op politieke gronden worden geadresseerd. In het politiek filosofische domein is er tussen verschillende stromingen discussie over wat fairness is en wat het vanuit een normatief oogpunt vermag.⁸³ Het is daarbij belangrijk om te onderstrepen dat het hier dus niet gaat om een discussie over *de middelen* om fairness te bewerkstelligen maar om *de waarden* die ermee gepaard gaan.⁸⁴ Het is dus niet zo vreemd dat er in het ML-domein tal van definities in de omloop zijn, deze representeren simpelweg verschillende waarden. In dat kader lijkt het dan ook ijdele hoop dat er uiteindelijk kan worden gekomen tot een universele definitie van fairness die vervolgens in wiskundige vorm kan dienen om AI fair te maken. Zonder nauwkeurige definities van overtuigingen over de toestand van de wereld, van welke schade er voorkomen dient te worden en voor welke vorm van rechtvaardigheid wordt geopteerd, is vooruitgang op dit punt moeilijk te boeken.⁸⁵ Een meer productieve insteek is om te erkennen dat ML-algoritmen niet op een objectieve manier *fair* gemaakt kunnen worden of dat er generieke beoordelingsmaatstaven te vinden zijn om *fairness* van algoritmen te meten, maar dat *fairness* net als discriminatie context afhankelijk is en dat er dus meer voor nodig is dan technische en wiskundige definities en standaarden om discriminatie door AI systemen te voorkomen.

1.3 Technische aandachtspunten

Een eerste probleem dat in de technische literatuur wordt beschreven is het Bias-in-bias-out fenomeen. Zo bestaan er simpelweg meer gegevens over een bepaalde groepen dan over anderen. Er zijn gegevens over wie er is gearresteerd, maar niet over wie er misdaden heeft gepleegd; wie de lening heeft terugbetaald of niet, maar niet over wie de lening zou hebben terugbetaald of niet bij mensen die niet voor een lening in aanmerking kwamen. Automatic Decision Making (ADM) met behulp van traditionele ML-benaderingen repliceert niet alleen de gegevens, maar zal ook de reeds in de gegevens aanwezige vooroordelen versterken.

Een tweede probleem is een vertekening in het proces van modelontwikkeling. De gegevens en fenomenen die ze weerspiegelen kunnen vrij complex en heterogeen zijn. Verschillende subpopulaties kunnen een verschillende verdeling hebben over kenmerken en die kunnen op hun beurt weer verschillende relaties hebben met de doelvariabele/het label dat een AI-model moet voorspellen. Traditionele ML-routines zijn gericht op het minimaliseren van de gemiddelde fout ten aanzien van de meerderheidspopulaties als het minimaliseren van fouten

⁸² P. Kalluri. Don't ask if artificial intelligence is good or fair, ask how it shifts power (2020). In Nature 583,169 https://www.nature.com/articles/d41586-020-02003-2?utm_source=twit_nnc&utm_medium=social&utm_campaign=naturenews

⁸³ R. Binns (2018). Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency (pp. 149-159).

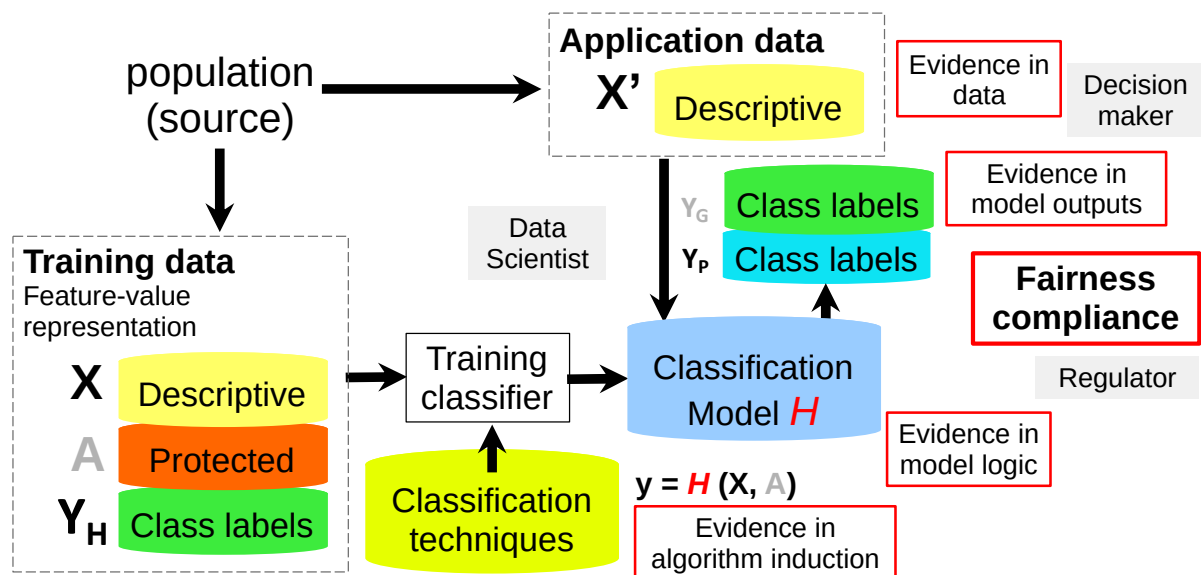
⁸⁴ P.-H. Wong (2020). "Democratizing algorithmic fairness." Philosophy & Technology 33: 228.

⁸⁵ S. A. Friedler, C. Scheidegger & S. Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236; p.14.

ten aanzien van beide populaties niet optimaal werkt. Dit leidt tot een verschillende verdeling van de fouten over de subpopulaties en de gemiddelde fout voor de minderheidspopulatie zal hoger zijn dan die voor de meerderheidspopulatie. Dat labeling fout kan gaan bleek bijvoorbeeld toen Google images bij het label gorilla's een foto van twee donkere mensen liet zien⁸⁶ en ook Amazon is gestopt met een AI-gedreven recruitment programma wegens bias naar mannen als gevolg van historische bias.⁸⁷

Een derde probleem, dat al werd aangestipt in de vorige paragraaf, dat wordt beschreven in de technische literatuur, is dat het problematisch is om fairness te formaliseren. Computerwetenschappers richten zich op generieke oplossingen, het ontwikkelen van funderingen, algoritmes en softwaresystemen die gebruikt kunnen worden voor verschillende taken en toepassingen, bijvoorbeeld algoritmes voor classificatie en andere voorspellende modelleringstaken kunnen worden toegepast voor tal van toepassingen. Om een dergelijke algemeenheid te bereiken wordt een vorm van abstractie en formalisering gebruikt. Het is dan gebruikelijk om een reductionistische benadering te kiezen voor het oplossen van reële problemen. Er zijn ten minste twee aspecten van het algoritme die hierbij relevant zijn: 1) het algoritme dat gebruikt wordt om een voorspellingsmodel te trainen en 2) het voorspellingsmodel zelf, bijvoorbeeld een convolutioneel neuraal netwerk dat gebruikt wordt om een voorspellingsmodel voor gezichtsherkenning af te leiden.

Het trainen van een AI-model kan worden beschouwd als een optimalisatieproces - gezien het doel om bijvoorbeeld nauwkeurige voorspellingen te doen voor toekomstige gevallen, wordt door biljoenen mogelijke modellen gezocht om het beste model te kiezen in termen van die geselecteerde doelstelling. Niet-uniforme nauwkeurigheid, $\text{Error}_{A=\text{males}} < \text{Error}_{A=\text{females}}$ of 'positieve' discriminatie bij het nemen van beslissingen: $P(+ | A=\text{'female'}) < P(+ | A=\text{'male'})$.



Datasets bestaan structureel uit verschillende groepen van kenmerken, die mogelijk niet allemaal toegankelijk zijn tijdens de training en het testen van het voorspellingsmodel:

⁸⁶ <https://www.bbc.com/news/technology-33347866>

⁸⁷ <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>

- $\mathbf{X}$ – predictive features (e.g. skills or job applicants or keywords appearing in tweets) used in ADM for model construction and application,
- $\mathbf{A}$ – protected features (gender, race etc) that is typically needed for fair model construction, but not used in model application; yet needed for ADM auditing,
- $\mathbf{Y_H}$ – historical labels that were possibly biased and/or noisy (hiring decisions, hate speech),
- $\mathbf{Y_G}$ – ground truth labels that are typically not available, unless we know the true data generation process as in the synthetically generated data, and
- $\mathbf{Y_P}$ – predictions generated by a model used in ADM.

Trainingsdata bestaan doorgaans uit een aantal potentieel voorspellende kenmerken, zoals vaardigheden, opleidingsniveau, ervaring uit het verleden, enzovoort. $\mathbf{A}$ is misschien helemaal niet beschikbaar, niet beschikbaar in de historische gegevens, maar wel beschikbaar tijdens de toepassing of andersom - beschikbaar tijdens de training, maar niet tijdens het testen. Als $\mathbf{A}$ niet beschikbaar is, kan het nog steeds mogelijk zijn om het af te leiden uit de proxy-informatie met een zekere mate van nauwkeurigheid. Let op: $\mathbf{Y_H}$ kan verwijzen naar mogelijke vertekening van de getrainde data als er een manier is om nauwkeurige $\mathbf{Y_G}$ voor de testset te construeren/krijgen (bijvoorbeeld in een simulatie op synthetisch gegenereerde data is het mogelijk om zowel de training als het genereren van testdata te controleren, waarbij er bewust vetekende labels in de trainingsdata worden gemaakt, en in de testset juist onbevooroordeeld). Echter, in de meeste simulatiestudies met historische data (niet synthetisch gegenereerd) is $\mathbf{Y_G}$ niet bekend en wordt cross-validatie gebruikt op dezelfde biased $\mathbf{Y_H}$.

De fairnessdefinities uit de informaticalliteratuur kan grofweg in de volgende categorieën worden samengevat:

	Parity	Preference
Treatment	Unawareness Counterfactual measures	Preferred treatment
Impact	Group fairness Individual fairness Equality of opportunity	Preferred impact

88

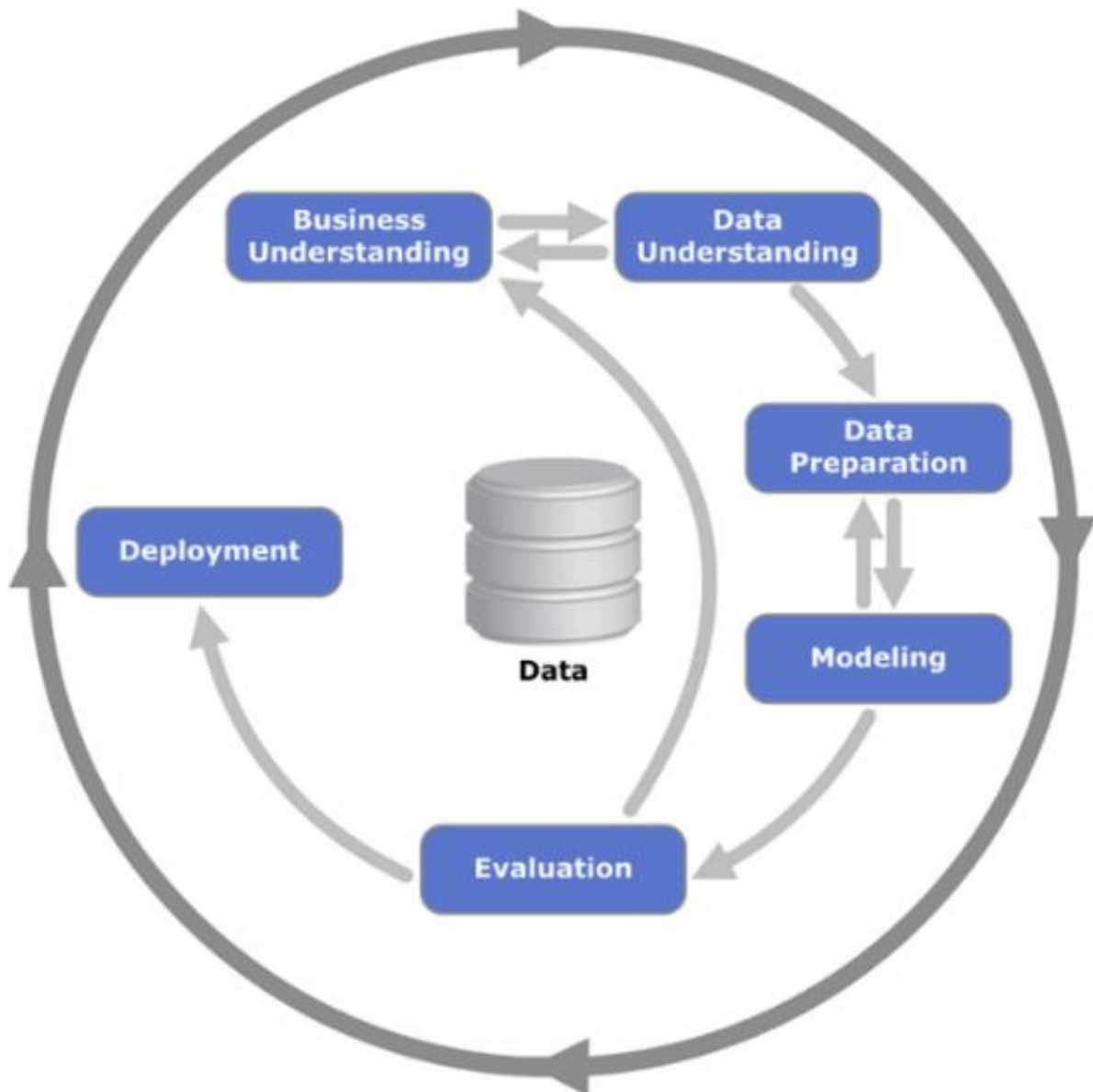
Fairness is bestudeerd in veel meer gerichte toepassingsgebieden, waaronder natuurlijke taalverwerking, computervisie, recommendersystemen en information retrieval gemeenschappen. In recommendersystemen is bijvoorbeeld het begrip "fairness" aan de aanbodzijde en aan de consumentzijde geïntroduceerd en bestudeerd. Onder meer de volgende problemen zijn daaromtrent beschreven:

- Onmogelijkheid van fairness. Verschillende statistische maatstaven van fairness zijn met elkaar in strijd, fairness op groepsniveau en op individueel niveau kan bijvoorbeeld met elkaar in strijd zijn.

⁸⁸ <https://arxiv.org/pdf/1710.03184.pdf>

- Daarbij is verklaarbare oneerlijkheid of statistische eerlijkheid niet voldoende. Verklaarbare discriminatie (Simpson paradox in de context van eerlijkheid; een waarneembare of niet-waarneembare eigenschap kan de discriminatie verklaren, bv. discipline kan de vooringenomenheid bij het nemen van aanwervingsbeslissingen verklaren; werkuren kunnen de inkomenskloof tussen mannen en vrouwen gedeeltelijk verklaren). Gegevens zijn ook niet genoeg; domeinkennis is nodig.
- De valkuil van fairness trade-offs
- Kosten van eerlijkheid

Het doen van AI-gebaseerde ADM is een proces, met inbegrip van het begrijpen van problemen en het in kaart brengen van de AI-probleemformulering. Toch ligt de focus op de goed gedefinieerde en goed geïsoleerde stap van voorspellende modelleringstaken, bijvoorbeeld classificatie. Dat kan in ieder geval worden uitgebreid met het aanpakken van fairness uitdagingen tijdens verschillende data voorbereidingsstappen, met inbegrip van bv. sampling, feature engineering, clustering en representatieleren.



89

Een vierde en laatste probleem is het dilemma van "een minderheid schaden ten gunste van de meerderheid/het algemene belang". Experimenteren is nodig om efficiënt te valideren of een bepaalde toepassing al dan niet werkt (bijvoorbeeld een medicijn) of welke behandeling het beste werkt, of welke behandeling het beste werkt voor wie. Dit vereist een vorm van gecontroleerd testen, dat wil zeggen dat alleen de werkzaamheid van een geneesmiddel bij patiënten aan wie het is toegewezen wordt bekeken en er vermoedelijk suboptimale opdrachten moeten worden uitgevoerd om de causale effecten te achterhalen (wat leidt tot een optimaler behandelbeleid) in plaats van voorspellende correlaties te maken. Terwijl het uitvoeren van dergelijke experimenten goedkoop/veilig is wanneer 'gepersonaliseerde' newsfeed aanbevelingen of advertenties worden aangeleerd, worden de kosten te hoog wanneer een beslissing met een hoge inzet wordt genomen (zet iemand die waarschijnlijk onschuldig en niet gevaarlijk is in de gevangenis om te wachten op een proef of laat iemand die potentieel gevaarlijk is weglopen; laat iemand potentieel gevaarlijke experimentele behandeling proberen of geef een placebo aan iemand die de ziekte had kunnen overleven met een experimentele

⁸⁹ https://upload.wikimedia.org/wikipedia/commons/b/b9/CRISP-DM_Process_Diagram.png

behandeling etc.). Maar het niet experimenteren op deze punten zal opportuniteitskosten veroorzaken (bij de hele bevolking of bepaalde subgroepen), waardoor de uiteindelijke kosten wellicht nog hoger zullen zijn.

1.4 Aandachtspunten binnen de beleidscontext

De beschreven risico's van AI zijn in een keur aan rapporten en beleidsdocumenten in kaart gebracht, zowel binnen de nationale als de internationale context. Zo refereert een kabinetsbrief aan discriminatiegevaaren die volgen uit: 'Bias in onderliggende data, hetgeen leidt tot discriminerende patronen; Bias in een algoritme, hetgeen leidt tot discriminerende patronen; Foutmarges die leiden tot onjuiste classificatie'.⁹⁰ Daarnaast wordt in beleidsrapporten ook met name gekeken naar de individu-overstijgende gevolgen van bias in databases, zoals ten aanzien van diversiteit, solidariteit en inclusiviteit. Daarbij zijn van belang rapporten zoals de iOverheid⁹¹ en Big Data rapporten⁹² van de Wetenschappelijke Raad voor Regeringsbeleid, de diverse rapporten van het Rathenau Instituut, zoals Beschaafde Bits,⁹³ De Datagedreven Samenleving,⁹⁴ Waardevol Innoveren,⁹⁵ Mensenrechten in het robottijdperk,⁹⁶ Regels voor het digitale mensenpark⁹⁷ en Opwaarderen: Borgen van publieke waarden in de digitale samenleving,⁹⁸ het rapport Profileren en selecteren van de Adviescommissie voor Vreemdelingenzaken (ACVZ),⁹⁹ het verkennend advies Waardevolle zorgtechnologie van de Raad voor Volksgezondheid en Samenleving¹⁰⁰ en het pas verschenen rapport van de Universiteit van Utrecht getiteld Algoritmes en grondrechten.¹⁰¹

Ook in de Tweede Kamer is er de nodige aandacht voor de kansen en risico's van de digitale samenleving, zo blijkt onder meer uit de Initiatiefnota *Onderlinge privacy* van Kamerlid Sven Koopmans.¹⁰² De regering heeft in een aantal kamerbrieven aangegeven het van belang te achten dat algoritmische besluitvorming transparant is¹⁰³ en publieke kernwaarden niet alleen waarborgt, maar ook versterkt.¹⁰⁴ Tot slot is ook op Europees en internationaal niveau de nodige aandacht voor de kansen en risico's van robotica en AI. Daarbij kan onder meer worden gewezen op het Sienna¹⁰⁵ en RoboLaw project¹⁰⁶ van de Europese Unie (EU), de ook in de startnotitie genoemde European Group on Ethics in Science and New Technologies (EGE),¹⁰⁷

⁹⁰ <https://zoek.officiëlebekeendmakingen.nl/kst-26643-642.html>

⁹¹ <https://www.wrr.nl/publicaties/rapporten/2011/03/15/ioverheid>

⁹² <https://www.wrr.nl/onderwerpen/big-data-privacy-en-veiligheid/documenten/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving>

⁹³ <https://www.rathenau.nl/sites/default/files/2018-07/Beschaafde%20Bits.pdf>

⁹⁴ https://www.rathenau.nl/sites/default/files/Rapport_Datagedreven_samenleving_Rathenau%20Instituut.pdf

⁹⁵ https://www.rathenau.nl/sites/default/files/2018-06/Rathenau%20Instituut_Waardevol_digitaliseren.pdf

⁹⁶ <https://www.rathenau.nl/nl/digitale-samenleving/mensenrechten-het-robottijdperk>

⁹⁷ https://www.rathenau.nl/sites/default/files/2018-06/Rathenau%20Instituut_Waardevol_digitaliseren.pdf

⁹⁸ https://www.rathenau.nl/sites/default/files/2018-02/Opwaarderen_FINAL.pdf

⁹⁹ <https://acvz.org/wp-content/uploads/2016/11/WEB-Profileren-en-Selecteren-.pdf>

¹⁰⁰ <https://www.raadrsv.nl/documenten/publicaties/2019/02/05/waardenvolle-zorgtechnologie>

¹⁰¹ <https://www.rijksoverheid.nl/documenten/rapporten/2018/03/01/algoritmes-en-grondrechten>

¹⁰² Kamerstukken II 2017/18, 34 926, nr. 2.

¹⁰³ Tweede Kamer, vergaderjaar 2018–2019, 26 643, nr. 570.

¹⁰⁴ Tweede Kamer, vergaderjaar 2017–2018, 26 643, nr. 537.

¹⁰⁵ <https://www.sienna-project.eu>

¹⁰⁶ <http://www.robolaw.eu/>

¹⁰⁷ Statement on Artificial Intelligence, Robotics and 'Autonomous Systems', Rapporteur: Jeroen van den Hoven, European Group on Ethics in Science and New Technologies, European Commission, DG for Research and Innovation, maart 2018, blz. 10.

de ‘Declaration for a Responsible Development of Artificial Intelligence’¹⁰⁸ en de zogenoemde ASILOMAR principles.¹⁰⁹

Het beeld dat hierin naar voren komt is dat de kansen van algoritmische besluitvorming niet alleen worden gezien in een efficiëntere publieke en private sector, waar middelen en mankracht beter en doeltreffender worden ingezet, maar ook dat organisaties beter in staat zijn hun kerntaken te verrichten. Algoritmes kunnen niet alleen worden gebruikt om historische analyses te doen of real-time analyses, maar kunnen ook probabilistische kennis over de toekomst opleveren. Zo kunnen overheidsdiensten beter voorspellen waar er problemen en knelpunten zullen optreden en daarop anticiperen; dat kan gaan om criminaliteit bij de politie en fraudegevallen bij de belastingdienst of DUO, maar het kan ook gaan om voorspellingen over wanneer welk type burger in de knel zal raken en hoe dit voorkomen kan worden.

De risico’s van de data-gedreven samenleving worden zowel gezien op het gebied van mensenrechten als op het gebied van publieke waarden. Naast het recht op privacy, de diverse vrijheidsrechten en het discriminatieverbod die onder druk kunnen komen te staan met het gebruik van automatische besluitvorming is een aantal veelgenoemde risico’s, die hier kort zullen worden besproken, omdat de specifieke problemen omtrent AI en discriminatie vaak een gevolg zijn van of samenhangen met de bredere context: effectiviteit en betrouwbaarheid (paragraaf 1.4.1), filter bubble en Mattheus effect (paragraaf 1.4.2) en legitimiteits- en vertrouwensvraagstukken (paragraaf 1.4.3).

1.4.1 Effectiviteit en betrouwbaarheid

Er zijn nog de nodige zorgen of AI, ML en DL ook daadwerkelijk leidt tot een effectievere publieke of private sector. Meer in het algemeen is niet zeker of AI en ML, waarover al decennia wordt gesproken, nu echt een succes zal worden. Ook wijzen sommigen er op dat veel van de digitaliseringsprojecten binnen de publieke en de private sector falen. Zo concludeerde de Commissie-Elias: ‘De commissie constateert dat veel problemen ontstaan aan het begin van ICTprojecten. Veel geplande projecten willen het onmogelijke voor elkaar krijgen. Projecten zijn te groot en te complex, terwijl juist deze grote projecten statistisch gezien vaker mislukken. Bovendien ontbreekt een goede zakelijke rechtvaardiging bij veel ICT-projecten. De rechtvaardiging wordt te veel gezien als een formaliteit – een manier om goedkeuring te krijgen om geld uit te geven, waarna het document in een la verdwijnt. «Businesscase, klaar is Kees», werd een gevleugelde uitdrukking bij de hoorzittingen. Het is juist belangrijk dat ook tijdens de uitvoering van een project de zakelijke rechtvaardiging van tijd tot tijd opnieuw bekeken wordt en vooral ook bijgewerkt.’¹¹⁰

Bij aanvang schetsen bedrijven die AI en data-gedreven producten ontwikkelen rooskleurige toekomstscenario’s, die later niet zelden niet uit blijken te komen. Zo heeft Google een computerprogramma ontwikkeld op basis waarvan het de uitbraak van griep kan voorspellen, onder meer door de zoektermen van gebruikers van zijn zoekmachine te analyseren op gezondheidsklachten en gerelateerde zaken. Google zelf melde al in 2009 dat zij de uitbraak van griep met maar liefst 97% zekerheid kon voorspellen. Een ongekeende accuraatheid voor voorspellend gezondheidsonderzoek. Al snel bleek echter uit een onafhankelijke studie dat er

¹⁰⁸Zie <http://nouvelles.umontreal.ca/en/article/2017/11/03/montreal-declaration-for-a-responsible-development-of-artificial-intelligence>.

¹⁰⁹ <https://futureoflife.org/ai-principles/>

¹¹⁰ https://www.tweedekamer.nl/sites/default/files/field_uploads/33326-5-Eindrapport_tcm181-239826.pdf

in het door hen geanalyseerde jaar 2013 twee keer zo veel griepgevallen werden voorspeld door Google als door de officiële instantie werd gemeten. Ook in een later artikel bleek dat meer in het algemeen, Google inaccurate voorspellingen deed met zijn Google Flu Trends service. Google's voorspellingen konden hoogstens de accuraatheid van bestaande voorspellende modellen verhogen.¹¹¹

Ook de Nederlandse belastingdienst die zo'n 5000 mensen ontsloeg en in plaats daarvan 1500 data-analisten aannam had grote verwachtingen van data-gedreven werken. Wat dat tot nu toe heeft opgeleverd, heeft de Algemene Rekenkamer berekend bij een doorlichting van delen van de belastingdienst. Interessant is dat daarbij juist de ICT projecten en de reorganisatie naar een data-gedreven wijze van opereren niet uit de verf komen. 'Uit analyse van de 42 projecten in de Investeringsagenda-projectportfolio blijkt dat 36% van de projecten de status 'vertraagd' of 'geannuleerd' heeft. Als we meer in detail kijken naar wijzigingen van planningsdata (ook van inmiddels afgeronde projecten), dan blijkt dat bijna 60% van de projecten vertraging heeft opgelopen of is geannuleerd.'¹¹² Dat het data-gedreven werken juist ook de nodige problemen met zich mee heeft gebracht moge gevoeglijk bekend zijn.

Voor dit rapport is van belang dat er onder meer door de Autoriteit Persoonsgegevens (AP) is geconcludeerd dat de risico-modellen die door de belastingdienst zijn gebruikt discriminerend waren.

Een ander voorbeeld is predictive policing, waar door de Nederlandse politie stevig in wordt geïnvesteerd. Ook hierbij was de gedachte dat als de politie maar meer data-gedreven zou opereren, het zowel efficiënter zou kunnen werken, en dus kosten zou kunnen besparen, als daadkrachtiger criminaliteit zou kunnen tegengaan. Uit onderzoek uit het buitenland waar gelijksoortige investeringen zijn gedaan was al duidelijk geworden dat er geen bewijs is dat deze techniek een significante invloed heeft op de criminaliteitscijfers, terwijl er wel veel geld en manuren in worden geïnvesteerd. Onlangs is door de Nederlandse Politieacademie een onderzoek gedaan naar de effectiviteit van de inzet van predictive policing in Nederland, en wat bleek, ook in Nederland is er geen bewijs dat de investeringen enig reëel effect sorteren. 'Aanwijzingen dat predictive policing uiteindelijk leidt tot minder (stijgende) criminaliteit hebben we niet kunnen vinden.'¹¹³

Dat ook het gebruik van AI in de strafrechtketen leidt tot discriminatoire vraagstukken is ook gevoeglijk bekend.¹¹⁴ Zo kan worden verwezen naar het COMPAS-systeem. Het systeem geeft een *risk-assessment* toepassing dat op basis van onder andere recidive data, dossiers van de verdachten en vragenlijsten ingevuld door verdachten, als output een score geeft voor de mogelijke kans op recidive. Het wordt door rechters in de V.S. gebruikt om te beslissen over de borgtocht van verdachten. Dit systeem bleek, onder andere, disproportioneel vaak en ten onrechte een hogere risicoscore toe te kennen aan verdachten met een African-American achtergrond. Een dergelijke problematische afwijking kan liggen aan de bias aanwezig in de data, maar het kan ook het resultaat zijn van de manier waarop het systeem getraind, getest en geëvalueerd wordt. De NGO die deze bias in eerste instantie aan het licht bracht, had gekeken

¹¹¹ L. Ottes (2017). Big Data in de zorg. Sdu Uitgevers.

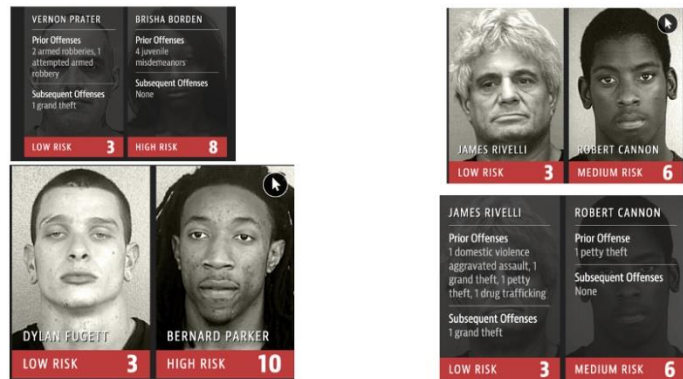
¹¹² <https://www.rekenkamer.nl/publicaties/rapporten/2017/10/11/tussenstand-investeringsagenda-belastingdienst>

¹¹³ <https://www.politieacademie.nl/kennisenonderzoek/kennis/mediatheek/PDF/93263.PDF>

¹¹⁴ J. Angwin, J. Larson, S. Mattu & L. Kirchner (2016). Machine Bias.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

naar de verhoudingen tussen de valse positieven en negatieve tussen twee groepen, maar het bedrijf achter het COMPAS-systeem gebruikte een ander *fairness* criterium, namelijk de kans dat een verdachte die wordt aangemerkt als mogelijke recidivist ook daadwerkelijk recidiveert gelijk is voor de verschillende groepen. Volgens het bedrijf voldeed het systeem aan dit criterium. Dit is dus een voorbeeld van hoe verschillende gehanteerde definities van *fairness* kunnen leiden tot verschillende uitkomsten en conclusies.



Het gebrek aan effectiviteit en de nodige fout(marges) zijn ten dele te verklaren door technische beperkingen en inherente en onvermijdelijke bias. Een ander deel is echter te verklaren door een gebrek aan standaarden voor data-analyse. De WRR wijst daarbij in haar Big Data rapport op de statistische beperkingen. Ze verwijst daarbij naar Anderson die in zijn veelbesproken artikel ‘The End of Theory: The Data Deluge Makes the Scientific Method Obsolete’ stelt dat het einde van theorie en de wetenschappelijke methode in zicht is door de data-gedreven ontwikkelingen.¹¹⁵ De Raad merkt echter op dat hier veel op valt af te dingen. Algoritmes vertellen niet uit zichzelf wat de data betekenen of hoe ze te interpreteren zijn. De onderzoeksvraag, de datasetselectie en de interpretatie van de uitkomsten bevatten onvermijdelijk vele theoretische vooronderstellingen. Ook wijst de Raad op het gevaar voor verwarring tussen correlatie en causaliteit en op het feit dat zonder theorie over het waarom van correlaties en zonder goede aanwijzingen dat deze causaal zijn, daarop gebaseerde interventies de plank mis kunnen slaan. Tot slot onderstreept de Raad op het zogenoemde multiple-comparisons problem, wat wil zeggen dat bij het analyseren van grote hoeveelheden data altijd wel correlaties zijn te vinden (data dredging of fishing expedition). Er kunnen (sterke) correlaties tussen twee variabelen worden gevonden die niets met elkaar te maken hebben, zoals het ‘verband’ tussen hoeveel kaas mensen eten en hoeveel mensen overlijden doordat ze vast komen te zitten in hun dekbed of de correlatie tussen films waarin Nicolas Cage speelt en het aantal mensen dat verdrinkt in een zwembad.¹¹⁶

De Raad stelt dan ook in haar aanbevelingen dat: ‘De algoritmes en methoden die bij data-analyses worden gebruikt moeten deugdelijk zijn en aan de wetenschappelijke criteria voor goed (statistisch) onderzoek voldoen.’¹¹⁷ Recent onderzoek uitgevoerd voor het WODC concludeert in dezelfde lijn dat de overheid er goed aan zou doen meer en betere standaarden vast te stellen ten aanzien van statistische analyses. ‘Het gaat hierbij om regelgeving die niet direct aan persoonlijke belangen is gerelateerd. Of een groepsprofiel of statistische correlatie met een bepaald waarschijnlijkheidspercentage als zodanig incorrect of discriminatoir is raakt immers als zodanig niemand; dat geschiedt pas als die worden gebruikt voor toepassingen in de praktijk. Toch kan het wenselijk zijn om reeds hier al minimale standaarden neer te leggen, omdat hiermee eventuele problemen in de toekomst kunnen worden voorkomen en omdat

¹¹⁵ C. Anderson (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired magazine, 16(7), 16-07.

¹¹⁶ L. Kool, J. Timmer & R. van Est (2015) De datagedreven samenleving, Den Haag: Rathenau Instituut

¹¹⁷ WRR, ‘Big Data in een vrije en veilige samenleving’, WRR-rapport, Amsterdam University Press, Amsterdam 2016.

deugdelijke statistische analyse zowel in het belang is van de burger als de efficiëntie van de Big Data partij ten goede komt. In deze fase worden immers veel van de wezenlijke keuzes gemaakt die later impact hebben op de burger.’¹¹⁸

1.4.2 Filter bubble en Mattheus effect

Een tweede gevaar waar op wordt gewezen hangt samen met de inzet van profiling. Een probleem dat hieraan is gekoppeld is dat oude patronen bepalend worden voor de mogelijkheden en kansen in de toekomst. Dit wordt ook wel eens het gevaar van de echokamer of de filter-bubble genoemd. ‘Personalisation of the web, with customised media and news services based on the individual's web behaviour, will also have an impact on the framework conditions for public debates and exchanges of ideas – important premises for a healthy democracy. This is not primarily a privacy challenge, but constitutes a challenge for society at large. The danger associated with so-called “echo chambers” or “filter bubbles” is that the population will only be exposed to content which confirms their own attitudes and values. The exchange of ideas and viewpoints may be curbed when individuals are more rarely exposed to viewpoints different from their own.’¹¹⁹ Een voorbeeld dat vaak wordt aangehaald als illustratie van dit fenomeen is dat de content die mensen te zien krijgen op het internet steeds meer geschiedt op basis van profielen. Een rechts-conservatieve-laagopgeleide man krijgt andere zoekresultaten, reclames, nieuws en video’s te zien dan een persoon die in het profiel van link-progressieve-hoogopgeleide vrouw valt. Dit zou kunnen leiden tot een moderne vorm van verzuiling, waarbij mensen steeds meer bevestigd worden in hun eigen wereldbeeld en groepen steeds meer uit elkaar groeien. Sommige experts leggen hier een verband met de polarisatie in onder meer de Verenigde Staten en het Verenigd Koninkrijk.¹²⁰

Ook de bestaande sociale ongelijkheid in de samenleving wordt door de algoritmes herhaald en daarmee versterkt. Dit wordt in de literatuur ook wel het Mattheus effect genoemd, oftewel de armen worden armer en de rijken worden rijker.¹²¹ Eerder is al het voorbeeld gegeven van het feit dat door in bepaalde wijken te surveilleren, er meer incidenten uit die wijk zullen worden geregistreerd, wat weer met zich mee kan brengen dat de data-analyse aangeeft dat er met name in die wijk meer moet worden gesurveilleerd. Hetzelfde zelfversterkende effect kan optreden bij bijvoorbeeld verzekeringen en banken. Als verzekeraars en banken op basis van data-gedreven processen beslissingen nemen over wie een verzekering of lening aan te bieden en tegen welke voorwaarden, dan zullen de personen uit de rijkere wijken leningen en verzekeringen krijgen tegen betere voorwaarden, terwijl de personen uit armere wijken hogere premies moeten betalen, omdat uit de data-gedreven analyse blijkt dat de bedrijven met de personen in die wijken een hoger risico lopen. De rijkere zijn daarmee nog beter af, terwijl de armeren er bekaaid afkomen.

¹¹⁸ B. van der Sloot & S. van Schendel, 'De Modernisering van het Nederlands Procesrecht in het licht van Big Data: Procedurele waarborgen en een goede toegang tot het recht als randvoorwaarden voor een data-gedreven samenleving', WODC 2019.

¹¹⁹ https://www.datenschutz-berlin.de/fileadmin/user_upload/pdf/publikationen/working-paper/2014/06052014_en.pdf

¹²⁰ <https://www.theguardian.com/media/2017/jan/08/eli-pariser-activist-whose-filter-bubble-warnings-presaged-trump-and-brexit>

¹²¹ A. van de Rijt (2017). Zelfversterkende processen in de samenleving. *Mens en maatschappij*, 92(3), 289-309. E. Snel, G. Custers, & G. Engbersen (2018). Ongelijkheid in de Participatiestad. *Mens en maatschappij*, 93(1), 31-57.

Daarom wordt er gewezen op het gevaar dat Big Data de solidariteit in de samenleving zal ondermijnen.¹²² Immers, steeds beter kan worden berekend welk gedrag leidt tot welk soort risico; dat gaat dan niet alleen om de gezondheidseffecten van roken of de kans op botbreuken bij wintersport; door middel van apps, fitbits en allerlei apparaten in de smart home kan in feite het hele leven van mensen in kaart worden gebracht. Zo kan potentieel aan ieder gedrag consequenties worden gekoppeld. Het idee dat iedereen voor elkaar betaalt – de kinderloze betaalt voor kinderopvang, de niet roker voor de medische behandeling van de roker, de man voor het zwangerschapsverlof van de vrouw, etc. – komt daarmee potentieel onder druk te staan.

Daarnaast heeft de Wetenschappelijke Raad voor Regeringsbeleid in haar rapport *Weten is nog geen doen* de nadruk gelegd op het doen-vermogen van mensen. De Raad vraagt aandacht voor het feit dat er steeds meer wordt verlangd van mensen in termen van autonomie en zelfredzaamheid. Hierdoor ervaren veel mensen druk en dreigt het gevaar dat met name mensen in een toch al zwakkere positie nog meer achter zullen blijven. De Raad wijst op een rationaliseringstendens binnen bestuur en beleid en zet dat af tegen de realiteit: de mens is geen rationele beslisboom.

‘Het klassieke beleidsperspectief van de overheid gaat uit van een burger bij wie de juiste kennis automatisch leidt tot de juiste daden. Het is het *rationalistische* perspectief dat aan de basis ligt van veel juridische en economische benaderingen van beleid. Daarnaast zetten wij een *realistisch* perspectief dat een vertaling is van de gedragswetenschappelijke bevindingen van dit rapport. Dat perspectief gaat ervan uit dat mensen niet altijd handelen, ondanks hun goede voornemens. Weten leidt niet altijd tot doen. Bovendien kan een keuze die op de lange termijn ‘onverstandig’ is, op de korte termijn gezien de situatie van het individu wel degelijk de ‘verstandige’ optie zijn. Het uitgangspunt van beide perspectieven is hetzelfde, te weten eigen verantwoordelijkheid, en het doel is eveneens hetzelfde: autonomie en redzaamheid voor iedere burger. De twee perspectieven verschillen echter in hun aannames over de mentale vermogens van burgers en over psychologische wetmatigheden. Omdat het eerste perspectief uitgaat van te hoge verwachtingen van het doenvermogen van burgers brengt dit het doel van redzaamheid niet altijd dichterbij.’¹²³ In zijn reactie op dit rapport heeft het kabinet onder meer aangegeven te willen onderzoeken in hoeverre AI een oplossing kan bieden voor dit aandachtspunt. Kan AI bijvoorbeeld worden ingezet voor een tijdige signalering van mensen die door omstandigheden (tijdelijk) niet zelfredzaam zijn of voor een bepaling wanneer de standaardbenadering plaats moet maken voor maatwerk? Hoe moet het juridische systeem hierop anticiperen?¹²⁴

Ook wordt een gevaar in de literatuur gesignaleerd dat organisaties die AI gebruiken voor beleid- en besluitvormingsprocessen voor burgers vaak steeds ongrijpbaarder worden. Voor de gemiddelde burger is het lastig om te doorgronden welke organisaties er allemaal gegevens over hem verzamelen, waarom die gegevens worden verwerkt en voor welke doeleinden ze vervolgens worden gebruikt. Bij AI wordt dit probleem versterkt. Een algoritme is voor burgers een *black box*. Data expert Frank Pasquale stelt dat dit een kenmerkend aspect is van onze hedendaagse samenleving, die hij dan ook de *Black Box Society* noemt.¹²⁵ Deze ‘transparantie

¹²² S. Klous, ‘Wij zijn Big Data’, oratie uitgesproken op 26 november om 16:00 in de aula van de Universiteit van Amsterdam.

¹²³ WRR, ‘Weten is nog geen doen: een realistisch perspectief op redzaamheid’, Den Haag 2017, p. 12.

¹²⁴ Zie o.a. <https://www.uu.nl/sites/default/files/tk-bijlage-onderzoeksrapport-juridische-aspecten-van-algoritmen-die-besluiten-nemen.pdf>

¹²⁵ <http://raley.english.ucsb.edu/wp-content/Engl800/Pasquale-blackbox.pdf>

paradox'¹²⁶ wordt extra problematisch als burgers zich willen verzetten tegen de AI-besluitvormingsprocessen die op hen van toepassing zijn. Ze zijn zich vaak niet bewust dat ze überhaupt onderdeel zijn van een AI-proces, als ze dat wel zijn dan weten ze vaak niet wie achter het proces zit of wil het desbetreffende overheidsorgaan of bedrijf geen openheid van zaken geven over hoe het algoritme werkt.¹²⁷

1.4.3 Legitimiteit en vertrouwen

Als laatste wordt er nog een aantal punten in de literatuur en de beleidscontext naar voren gebracht omtrent de legitimiteit van organisaties die AI gebruiken voor beleids- en besluitvorming en het vertrouwen van burgers in deze organisaties. Het eerste punt dat hiermee samenhangt is wat de WRR heeft genoemd de 'ontschotting' van de samenleving. Aangezien alles in feite iets zeggen over al het andere (de kleur van iemands bank kan iets zeggen over zijn karakter, iemands muzieksmaak kan een voorspeldende waarde hebben voor zijn politieke voorkeur, de naam van iemands kat kan iets zeggen over zijn gezondheid) worden schotten en contexten die normaal wezenlijke zijn in sociale relaties en maatschappelijke processen doorbroken.¹²⁸ Omdat er in data-analyses traditioneel werd gekeken naar causale verbanden waren de data vaak gerelateerd aan datgene waarover iets werd gezegd; bij statische correlaties hoeft dat echter geenszins het geval te zijn.

Het feit dat organisaties deze contextualiteit niet meenemen, maar data als zodanig bekijken wordt ook wel eens vergeleken met een vorm van autisme, wat door experts wordt gedefinieerd als een vorm van context-blindheid.¹²⁹ Een decontextualisering en verabsolutering van regels kan worden gestimuleerd door data-gedreven werken. Niet alleen worden zeer grote hoeveelheden data verzameld zonder hoofd- en bijzaken te onderscheiden, ook worden de verkregen inzichten niet zelden veralgemeniseerd en verabsoluteerd en worden regels en patronen op een niet-contextspecifieke wijze geïnterpreteerd. Het kan afbreuk doen aan de steun voor beslissingen, bijvoorbeeld als feiten die burgers zien als ongerelateerd aan het punt waarover een beslissing wordt genomen (iemands muzieksmaak wordt meegenomen bij de vraag of een bank hem een lening vertrekt) toch een rol spelen.¹³⁰

Daarbij komt dat het menselijk oordeel steeds meer door computers en machines kan worden overgenomen en vervangen. Ook het simpele feit dat niet een mens, maar een computer een beslissing neemt kan als zodanig een legitimiteits- en vertrouwensvraag oproepen.¹³¹ Daarnaast is één van de problemen dat organisaties die AI inzetten voor beleids- en besluitvormingsprocessen niet altijd zelf de data hebben verzameld waarvan zij gebruik maken. Dat betekent dat zij niet altijd kunnen verantwoorden hoe data is verzameld en waarom; dit geldt zeker als data worden geschraapt van het internet, maar ook als databases worden gekoppeld of opgekocht. Interessant is daarbij dat het rapport Waarde(n)volle Zorg daarbij ook wijst op het belang van data-continuïteit. 'Bedrijven kunnen failliet gaan. In dat geval moeten

¹²⁶ N. M. Richards & J. King (2014). 'Big Data ethics', Wake Forest Law Review, 49: 393-432.

¹²⁷ <https://www.technologyreview.com/s/608011/secret-algorithms-threaten-the-rule-of-law/>
<https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/>

¹²⁸ Nissenbaum, H. (2009). Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

¹²⁹ P. Vermeulen, 'Autisme als contextblindheid', in: Epo 2009.

¹³⁰ WRR, 'Big Data in een vrije en veilige samenleving', WRR-rapport, Amsterdam University Press, Amsterdam 2016.

¹³¹ <https://www.thebureauinvestigates.com/stories/2017-01-17/obamas-covert-drone-war-in-numbers-ten-times-more-strikes-than-bush>

patiënten over hun gegevens kunnen blijven beschikken. Een verplichte (gratis) downloadmogelijkheid voor patiënten van hun gegevens in zorgsystemen – eventueel gecombineerd met een van overheidswege opgelegde ‘health data escrow’ of een erkenning voor bedrijven die een dergelijke escrow-regeling hebben waarmee patiënten kunnen zien dat zij altijd hun gegevens terug kunnen krijgen – kan dit garanderen en de continuïteit van zorg waarborgen.’¹³²

Daarnaast ontbreekt het burgers vaak aan een goed klachtrecht: als zij al kunnen klagen dan gaat het vaak om hun particuliere situatie en niet om het profiel, het algoritme of de dataset als zodanig, terwijl daar de oorsprong van eventuele discriminatie is te vinden. Toezicht op algoritmische besluitvorming is (zeer) beperkt, zo bleek ook uit de antwoorden van respondenten als besproken in paragraaf 0.3. Het Rathenau Instituut concludeert in zijn rapport Opwaarderen dat het meer in het algemeen ontbreekt aan een goede governance structuur voor digitalisering. ‘Onze analyse van het governancelandschap laat zien dat de bescherming van publieke waarden op dit moment sterk tekortschiet. Digitalisering leidt tot conceptuele verwarring over welke regels gelden en over hoe deze moeten worden toegepast. In sommige gevallen zien we dat bestaande kaders daardoor (aanvankelijk) niet worden toegepast terwijl deze wel van toepassing zijn. In andere gevallen zien we dat digitalisering vraagt om nieuwe kaders over hoe belangrijke waarden worden geborgd in de digitale samenleving. Op verschillende onderdelen van het landschap zien we dat er blinde vlekken zijn ontstaan:

- Vertaling van nieuwe maatschappelijke en ethische vraagstukken naar beleid, interdepartementale afstemming en coördinatie op digitalisering, en het politieke debat over die nieuwe vraagstukken.
- Borging van grond- en mensenrechten in de digitale samenleving.
- Versterken van toezichthouders en zorgen voor onderlinge afstemming tussen toezichthouders.
- Nieuwe verantwoordelijkheden van ontwikkelaars van digitale diensten en producten.
- Organiseren van maatschappelijk tegengeluid: versterken maatschappelijk middenveld, kennis en vaardigheden van burgers en maatschappelijk debat over digitalisering.’¹³³

Tot slot laten interne audits niet zelden te wensen over. Zo concludeerde de Adviescommissie voor Vreemdelingenzaken dat maar enkele van de onderzochte instanties het vier-ogen principe hanteerde en dat interne audits, bijvoorbeeld in de vorm van steekproeven, nog niet door alle organisaties werden toegepast. ‘Door het ontbreken van aselechte steekproeven kunnen profielen niet goed bijgesteld worden en treedt een zelfversterkend effect op van het beeld. De meeste instanties voeren geen aselechte steekproeven ter controle op de profielen uit. [Twee instanties] gebruiken het criterium risicowijk in hun profielen. Die wijken zijn op grond van een aantal sociaal-economische indicatoren geselecteerd, met een kleine correctie om ook een landelijke spreiding van wijken te genereren. Door het ontbreken van een aselechte steekproef treedt een zelfversterkend effect op van het beeld van risicowijken ten opzichte van wijken die niet als risicowijk bestempeld zijn. In het kader van profilering met Advance Passenger Information en Mobiel Toezicht Veiligheid worden ook geen aselechte controles uitgevoerd. Bij de Inspectie SZW [Sociale Zaken en Werkgelegenheid] krijgen inspecteurs op basis van een profiel een aantal adressen aangeleverd waaruit zij weer een keuze maken op grond van hetgeen zij ter plekke waarnemen en op grond van kennis en ervaring. Ook bij Inspectie SZW worden geen

¹³² https://www.raadrvs.nl/binaries/raadrvs/documenten/publicaties/2019/02/05/waardenvolle-zorgtechnologie/RVS_advies_Waarde%28n%29volle_zorgtechnologie_online+versie.pdf

¹³³ https://www.rathenau.nl/sites/default/files/2018-02/Opwaarderen_FINAL.pdf

aselecte controles gehouden. Wel heeft de inspectie een intelligence traject dat als doel heeft zicht te krijgen op blinde vlekken. De Belastingdienst voert wel ieder jaar standaard een aselecte controle uit om zo haar profielen te kunnen bijstellen. Zij ervaart dit als nuttig en noodzakelijk.¹³⁴

134

https://www.adviescommissievoorvreemdelingenzaken.nl/binaries/adviescommissievoorvreemdelingenzaken/documenten/publicaties/2016/11/08/profileren-en-selecteren/Profileren_en_Selecteren_ACVZ_beleidsadvies_20161108.pdf

Casus: Gepersonifieerde online prijsdiscriminatie gezien vanuit het gelijke behandelingsrecht

Van prijsdifferentiatie in de brede, economische zin is sprake waar hetzelfde product of dezelfde dienst aan verschillende personen tegen verschillende prijzen wordt verkocht, voor redenen die niet herleidbaar zijn tot (een verschil in) kosten¹ of resterende voorraad. Voor deze definitie is niet relevant of de differentiatie offline of online plaatsvindt.

De situatie waarover advies wordt gevraagd ziet op een vorm van prijsdifferentiatie die tot stand komt door de (1) toepassing van algoritmes op (2) (big) data om zo (3) sneller, gericht (rekening houdend met individuele kenmerken) en aan meer personen een individueel prijsaanbod te kunnen doen. Dit kan overigens zowel analoog als digitaal tot uiting te komen: immers kan met de persoonlijke bonuskaartaanbiedingen de prijs voor een pot pindakaas in de winkel voor twee klanten van de AH ook verschillend zijn, en kunnen de bordjes met digitale prijzen in winkels makkelijk variëren naargelang de tijd van de dag.¹

Prijsdifferentiatie is onderdeel van de reguliere marktwerking. Een onderneming zal immers proberen een product aan te bieden die aansluit bij de vraag en daarbij streven naar winstmaximalisatie. Consumenten zullen bij het consumeren van producten zo goed mogelijk in hun behoeftes proberen te voorzien en daarbij streven naar nutmaximalisatie. Gegeven verschillen in behoeftes en budget zullen verschillende consumenten bereid zijn verschillende prijzen voor hetzelfde product te betalen: de betalingsbereidheid. Een onderneming kan hierop proberen in te spelen door meer van de klant te weten te komen. Een consument kan veelal tussen verschillende aanbieders kiezen en diegene kiezen die het product voor de laagste prijs kan leveren.

Gebruik van algoritmes en big data heeft tot gevolg dat ondernemingen steeds beter kunnen inschatten wat de maximale prijs is die een individuele consument voor een product wenst te betalen. Dit kan leiden tot een disbalans in die zin dat de *consumer surplus* (het verschil tussen de prijs die een persoon betaalt en de prijs die hij of zij maximaal had willen betalen) steeds kleiner wordt. De *producer surplus* (het verschil tussen de prijs die een onderneming voor een product krijgt en de laagste prijs die de

ondernemer zou hebben geaccepteerd) wordt steeds groter.

In de volgende secties wordt uitgelegd onder welke voorwaarden en onder welke omstandigheden deze disbalans problematisch kan zijn vanuit het oogpunt van het juridische bindende gelijke behandelingsrecht dat – net als veel wetgeving m.b.t. de markt – grotendeels van Europese origine is.

Toetsingskader gelijke behandelingsrecht

Het toetsingskader gelijke behandelingsrecht valt uiteen in grofweg vier elementen: het bepalen of een situatie onder de reikwijdte valt, de kwalificatie van het onderscheid, de toepasselijkheid van uitzonderingen of een rechtvaardigingsgrond en een proportionaliteitanalyse.

Qua reikwijdte verbiedt het gelijke behandelingsrecht het maken van onderscheid op bepaalde terreinen (zogenoemde ‘beschermd terreinen’) op basis van de in de wet genoemde gronden (zogenoemde ‘beschermd gronden’). Online prijsdifferentiatie bij het aanbieden van goederen en diensten valt onder de beschermd terreinen,¹ wat betekent dat geen onderscheid gemaakt mag worden op grond van burgerlijke staat, geslacht, godsdienst, handicap/chronische ziekte, hetero- of homoseksuele gerichtheid, levensovertuiging, nationaliteit, politieke gezindheid en ras/ethniciteit.¹

De volgende stap is het bepalen van het type onderscheid. Van direct onderscheid is sprake als een persoon op grond van een beschermd grond wordt of zou worden benadeeld ten opzichte van een ander in een vergelijkbare situatie.

Anne is woonwagenbewoonster en wil graag online kleding bestellen. Omdat zij op een woonwagenlocatie woont zijn echter slechts een beperkt aantal diensten beschikbaar (alleen vooraf betalen, of ophalen bij bezorgdepot). Personen die niet op een woonwagenlocatie wonen lopen niet tegen deze beperkingen aan. De verkoper maakt hier direct onderscheid op grond van ethniciteit.¹

Van indirect onderscheid is sprake als een ogenschijnlijk neutrale bepaling, maatstaf of handelwijze een groep personen gekenmerkt door een beschermd grond in het bijzonder treft in vergelijking met een andere groep personen die deze kenmerken niet heeft.

Jan – een Nederlander – zoekt een woning. Een makelaar biedt een appartement online aan en schrijft daarbij dat deze alleen voor ‘expats’ beschikbaar is. Naar gangbaar gebruik is een expat iemand die tijdelijk in het buitenland werkt. Een expat die een woonruimte in Nederland zoekt zal veelal een buitenlandse nationaliteit hebben. Het door de makelaar gehanteerde

criterium werkt daarmee voornamelijk ten nadele van personen met de Nederlandse nationaliteit.¹

Niet elk onderscheid is echter per definitie verboden. Voor direct onderscheid kan er sprake zijn van een in de wet opgenomen uitzondering. Voor indirect onderscheid geldt de ruimere mogelijkheid van objectieve rechtvaardiging: het hebben van een goede reden om het onderscheid te maken. Vervolgens wordt getoetst of het aangewende middel om dit doel te bereiken geschikt en noodzakelijk is.

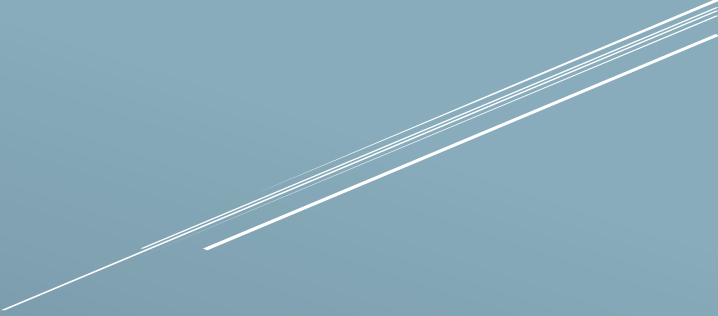
Risico's in de sfeer van het gelijke behandelingsrecht.

Voorgaande betekent dat een volledig op het individu toegesneden prijs die voor persoon A hoger is dan B als gevolg van een (hypothetisch) neutraal algoritme vanuit het perspectief van het gelijke behandelingsrecht niet altijd problematisch hoeft te zijn. Immers ziet dat nu juist enkel op situaties waar mensen direct of indirect negatieve gevolgen verbinden aan bepaalde door de wet beschermde persoonskenmerken van een ander.

Indien het verschil in prijs niet herleidbaar is tot een van de genoemde beschermde gronden is er geen (voor de gelijke behandelingswetgeving relevant) discriminatieprobleem. Wordt in deze neutrale setting de verschuiving in het surplus-equilibrium onwenselijk geacht, dan kan de marktbalans gepoogd te worden hersteld via mededingingsrecht, consumentenrecht en gegevensbescherming. Bijvoorbeeld door consumenten meer controle over hun data (en dus de dataverzameling) te geven en concurrentiebeperkend algoritmegebruik of misbruik van marktmacht¹ tegen de gaan.

Prijsdifferentiatie die ontstaat door inzet van algoritmes is echter niet noodzakelijkerwijs neutraal.¹ Bij datagedreven prijs- en aanbodbepaling bestaat het risico dat zowel bij het *design* van het algoritme direct of indirect wordt aangesloten bij beschermde gronden, de gebruikte *data* inherente vooroordelen (biases) bevat en als gevolg van de digital divide (tussen gebruikers met en zonder toegang tot de online wereld) de op de *gebruikers* toegesneden prijs- en dienstverlening bepaalde groepen systematisch marginaliseert:¹

1. Design. De schrijvers van een algoritme maken (on)bewuste keuzes ten aanzien van het doel van het algoritme, en schrijven aan de hand daarvan een sorteringsmechanisme dat data indeelt in categorieën die zij denken dat bij het gedefinieerde doel behoren (middel).¹

- 
- a. **Voorbeeld.** Doel is kredietwaardigheid vast te stellen. Middel is dat personen die tenminste 5 jaar een voltijdbaan hadden voorafgaand aan de aanvraag als kredietwaardig moeten worden beschouwd. Mogelijk discriminatoir gevolg: vrouwen – zijnde vaker parttimer - worden disproportioneel vaak hogere premies voorgeschoteld. Een dergelijk design zou een onrechtvaardigbaar indirect onderscheid kunnen bestendigen.
 2. Data. Algoritmes worden ‘getraind’ op een dataset en kunnen zo daarin bestaande *biases* (vooroordelen) reproduceren of zelfs verder verdiepen: garbage in, garbage out.¹
 - a. **Voorbeeld.** Indien de gebruikte data (on)bewust discriminatoir tot stand is gekomen, bijvoorbeeld door gebruik van direct of indirecte gegevens over ras in onderzoek naar fietsdiefstallen, dan zal een analyse van deze data voor de bepaling van de premie voor fietsverzekeringen voor personen met een bepaalde etnische afkomst slechter uitvallen.
 3. Gebruikers. De digital divide houdt in dat sommige gebruikers meer vaardigheden en betere toegang hebben tot de onlinewereld dan anderen,¹ zoals ouderen, personen met een beperking en lager opgeleiden.¹ Dit heeft gevolgen voor 1. het type data dat wordt verzameld en 2. hoe kwetsbaar personen wel of niet zijn voor (de gevolgen van) (prijs)discriminatie: Hoe meer de dienstverlening wordt ingericht op basis van (een automatische) analyse van de grote hoeveelheid data die internetgebruikers produceren, hoe beter deze zal aansluiten op de (wensen en behoeftes van de) gemiddelde internetgebruiker. Dit kan echter tot gevolg hebben dat diegene die afwijken van de gemiddelde norm worden gemarginaliseerd terwijl datzelfde gebrek aan toegang en/of vaardigheden betekent dat zij zich moeilijker kunnen wenden tot (digitale) alternatieven.
 - a. **Voorbeeld.** Een pakketdienst specialiseert zich in het bezorgen van online aankopen en heeft zijn dienstverlening volledig digitaal ingericht. Indien de ontvanger van het pakket niet thuis is wordt een briefje achtergelaten met een code die in de app van het bedrijf ingevoerd kan worden om de locatie van het pakket (e.g. bureu, depot) na te gaan. Dit sluit aan bij de wensen van veel klanten, maar zal voor sommige ontvangers problematisch zijn, bijvoorbeeld omdat zij een visuele beperking hebben of gebrekkige toegang tot het internet.

De oordelenlijnen van het College voor de Rechten van de Mens vis-a-vis (online) prijsbepaling: wanneer wordt differentiatie discriminatie?

Waar algoritmes leiden tot onderscheid in online prijsbepalingen herleidbaar tot een beschermde grond dan is de gelijke behandelingswetgeving daarop volledig en onverkort van toepassing. Het College voor de Rechten van de Mens (en daarvoor de Commissie Gelijke Behandeling) heeft, op grond van Europese gelijke behandelingsrichtlijnen, de verantwoordelijkheid om individuele klachten over mogelijk verboden onderscheid in behandeling te nemen, en waar mogelijk beleidsadvies te geven over deze materie om zo problematische praktijken aan de kaak te stellen.

Het College heeft zich in de analoge wereld een aantal keer direct en indirect uitgesproken over (offline) prijsverschillen die verband hielden met een beschermde grond. Hoewel de prijsverschillen in deze oordelen geen betrekking hadden op de digitale omgeving, zijn de principes ontleend aan deze oordelen evenzeer relevant voor die omgeving.

Uitgangspunt is dat een beschermde grond, welke dat ook is, in principe geen zelfstandig element mag zijn in de prijsbepaling van goederen en diensten. Dit betekent onder meer het volgende:

- Het hanteren van een prijsverschil tussen persoon A en persoon B met betrekking tot een goed of dienst dat direct of indirect herleidbaar is tot een beschermde grond is in beginsel niet toelaatbaar, voor zover A en B zich in een vergelijkbare situatie bevinden.

Uitgaande van geslacht kunnen bijvoorbeeld de volgende situaties zich voordoen. Wanneer ondernemer op internationale vrouwendag aan zijn online klantenbestand alleen personen die (vermeend) vrouw zijn een korting aan zou bieden kan dit leiden tot verboden onderscheid.¹ Het kan ook indirecter, bijvoorbeeld wanneer een algoritme overwegend en structureel kortingen zou aanbieden aan personen van een bepaald geslacht op basis van ogenschijnlijk neutrale koperskenmerken (e.g. uitgaande van de lengte van de koper, of het type goederen en diensten dat wordt gekocht) omdat het algoritme inschat dat het ene geslacht nu eenmaal een lagere betalingsbereidheid heeft dan het andere.¹ Ook het meewegen van geslacht in premiebepaling voor e.g. verzekeringen – ondanks eventuele statistische onderbouwingen – zou strijd met de gelijke behandelingswetgeving opleveren.¹

- De keuzevrijheid van A ten opzichte van B om voor bepaalde goederen of diensten te kiezen mag niet worden beperkt als deze beperking is gebaseerd op, of direct of indirect herleidbaar is tot, een beschermde grond, voor zover A en B zich in een vergelijkbare situatie bevinden.

Een vergelijkbaar probleem kan zich voordoen waar het aanbod van onlinediensten of goederen wordt toegesneden op een klant waarbij beschermde gronden een rol hebben gespeeld in het (getoonde) aanbod, samenstelling of omvang van de diensten/goederen.¹ Stel dat een aanbieder een uitgebreidere, duurder set diensten tezamen met een set basisdiensten levert. Een algoritme dat op basis van het (vermeende) ras, de afkomst of nationaliteit van de consument de dienstverlening aanpast door alleen producten uit het basissegment te tonen of te verkopen levert hoogstwaarschijnlijk verboden onderscheid op.¹ Meer grofmazig kunnen ook restricties worden toegepast waardoor sommige onlinedienstverlening niet toegankelijk is op basis van woonplaats (e.g. over de grens – hetgeen indirect onderscheid op grond van nationaliteit kan opleveren)¹ of een variant zoals de eerdergenoemde weigering om onder bepaalde voorwaarden online gekochte goederen te leveren op woonwagenlocaties.¹

Uit de gelijke behandelingswetgeving vloeit overigens niet voort dat een goed of dienst slechts mag worden aangeboden wanneer iedereen er evenveel gebruik van kan maken, of er een gelijkwaardig alternatief wordt aangeboden.¹ Echter, indien een bepaalde dienst of goed aangeboden wordt mag er geen sprake zijn van het beperken van de kring van personen aan wie deze wordt opengesteld als die beperking direct of indirect is gebaseerd op een verboden grond.¹

Conclusie

Data-gedreven (online) prijsbepaling, en differentiëren in prijs, kan (onbedoeld) discriminatoire gevolgen hebben. Het is van belang dat aanbieders van goederen en diensten zich goed bewust zijn van, of waar nodig door de overheid worden gewezen op, de kaders die het gelijke behandelingsrecht stelt en daar rekening mee houden in ontwerp en gebruik van hun algoritmes. Zij moeten daar rekening mee houden in ontwerp en gebruik van hun algoritmes, en alleen niet-discriminerende alternatieven (laten) inzetten.

2. Bestaande standaarden en principes

Er bestaan al de nodige standaarden en best practices. Uiteraard is er wetgeving en jurisprudentie ten aanzien van discriminatie en ongelijke behandeling die in acht moet worden genomen (sectie 2.1) en er is binnen organisaties en in de literatuur een aantal aanvullende normen ontwikkeld (sectie 2.2). Daarnaast zijn twee type normatieve stelsels zijdelings van belang. Allereerst zijn de nodige richtsnoeren met statistische standaarden en methodiek neergelegd (sectie 2.3); tot slot zijn ook binnen de privacy- en gegevensbeschermingscontext richtsnoeren gepubliceerd die raken aan datakwaliteit, transparantie en fairness (sectie 2.4).

2.1 Discriminatie/gelijke behandeling

Het toetsingskader gelijke behandelingsrecht valt uiteen in grofweg vier elementen:

- het bepalen of een situatie onder de reikwijdte valt,
- de kwalificatie van het al dan niet gemaakte onderscheid,
- de toepasselijkheid van uitzonderingen of een rechtvaardigingsgrond,
- een proportionaliteitanalyse.

Deze stappen zullen hieronder puntsgewijs worden uitgewerkt. Hieronder kort eerst een aantal belangrijke wetsbepalingen:

Artikel 26 IVBPR ¹³⁵	Artikel 14 EVRM	Artikel 5 Protocol 7 EVRM ¹³⁶	Artikel 1 Protocol 12 EVRM ¹³⁷
All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.	Het genot van de rechten en vrijheden die in dit Verdrag zijn vermeld, moet worden verzekerd zonder enig onderscheid op welke grond ook, zoals geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst, het behoren tot een nationale minderheid, vermogen, geboorte of andere status.	Echtgenoten hebben gelijke rechten en verantwoordelijkheden van civielrechtelijke aard, zowel onderling als in hun betrekkingen met hun kinderen, wat betreft het huwelijk, tijdens het huwelijk en bij de ontbinding ervan. Dit artikel belet de Staten niet de in het belang van de kinderen noodzakelijke maatregelen te nemen.	1. Het genot van elk in de wet neergelegd recht moet worden verzekerd zonder enige discriminatie op welke grond dan ook, zoals geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst, het behoren tot een nationale minderheid, vermogen, geboorte of andere status. 2. Niemand mag worden gediscrimineerd door enig openbaar gezag op, met name, een van de in het eerste lid vermelde gronden.

¹³⁵ Internationaal Verdrag inzake burgerrechten en politieke rechten, New York, 16-12-1966.

¹³⁶ Protocol Nr. 7 bij het Verdrag tot Bescherming van de Rechten van de Mens en de Fundamentele Vrijheden Straatsburg, 22.XI.1984.

¹³⁷ Protocol Nr. 12 bij het Verdrag tot Bescherming van de Rechten van de Mens en de Fundamentele Vrijheden Rome, 4.XI.2000.

Artikel 20 Handvest EU ¹³⁸	Artikel 21 Handvest EU	Artikel 22 Handvest EU	Artikel 23 Handvest EU	Artikel 24 Handvest EU	Artikel 25 Handvest EU	Artikel 26 Handvest EU
Eenieder is gelijk voor de wet.	1. Iedere discriminatie, met name op grond van geslacht, ras, kleur, etnische of sociale afkomst, genetische kenmerken, taal, godsdienst of overtuiging, politieke of andere denkbeelden, het behoren tot een nationale minderheid, vermogen, geboorte, een handicap, leeftijd of seksuele gerichtheid, is verboden. 2. Binnen de werkingssfeer van de Verdragen en onverminderd de bijzondere bepalingen ervan, is iedere discriminatie op grond van nationaliteit verboden.	De Unie eerbiedigt de verscheidenheid van cultuur, godsdienst en taal.	De gelijkheid van vrouwen en mannen moet worden gewaarborgd op alle gebieden, met inbegrip van werkgelegenheid, beroep en beloning. Het beginsel van gelijkheid belet niet dat maatregelen worden gehandhaafd of genomen waarbij specifieke voordelen worden ingesteld ten gunste van het ondervertegenwoordigde geslacht.	1. Kinderen hebben recht op de bescherming en de zorg die nodig zijn voor hun welzijn. Zij mogen vrijelijk hun mening uiten. Aan hun mening in hen betreffende aangelegenheden wordt in overeenstemming met hun leeftijd en rijpheid passend belang gehecht. 2. Bij alle handelingen in verband met kinderen, ongeacht of deze worden verricht door overheidsinstanties of particuliere instellingen, vormen de belangen van het kind een essentiële overweging. 3. Ieder kind heeft het recht, regelmatig persoonlijke betrekkingen en rechtstreekse contacten met zijn beide ouders te onderhouden, tenzij dit tegen zijn belangen indruist.	De Unie erkent en eerbiedigt het recht van ouderen, een waardig en zelfstandig leven te leiden en aan het maatschappelijk en cultureel leven deel te nemen.	De Unie erkent en eerbiedigt het recht van personen met een handicap op maatregelen die beogen hun zelfstandigheid, hun maatschappelijke en beroepsintegratie en hun deelname aan het gemeenschap s leven te bewerkstelligen.

¹³⁸ Handvest Grondrechten Europese Unie 2012/C 326/02.

Artikel 6 Richtlijn gelijke behandeling arbeid en beroep ¹³⁹	Artikel 4 Richtlijn gelijke behandeling mannen en vrouwen arbeid en beroep ¹⁴⁰	Artikel 4 Richtlijn gelijke behandeling mannen en vrouwen aanbod van goederen en diensten ¹⁴¹	Artikel 4 Voorstel voor Richtlijn gelijke behandeling betreffende godsdienst of overtuiging, handicap, leeftijd of seksuele geaardheid ¹⁴²	Artikel 2 Richtlijn gelijke behandeling ras en etniciteit ¹⁴³
1. Niettegenstaande artikel 2, lid 2, kunnen de lidstaten bepalen dat verschillen in behandeling op grond van leeftijd geen discriminatie vormen indien zij in het kader van de nationale wetgeving objectief en redelijk worden gerechtvaardigd door een legitiem doel, met inbegrip van legitieme doelstellingen van het beleid op het terrein van de werkgelegenheid, de arbeidsmarkt of de beroepsopleiding, en de middelen voor het bereiken van dat doel passend en noodzakelijk zijn. Dergelijke verschillen in behandeling kunnen onder meer omvatten: a) het creëren van bijzondere voorwaarden voor toegang tot arbeid en beroepsopleiding, van bijzondere arbeidsvoorwaarden en -omstandigheden, met inbegrip van voorwaarden voor ontslag en beloning voor jongeren, oudere	Voor gelijke arbeid of voor arbeid waaraan gelijke waarde wordt toegekend, moet directe en indirecte discriminatie op grond van geslacht ten aanzien van alle elementen en voorwaarden van beloning worden afgeschaft. In het bijzonder, wanneer voor de vaststelling van de beloning gebruik wordt gemaakt van een systeem van werkclassificatie, moet dit systeem berusten op criteria die voor mannelijke en vrouwelijke werknemers hetzelfde zijn, en zodanig zijn opgezet dat	1. Voor de toepassing van deze richtlijn houdt het beginsel van gelijke behandeling van mannen en vrouwen in dat a) directe discriminatie op grond van geslacht, inclusief de ongunstigere behandeling van vrouwen wegens zwangerschap en moederschap, verboden is; b) indirecte discriminatie op grond van geslacht verboden is. 2. Deze richtlijn laat gunstiger bepalingen inzake de bescherming van vrouwen in verband met zwangerschap	1. Teneinde te waarborgen dat het beginsel van gelijke behandeling met betrekking tot personen met een handicap wordt nageleefd: a) moeten de nodige maatregelen om personen met een handicap effectieve toegang te geven tot sociale bescherming, sociale voordelen, gezondheidszorg, onderwijs en toegang tot en aanbod van goederen en diensten die publiekelijk beschikbaar zijn, met inbegrip van huisvesting en vervoer, van tevoren worden genomen, onder meer via passende wijzigingen of aanpassingen. Dergelijke maatregelen mogen geen onevenredige belasting vormen, geen fundamentele wijziging van de sociale bescherming, de sociale voordelen,	1. Voor de toepassing van deze richtlijn wordt onder het beginsel van gelijke behandeling verstaan de afwezigheid van elke vorm van directe of indirecte discriminatie op grond van ras of etnische afstamming. 2. Voor de toepassing van lid 1 is er: a) "directe discriminatie", wanneer iemand op grond van ras of etnische afstamming ongunstiger wordt behandeld dan een ander in een vergelijkbare situatie wordt, is of zou worden behandeld; b) "indirecte discriminatie", wanneer een ogenschijnlijk neutrale bepaling, maatstaf of handelwijze

¹³⁹ Richtlijn 2000/78/EG van de Raad van 27 november 2000 tot instelling van een algemeen kader voor gelijke behandeling in arbeid en beroep

¹⁴⁰ Richtlijn 2006/54/EG van het Europees Parlement en de Raad van 5 juli 2006 betreffende de toepassing van het beginsel van gelijke kansen en gelijke behandeling van mannen en vrouwen in arbeid en beroep (herschikking)

¹⁴¹ Richtlijn 2004/113/EG van de Raad van 13 december 2004 houdende toepassing van het beginsel van gelijke behandeling van mannen en vrouwen bij de toegang tot en het aanbod van goederen en diensten

¹⁴² Richtlijn van de Raad betreffende de toepassing van het beginsel van gelijke behandeling van personen ongeacht godsdienst of overtuiging, handicap, leeftijd of seksuele geaardheid {SEC(2008) 2180} {SEC(2008) 2181}

¹⁴³ Richtlijn 2000/43/EG van de Raad van 29 juni 2000 houdende toepassing van het beginsel van gelijke behandeling van personen ongeacht ras of etnische afstamming

werknemers en werknemers met personen ten laste, teneinde hun opnemings in het arbeidsproces te bevorderen, en hun bescherming te verzekeren; b) de vaststelling van minimumvoorwaarden met betrekking tot leeftijd, beroepservaring of -anciënniteit in een functie voor toegang tot de arbeid of bepaalde daaraan verbonden voordelen; c) de vaststelling van een maximumleeftijd voor aanwerving, gebaseerd op de opleidingseisen voor de betrokken functie of op de noodzaak van een aan pensionering voorafgaand redelijk aantal arbeidsjaren. 2. Niettegenstaande artikel 2, lid 2, kunnen de lidstaten bepalen dat de vaststelling, in ondernemings- en sectorale regelingen inzake sociale zekerheid, van een toetredingsleeftijd of van een leeftijd voor het verkrijgen van het recht op pensioen- of invaliditeitsuitkeringen, inclusief de vaststelling van verschillende leeftijden voor werknemers of voor groepen of categorieën werknemers, in de ondernemings- en sectorale regelingen inzake sociale zekerheid, en het gebruik, in het kader van die regelingen, van leeftijdsriteria in de actuariële berekeningen, geen discriminatie op grond van leeftijd vormt, mits dat niet leidt tot discriminatie op grond van geslacht.	elke discriminatie op grond van geslacht is uitgesloten.	en moederschap onverlet. 3. Intimidatie en seksuele intimidatie in de zin van deze richtlijn worden beschouwd als discriminatie op grond van geslacht en zijn derhalve verboden. Het feit dat een persoon zulk gedrag afwijst of lijdzzaam ondergaat, mag niet ten grondslag liggen aan een beslissing die die persoon treft. 4. Een opdracht tot directe of indirecte discriminatie op grond van geslacht wordt beschouwd als discriminatie in de zin van deze richtlijn. 5. Deze richtlijn vormt geen beletsel voor verschillen in behandeling indien de levering van de goederen en diensten aan, uitsluitend of hoofdzakelijk, leden van een bepaald geslacht door een legitiem doel wordt gerechtvaardigd en de middelen voor het bereiken van dat doel passend en noodzakelijk zijn.	de gezondheidszorg, het onderwijs, of de betrokken goederen en diensten vereisen, noch vereisen dat daarvoor alternatieve oplossingen worden aangeboden; b) wordt voorzien in de verplichting om effectieve niet-discriminerende toegang - alsook de nodige redelijke aanpassingen in specifieke gevallen - te waarborgen, tenzij dit een onevenredige belasting vormt. 2. Bij de beoordeling of de nodige maatregelen om aan lid 1 te voldoen een onevenredige belasting vormen, moet in het bijzonder rekening worden gehouden met de omvang en de middelen van de organisatie, haar aard, de geraamde kosten, de levensduur van de goederen en diensten, en de mogelijke voordelen van een betere toegang voor personen met een handicap. Wanneer die belasting in voldoende mate wordt gecompenseerd door bestaande maatregelen in het kader van het door de lidstaten gevoerde beleid inzake gelijke behandeling, mag zij niet als onevenredig worden beschouwd. 3. Deze richtlijn laat de bepalingen van het Gemeenschapsrecht en nationale regels die van toepassing zijn op de toegankelijkheid van goederen of diensten onverlet.	personen van een bepaald ras of een bepaalde etnische afstamming in vergelijking met andere personen bijzonder benadeelt, tenzij die bepaling, maatstaf of handelwijze objectief wordt gerechtvaardigd door een legitiem doel en de middelen voor het bereiken van dat doel passend en noodzakelijk zijn. 3. Intimidatie wordt als discriminatie in de zin van lid 1 beschouwd als er sprake is van ongewenst gedrag dat met ras of etnische afstamming verband houdt, en tot doel of gevolg heeft dat de waardigheid van een persoon wordt aangetast en een bedreigende, vijandige, beledigende, vernederende of kwetsende omgeving wordt gecreëerd. Het begrip intimidatie kan in dit verband worden gedefinieerd in overeenstemming met de nationale wetgeving en praktijken van de lidstaten. 4. Een opdracht tot het discrimineren van personen op grond van hun ras of etnische afstamming wordt beschouwd als discriminatie in de zin van lid 1.

Artikel 1 GW ¹⁴⁴	Artikel 5 Awbg ¹⁴⁵	Artikel 6 Awbg
Allen die zich in Nederland bevinden, worden in gelijke gevallen gelijk behandeld. Discriminatie wegens godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht of op welke grond dan ook, is niet toegestaan.	1 Onderscheid is verboden bij: a. de aanbidding van een betrekking en de behandeling bij de vervulling van een openstaande betrekking; b. arbeidsbemiddeling; c. het aangaan en het beëindigen van een arbeidsverhouding; d. het aanstellen of ontslaan van personen, op wie artikel 3 van de Ambtenarenwet 2017 van toepassing is; e. arbeidsvoorwaarden; f. het laten volgen van onderwijs, scholing en vorming tijdens of voorafgaand aan een arbeidsverhouding; g. bevordering; h. arbeidsomstandigheden. 2 Het eerste lid laat onverlet dat: a. een instelling op godsdienstige of levensbeschouwelijke grondslag, b. een instelling van bijzonder onderwijs, of c. een instelling op politieke grondslag, ten aanzien van personen die voor haar werkzaam zijn onderscheid mag maken op grond van godsdienst, levensovertuiging of politieke gezindheid, voor zover deze kenmerken vanwege de aard van de betrokken specifieke beroepsactiviteit of de context waarin deze wordt uitgeoefend een wezenlijk, legitiem en gerechtvaardigd beroepsvereiste vormen, gezien de grondslag van de instelling. Een zodanig onderscheid mag niet verder gaan dan passend is, gelet op de houding van goede trouw en loyaliteit aan de grondslag van de instelling die van de voor haar werkzame personen mag worden verlangd, en mag niet leiden tot onderscheid op een andere in artikel 1 genoemde grond, onverminderd artikel 2, eerste lid. 2a Het eerste lid laat tevens onverlet de vrijheid van gemeenten om onderscheid te maken op grond van godsdienst of levensovertuiging ten aanzien van een ambtenaar of buitengewoon ambtenaar van de burgerlijke stand die in de uitoefening van zijn ambt onderscheid maakt, tenzij het door hem gemaakte onderscheid is gebaseerd op een algemeen verbindend voorschrift. 3 Het eerste lid is niet van toepassing indien: a. de werkverhouding een privé-karakter heeft, b. het verschil in behandeling berust op een kenmerk dat verband houdt met godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat, en c. dat kenmerk vanwege de aard van de betrokken specifieke beroepsactiviteit of de context waarin deze wordt uitgeoefend, een wezenlijk en bepalend beroepsvereiste vormt, mits het doel legitiem is en het vereiste evenredig is aan dat doel. 4 Het eerste lid is niet van toepassing op eisen met betrekking tot de politieke gezindheid die in redelijkheid kunnen worden gesteld in verband met de vervulling van functies in bestuursorganen en adviesorganen. 5 Het eerste lid is niet van toepassing op eisen met betrekking tot de politieke gezindheid die in redelijkheid kunnen worden gesteld in verband met de vervulling van vertrouwensfuncties. 6 Het eerste lid, onderdeel e, is niet van toepassing op onderscheid op grond van burgerlijke staat met betrekking tot nabestaandenpensioen-voorzieningen en met betrekking tot aanspraken op pensioen die vóór de datum van inwerkingtreding van artikel I, onderdeel B, van de wet van 21 december 2000, houdende wijziging van de Pensioen- en spaarfondsenwet en enige andere wetten in verband met het recht van keuze voor ouderdomspensioen in plaats van nabestaandenpensioen en gelijke behandeling van mannen en vrouwen (Stb. 625), zijn opgebouwd.	Onderscheid is verboden met betrekking tot de voorwaarde en voor en de toegang tot het vrije beroep en de mogelijkheden tot uitoefening van en ontplooiing binnen het vrije beroep.

¹⁴⁴ Grondwet voor het Koninkrijk der Nederlanden van 24 augustus 1815

¹⁴⁵ Wet van 2 maart 1994, houdende algemene regels ter bescherming tegen discriminatie op grond van godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat

Artikel 6a	Artikel 7 Awbg	Artikel 7a Awbg
1Onderscheid is verboden bij het lidmaatschap van of de betrokkenheid bij een werkgevers- of werknemersorganisatie of een vereniging van beroepsgenoten, alsmede bij de voordelen die uit dat lidmaatschap of uit die betrokkenheid voortvloeien. 2Het eerste lid laat onverlet dat: a.een op godsdienstige of levensbeschouwelijke grondslag gebaseerde organisatie of vereniging, of b.een op politieke grondslag gebaseerde organisatie of vereniging, ten aanzien van personen die door een lidmaatschap of anderszins bij haar betrokken zijn onderscheid mag maken op grond van godsdienst, levensovertuiging of politieke gezindheid, voor zover deze kenmerken vanwege de aard van de betrokkenheid of de context waarin specifieke activiteiten worden uitgeoefend een wezenlijk, legitiem en gerechtvaardigd vereiste vormen, gezien de grondslag van de organisatie of vereniging. Een zodanig onderscheid mag niet verder gaan dan passend is, gelet op de houding van goede trouw en loyaliteit aan de grondslag van de organisatie of vereniging die van de daarbij betrokkenen mag worden verlangd, en mag niet leiden tot onderscheid op een andere in artikel 1 genoemde grond, onverminderd artikel 2, eerste lid.	1Onderscheid is verboden bij het aanbieden van of verlenen van toegang tot goederen of diensten en bij het sluiten, uitvoeren of beëindigen van overeenkomsten ter zake, alsmede bij het geven van loopbaanoriëntatie en advies of voorlichting over school- of beroepskeuze, indien dit geschiedt: a.in de uitoefening van een beroep of bedrijf; b.door de openbare dienst; c.door instellingen die werkzaam zijn op het gebied van volkshuisvesting, welzijn, gezondheidszorg, cultuur of onderwijs of d.door natuurlijke personen die niet handelen in de uitoefening van een beroep of bedrijf, voor zover het aanbod in het openbaar geschiedt. 2Het eerste lid, onderdeel c, laat onverlet dat een instelling van bijzonder onderwijs bij de toelating en ten aanzien van de deelname aan het onderwijs onderscheid mag maken op grond van godsdienst, levensovertuiging of geslacht, voor zover deze kenmerken vanwege de aard van het onderwijs een wezenlijk, legitiem en gerechtvaardigd vereiste vormen, gezien de grondslag van de instelling. Onderscheid op grond van geslacht is slechts toegestaan, indien voor alle leerlingen, ongeacht hun geslacht, gelijkwaardige voorzieningen aanwezig zijn. Een zodanig onderscheid mag niet verder gaan dan passend is, gelet op de houding van goede trouw en loyaliteit aan de grondslag van de instelling die van leerlingen mag worden verlangd en mag niet leiden tot onderscheid op een andere in artikel 1 genoemde grond, onverminderd artikel 2, eerste lid. 3Het eerste lid, onderdelen a en d, is niet van toepassing indien: a.de rechtsverhouding een privékarakter heeft, b.het verschil in behandeling berust op een kenmerk dat verband houdt met godsdienst, levensovertuiging, politieke gezindheid, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat, en c.het verschil in behandeling door een legitiem doel wordt gerechtvaardigd en de middelen voor het bereiken van dat doel passend en noodzakelijk zijn.	1Onverminderd artikel 7 is onderscheid op grond van ras verboden bij sociale bescherming, daaronder begrepen sociale zekerheid, en sociale voordelen. 2Bij algemene maatregel van bestuur kunnen de begrippen sociale bescherming, sociale zekerheid en sociale voordelen, bedoeld in het eerste lid, worden omschreven. De voordracht voor een krachtens de eerste volzin vast te stellen algemene maatregel van bestuur wordt niet eerder gedaan dan vier weken nadat het ontwerp aan beide kamers der Staten-Generaal is overgelegd.

2.1.1 Reikwijdte

Het gelijkebehandelingsrecht – in brede zin – verbiedt discriminatie op bepaalde terreinen ('beschermde terreinen') met betrekking tot de in de internationale instrumenten en/of wet genoemde gronden (zogenoemde 'beschermde gronden'). Het recht heeft een zeer heterogene grondslag in internationale, EU en nationale instrumenten. Een belangrijk element ter bepaling van de reikwijdte is de bepaling of er sprake is van overheidshandelen dan wel relaties in het economisch verkeer (in het bijzonder: arbeid en het aanbieden van goederen en diensten).

Overheidshandelen

Bij overheidshandelen gelden de normen zoals vervat in Artikel 26 IVPBR, Artikel 14 EVRM, Artikel 1 Protocol 12 bij het EVRM en, voor Nederland, Artikel 1 Grondwet en de nadere uitwerking daarvan in verschillende gelijkebehandelingswetten. Daarnaast is er een veelheid aan (specialistische) internationale en EU-instrumenten die een en ander voor specifieke terreinen nader reguleren. Daarbij moet in het bijzonder het EU non-discriminatiebeginsel op grond van nationaliteit¹⁴⁶ en/of ras en etniciteit (wat betreft overheidshandelen op het terrein van de sociale bescherming)¹⁴⁷ genoemd worden. Daarnaast zijn de instellingen van de EU en de EU-lidstaten waar zij Unierecht ten uitvoer brengen verplicht de non-discriminatie-normen¹⁴⁸ vervat in het EU Handvest van de grondrechten van de Europese Unie na te leven.¹⁴⁹

Voor toepasselijkheid van deze normen is het voornaamste aanknopingspunt dat er sprake moet zijn van discriminatoir overheidsoptreden (het beschermde terrein): benadeling door de overheid zelf,¹⁵⁰ of nalaten op te treden in horizontale situaties waar ingrijpen noodzakelijk was.¹⁵¹ De hierboven genoemde artikelen komen overeen in die zin dat er sprake is van een niet-limitatieve lijst is van gronden die door het verbod beschermd worden. Het EHRM heeft de reikwijdte uitgedrukt als omvattende alle: 'differences based on an identifiable, objective, or personal characteristic, or "status", by which individuals or groups are distinguishable from one another'. De potentiële reikwijdte moet dus breed worden uitgelegd¹⁵² en is niet beperkt tot

¹⁴⁶ Zie Artikel 18 VWEU. Overigens geldt het non-discriminatiebeginsel op grond van nationaliteit alleen voor onderdanen van de EU-lidstaten, tenzij specifieke EU instrumenten de werkingssfeer uitbreiden. Zie ook: HvJ, 4 juni 2009, C-22/08 en C-23/08, ECLI:EU:C:2009:344 (Vatsouras en Koupatantze v Arbeitsgemeinschaft (ARGE) Nürnberg 900), para. 48ff.

¹⁴⁷ Richtlijn 2000/43/EG van de Raad van 29 juni 2000 houdende toepassing van het beginsel van gelijke behandeling van personen ongeacht ras of etnische afstamming, EU Publ. [2000] L 180/22.

¹⁴⁸ Artikel 21 EU Handvest.

¹⁴⁹ Artikel 51 EU Handvest. Zie voor de reikwijdte van dit artikel FRA, Applying the Charter of Fundamental Rights of the European Union in law and policymaking at national level, (European Union Agency for Fundamental Rights, 2018) en HvJ, 26 februari 2013, C-617/10, ECLI:EU:C:2013:105 (Åklagaren v Hans Åkerberg Fransson).

¹⁵⁰ Council of Europe, Guide on Article 14 of the European Convention on Human Rights and on Article 1 of Protocol No. 12 to the Convention, (Council of Europe, 2020), para. 21ff.

¹⁵¹ Zie voor het laatste punt e.g. EHRM, 9 juli 2019, nr. 41261/17, ECLI:CE:ECHR:2019:0709JUD004126117 (Volodina v. Russia), para. 109ff.

¹⁵² Zie EHRM, 16 maart 2010, nr. 42184/05, ECLI:CE:ECHR:2010:0316JUD004218405 (Carson a.o. v. United Kingdom), para. 70 en de daar genoemde voorbeelden. Zie ook: Council of Europe, Guide on Article 14 of the European Convention on Human Rights and on Article 1 of Protocol No. 12 to the Convention, (Council of Europe, 2020) p. 37-38.

‘characteristics which are personal in the sense that they are innate or inherent’.¹⁵³ Tegelijkertijd omvat het niet elk mogelijk onderscheid.¹⁵⁴

Relaties in het economisch verkeer

Het is een axioma dat internationaal recht individuen (onderling) niet bindt. Dit geldt ook voor de EU-regelgeving die worden gerekend tot het EU non-discriminatie acquis¹⁵⁵ daar deze regelgeving uit Richtlijnen bestaat.¹⁵⁶ Voor horizontale situaties moet dus aansluiting gezocht worden bij nationale wetgeving ter implementatie van deze verplichtingen. Voor Nederland zijn de *beschermd gronden* limitatief neergelegd in verschillende wetten en besluiten. Daarnaast moet de discriminatie hebben plaatsgevonden op een *beschermd terrein*: het economisch verkeer. Dat valt uiteen in *arbeid* (in brede zin,¹⁵⁷ inclusief beroepsopleidingen) en/of het *aanbieden van goederen en diensten* (inclusief bepaalde door de overheid aangeboden diensten, zoals onderwijs of gezondheidszorg).¹⁵⁸

Bron	Beschermd Grond	Beschermd terrein
Algemene wet gelijke behandeling (AWGB) (Besluit gelijke behandeling)	Burgerlijke staat, geslacht, godsdienst, hetero- of homoseksuele gerichtheid, levensovertuiging, nationaliteit, politieke gezindheid en ras/ethniciteit	Arbeid Aanbieden van goederen en diensten (Sociale bescherming ras/ethniciteit → overheidshandelen)
Wet gelijke behandeling op grond van handicap of chronische ziekte (WGBH/CZ)	Handicap/chronische ziekte	Arbeid Aanbieden van goederen en diensten ¹⁵⁹
Wet gelijke behandeling van mannen en vrouwen (WGBMV); artikel 7:646 BW	Geslacht	Arbeid
Wet gelijke behandeling op grond van leeftijd bij de arbeid (WGBL)	Leeftijd	Arbeid
Wet onderscheid arbeidsduur (WOA), Wet Onderscheid bepaalde en onbepaalde tijd (WOBOT)	Deeltijd/volgtijd en vaste/tijdelijke contracten	Arbeid

¹⁵³ EHRM, 25 mei 2016, nr. 38590/10, ECLI:CE:ECHR:2016:0524JUD003859010 (Biao v. Denmark), para. 89.

¹⁵⁴ Bijvoorbeeld de regel van de Zweedse overheid dat het alleen onderhandelde met representatieve vakbonden (waarbij kleine vakbonden dus soms buiten de boot vallen); EHRM, 6 februari 1976, nr. 5614/72

ECLI:CE:ECHR:1976:0206JUD000561472 (Swedish Engine Drivers' Union v Sweden), para. 46. Zie voor een kritische beschouwing J. Gerards, 'The Discrimination Grounds of Article 14 of the European Convention on Human Rights', 13(1) Human Rights Law Review (2013) 99.

¹⁵⁵ Zie https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en laatst bezocht

¹⁵⁶ Zie HvJ, 26 februari 1986, 152/84, ECLI:EU:C:1986:84.

¹⁵⁷ Zie Artikel 5 Awgb e.v.

¹⁵⁸ Zie Artikel 7 Awgb e.v.

¹⁵⁹ Speciale regels gelden voor het aanbieden van woon- en openbaarvervoersdiensten.

Concreet betekent een en ander dat de weigering om iemand promotie te laten maken vanwege zijn hanenkam (geen beschermde grond),¹⁶⁰ of de weigering van een dierenasiel om een kat mee te geven aan een vrouw van 75 jaar omdat zij te oud zou zijn (geen beschermd terrein) niet getoetst kan worden op grond van de huidige regelgeving. Het Nederlandse gelijkebehandelingsrecht kent daarmee in het economisch verkeer een zogenoemd gesloten stelsel van (beschermde) gronden.

2.1.2 Bepaling en kwalificatie van het onderscheid

De volgende stap is het bepalen van het type onderscheid.

2.1.2.1 Direct en indirect onderscheid

Van direct onderscheid is sprake als een persoon op basis van (kenmerken behorend tot) een beschermde grond wordt of zou worden benadeeld ten opzichte van een ander, in een vergelijkbare situatie,¹⁶¹ of de benadeling erin besloten ligt dat verschillende situaties gelijk worden behandeld.¹⁶²

Een man solliciteerde twee keer op dezelfde functie, eerst onder zijn eigen (Marokkaans klinkende) naam, en later onder een Nederlands klinkende naam met een voor het overige gelijke CV en brief. De man werd enkel uitgenodigd onder zijn Nederlandse naam. Het College stelt vast dat het bedrijf direct onderscheid op grond van (etnische) afkomst heeft gemaakt.¹⁶³

Van indirect onderscheid is sprake als een ogenschijnlijk neutrale bepaling, maatstaf of handelwijze een groep personen gekenmerkt door (elementen herleidbaar tot) een beschermde grond in het bijzonder treft in vergelijking met een andere groep personen in een vergelijkbare situatie.

Een DigiD – een digitaal authenticatie en communicatiemiddel – was alleen beschikbaar voor personen met een Nederlandse nationaliteit of personen die in Nederland wonen. Personen met een niet-Nederlandse nationaliteit die in het buitenland wonen maar in Nederland werken kwamen niet in aanmerking. Dit leverde indirecte discriminatie op grond van nationaliteit op, nu in Nederland

¹⁶⁰ Zie e.g. CRM, 1 januari 1997, 1997-15, <https://mensenrechten.nl/nl/oordeel/1997-15>. Overigens is dat anders als de haardracht met een bepaald geloof of ras/etniciteit samenhangt: CRM, 1 januari 2005, 2005-162, <https://mensenrechten.nl/nl/oordeel/2005-162>. Ook het hanteren van verschillende regels voor mannen en vrouwen voor uiterlijk kan onderscheid opleveren: CRM, 11 augustus 2011, 2011-126, <https://mensenrechten.nl/nl/oordeel/2011-126>.

¹⁶¹ EHRM, 29 april 2008, nr. 13378/05, ECLI:CE:ECHR:2008:0429JUD001337805 (Burden v. The United Kingdom) para. 60.

¹⁶² HvJ, 16 december 2008, C-127/07, ECLI:EU:C:2008:728 (*Société Arcelor Atlantique et Lorraine and Others v Premier ministre, Ministre de l'Écologie et du Développement durable and Ministre de l'Économie, des Finances et de l'Industrie*).

¹⁶³ CRM, 6 december 2018, 2018-129, <https://mensenrechten.nl/nl/oordeel/2018-129>.

*overwegend Nederlanders wonen, en in het buitenland overwegend niet-Nederlanders.*¹⁶⁴

Of zoals het EHRM optekende ten aanzien van het gebrek aan wetgeving en beleid in Rusland om huiselijk geweld te bestraffen:

*The Court reiterates that a general policy or a de facto situation which has disproportionately prejudicial effects on a particular group may constitute discrimination within the meaning of Article 14 of the Convention, even where it does not specifically target that group and there is no discriminatory intent. Violence against women, including domestic violence, is a form of discrimination against women. The State's failure to protect women against domestic violence breaches their right to equal protection of the law, irrespective of whether such failure is intentional or not (...).*¹⁶⁵

Voor wat betreft de te hanteren maatstaf met betrekking tot 'in het bijzonder treffen' heeft het Hof van Justitie van de EU in een recente zaak als volgt geoordeeld:

*Het bestaan van een bijzonder nadeel kan met name worden vastgesteld wanneer wordt bewezen dat een regeling als die welke in het hoofdgeding aan de orde is, **een aanzienlijk groter percentage personen** van het ene dan van het andere geslacht treft (...).*¹⁶⁶

Een klein verschil in de verhouding zal dus geen indirecte discriminatie opleveren. Waar de grens echter precies loopt is niet bepaald.

2.1.2.2 Vergelijkbare situaties

Uitgangspunt is dat de vergelijkbaarheid van de situaties onderzocht moeten worden aan de hand van de specifieke omstandigheden van het geval en de (rechts)context:

(...) the requirement to demonstrate an analogous position does not require that the comparator groups be identical. An applicant must demonstrate that, having regard to the particular nature of his or her complaint, he or she was in a relevantly similar situation to others treated differently (...) taking into account the elements that characterise their circumstances. (...) The elements which characterise different situations must be assessed in the light of the

¹⁶⁴ ABRvS, 24 december 2019, ECLI:NL:RVS:2019:4434.

¹⁶⁵ EHRM, 16 juni 2020, nr. 65557/14, ECLI:CE:ECHR:2020:0616JUD006555714 (Polshina v. Russia), para. 44.

¹⁶⁶ HvJ, 3 oktober 2019, C-274/18, ECLI:EU:C:2019:828 (Minoo Schuch-Ghannadan v Medizinische Universität Wien), para. 45.

*subject-matter and purpose of the measure which makes the distinction in question.*¹⁶⁷

Het Hof van Justitie hangt eenzelfde interpretatie aan:

*Dienaangaande zij gepreciseerd dat enerzijds niet is vereist dat de situaties identiek zijn, maar alleen dat zij vergelijkbaar zijn, en anderzijds de vergelijkbaarheid niet algemeen en abstract, maar specifiek en concreet dient te worden onderzocht voor de betrokken [handeling die het onderscheid maakt].*¹⁶⁸

De mee te wegen aspecten:

*moeten met name worden bepaald en beoordeeld tegen de achtergrond van het voorwerp en het doel van de [handeling] die het betrokken onderscheid invoert. Bovendien moet rekening worden gehouden met de beginselen en doelstellingen van het gebied waaronder de betrokken handeling valt*¹⁶⁹

Dit betekende, naar mening van de twee rechters in een recente EHRM-zaak dat personen met een beperking als gevolg van een ongeval vergelijkbaar konden worden geacht aan personen met een beperking opgelopen in het kader van hun militaire dienst wanneer de door hun opgelopen beperking en de daaraan gelinkte behoeftes vergelijkbaar waren (de ‘oorsprong’ van de beperking, noch het toepasselijk zijn van verschillende wettelijke regimes, was dus niet doorslaggevend).¹⁷⁰

Aan de andere kant kwam het Hof van Justitie tot de conclusie dat een rijbewijs voor vrachtwagens, gezien de karakteristieken van de machine, de manier van rijden en de daarmee gepaarde verantwoordelijkheden inzake verkeersveiligheid, wezenlijk anders was dan een rijbewijs voor een personenauto. Dit maakte bestuurders van de eerste categorie (en de eisen daaraan gesteld) niet vergelijkbaar met bestuurders van de tweede categorie.¹⁷¹

Overigens is het niet voor alle vormen van ongelijke behandeling verplicht om een ‘comparator’ te identificeren. Zwangerschapsdiscriminatie is daar een voorbeeld van.¹⁷²

Volgens VJV dient de tweede vraag bevestigend te worden beantwoord, omdat het niet gaat om het discriminerend effect van een maatregel in abstracto, maar om de concrete beslissing van een werkgever om een bepaalde sollicitant niet in dienst te nemen. Wanneer een werkgever een keuze maakt tussen uitsluitend vrouwelijke

¹⁶⁷ EHRM, 5 september 2017, nr. 78117/13, ECLI:CE:ECHR:2017:0905JUD007811713 (Fábián v. Hungary), para. 113, 121.

¹⁶⁸ HvJ, 12 december 2013, C-267/12, ECLI:EU:C:2013:823 (Frédéric Hay v Crédit agricole mutuel de Charente-Maritime et des Deux-Sèvres), para. 33.

¹⁶⁹ HvJ, 19 december 2019, C-460/18 P, ECLI:EU:C:2019:1119 (HK v European Commission), para. 67.

¹⁷⁰ EHRM, 30 juni 2020, nr. 26944/13, ECLI:CE:ECHR:2020:0630JUD002694413 (Popović a.o. v. Serbia).

¹⁷¹ HvJ, 22 mei 2014, C-356/12 (Wolfgang Glatzel v Freistaat Bayern), para. 80ff.

¹⁷² HvJ, 8 november 1990, C-177/88, ECLI:EU:C:1990:383 (Elisabeth Johanna Pacifica Dekker v Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus), para. 17.

sollicitanten, zou bij het nemen van die beslissing geen sprake kunnen zijn van onderscheid naar geslacht, omdat een werkgever zich in dat geval laat leiden door andere, financiële of beleidsmatige overwegingen.

Wat dit aangaat, zij eraan herinnerd, dat het antwoord op de vraag of de weigering om een vrouw aan te stellen, een directe of indirecte discriminatie oplevert, afhangt van het motief van deze weigering. Is dit motief de omstandigheid dat betrokkene zwanger is, dan houdt het besluit rechtstreeks verband met het geslacht van de sollicitant. Zo gezien is het voor het antwoord op de eerste prejudiciële vraag niet van belang, dat er geen mannelijke sollicitanten waren.

Het gaat in dit soort situaties normaliter om vormen van benadeling als gevolg van stereotypering.

2.1.2.3 Bijzondere vormen van onderscheid

Behalve het al genoemde directe en indirecte onderscheid zijn er nog specifieke vormen van onderscheid. Zo verwijst Artikel 1a AWGB naar *intimidatie*:

- 1 Het in deze wet neergelegde verbod van onderscheid houdt mede in een verbod van intimidatie en een verbod van seksuele intimidatie.
- 2 Onder intimidatie als bedoeld in het eerste lid wordt verstaan: gedrag dat met de hoedanigheden of gedragingen, bedoeld in artikel 1, eerste lid, onderdeel b, verband houdt en dat tot doel of gevolg heeft dat de waardigheid van de persoon wordt aangetast en dat een bedreigende, vijandige, beledigende, vernederende of kwetsende omgeving wordt gecreëerd.
- 3 Onder seksuele intimidatie als bedoeld in het eerste lid wordt verstaan: enige vorm van verbaal, non-verbaal of fysiek gedrag met een seksuele connotatie dat als doel of gevolg heeft dat de waardigheid van de persoon wordt aangetast, in het bijzonder wanneer een bedreigende, vijandige, beledigende, vernederende of kwetsende omgeving wordt gecreëerd.
(...)

In de Wet gelijke behandeling op grond van handicap en chronische ziekte staan daarnaast nog twee varianten opgenomen. Uitgangspunt van de WGBH/CZ is dat ‘ieder mens in staat [moet worden] gesteld aansluitend bij zijn eigen mogelijkheden autonoom te zijn’.¹⁷³ Het in die wet geformuleerde discriminatieverbod, het recht op een doeltreffende aanpassing en het recht op algemene toegankelijkheid moet garanderen dat mensen met een beperking of chronische ziekte op een volwaardige, effectieve en gelijkwaardige wijze kunnen deelnemen aan het maatschappelijk leven.¹⁷⁴

Het recht op een *doeltreffende aanpassing* omvat:

*Het verbod van onderscheid houdt mede in dat degene, tot wie dit verbod zich richt, gehouden is naar gelang de behoefte doeltreffende aanpassingen te verrichten, tenzij deze voor hem een onevenredige belasting vormen.*¹⁷⁵

¹⁷³ Artikel 01 WGBH/CZ.

¹⁷⁴ Preambule E, artikel 1 en 3 VN Verdrag inzake de rechten van personen met een handicap, Trb. 2007, 169. Zie ook Artikel CGB Advies 2006/14, Wet gelijke behandeling op grond van handicap of chronische ziekte (WGBH/CZ) in verband met uitbreiding van de reikwijdte tot het primair en voortgezet onderwijs, p. 3.

¹⁷⁵ Artikel 2 WGBH/CZ.

Dit betekent dat een persoon een op zijn of haar individuele situatie toegesneden aanpassing kan verlangen met het doel een belemmering of nadeel ondervonden in de deelname aan het maatschappelijk leven als gevolg van een handicap of chronische ziekte (H/CZ) op te heffen.

Voorbeeld: Het verzoek van een vrouw om de bibliotheek te betreden met haar scootmobiel in plaats van de (handaangedreven) rolstoelen te gebruiken is een voorbeeld van een doeltreffende aanpassing. Het stelt haar in staat om autonoom gebruik te maken van de diensten van de bibliotheek.¹⁷⁶

Toegankelijkheid ligt in het verlengde van de doeltreffende aanpassing en streeft naar volledige inclusie: hoe toegankelijker de samenleving, hoe minder aanpassing in het individuele geval nodig is.¹⁷⁷ Artikel 2a WGBH/CZ omvat daarom de verplichting om anticiperend belemmeringen in de sociale, fysieke of economische sfeer, of op informatie en communicatievlak¹⁷⁸ voor personen met een beperking weg te nemen. Geëigende voorbeelden zijn standaardisatie initiatieven met het doel ervoor te zorgen dat gebouwen toegankelijk zijn voor mensen in een rolstoel, het gebruik van hoog contrast en grote lettertypen op websites voor slechtzienden, of het gebruik van eenvoudige taal voor communicatiedoeleinden zodat de boodschap begrijpelijk is voor personen met een verstandelijke beperking.

2.1.2.4 Samenloop van gronden

Discriminatie kan complexe vormen aannemen bij samenloop van gronden. Er worden normaliter drie vormen van samenloop onderscheiden:¹⁷⁹

- Twee (of meer) gronden kunnen beide, onafhankelijk van elkaar aan de orde zijn (*meervoudige discriminatie*);
In zaak CRM, 23 juli 2019, 2019-73¹⁸⁰ stelde verzoeker dat de decaan van onderwijsinstelling tijdens het afwijzingsgesprek en in het afwijzingsbesluit heeft verwezen naar zowel de leeftijd van de man (ouder dan de gemiddelde scholier), zijn autisme (handicap/chronische ziekte) en zijn Nederlandse afkomst (ras).
- de ene grond kan de andere versterken (*samengestelde discriminatie*);
In CRM, 22 december 2006, 2006-256 ondervond verzoekster nadeel van het feit dat zij een mondeling examen moest doen in plaats van schriftelijk (waarbij de vragen van het schriftelijke examen werden voorgelezen en de vrouw moest antwoorden). Ze ondervond hierin een grieppende benadeling nu zij niet de mogelijkheid had, net als niet-visueel beperkte examendeelnemers, de vragen te kunnen herlezen/nalezen (wat wel had gekund indien het examen in braille zou zijn geschreven). Dit nadeel werd versterkt door het feit dat haar moedertaal niet Nederlands, maar Turks was (hetgeen bij een schriftelijke toets minder nadelig is dan bij een mondelinge).

¹⁷⁶ CRM, 4 juli 2019, 2019-64, <https://mensenrechten.nl/nl/oordeel/2019-64>.

¹⁷⁷ Kamerstukken II 2013/14, 33992-(R2034), 3 (MvT), p. 33.

¹⁷⁸ S. Charitakis, Access Denied: The Role of the European Union in Ensuring Accessibility under the United Nations Convention the Rights of Persons with Disabilities, (Intersentia, 2018), p. 90, 25-31.

¹⁷⁹ CRM, 28 april 2011, 2011-68, para. 3.3.

¹⁸⁰ <https://mensenrechten.nl/nl/oordeel/2019-73>.

- de beide gronden kunnen zodanig nauw zijn verweven dat het nadeel zich uitsluitend voordoet wanneer beide gronden aan de orde zijn (*intersectionele discriminatie*).

In zaak *Parris v Trinity College and Others*¹⁸¹ verzocht Dr. Parris dat zijn partner – een persoon van hetzelfde geslacht - aanspraak kon maken op een weduwnaarspensioen ingeval van zijn overlijden. Dit werd echter geweigerd omdat hij slechts na zijn 60^{ste} een geregistreerd partnerschap was aangegaan. Het geregistreerd partnerschap voor personen van hetzelfde geslacht was in Ierland echter slechts kort ervoor opengesteld voor personen van hetzelfde geslacht, toen Dr. Parris de leeftijd van 60 al had bereikt. Dit betekent dat Dr. Parris niet zozeer van wege zijn homoseksuele gerichtheid werd benadeeld (immers zouden ook heteroseksuele personen die geen geregistreerd partnerschap voor hun 60^{ste} waren aangegaan niet in aanmerking komen), noch zijn leeftijd (er bestaat een uitzondering¹⁸² die het toelaat dat een lidstaat voor het verkrijgen van bepaalde pensioenuitkeringen een toetredingsleeftijd hanteren). Samengenomen (seksuele gerichtheid en zijn leeftijd), echter, was duidelijk dat Parris echter een bijzonder nadeel ondervond nu hij nooit de mogelijkheid heeft gehad om aan de regel te voldoen, terwijl personen zonder deze kenmerken wel deze mogelijkheid hebben gehad.

Overigens oordeelde het Hof van Justitie nu de Richtlijnen de discriminatiegronden limitatief opsommen er geen mogelijkheid is tot de erkenning, in het EU recht, van intersectionele discriminatie (dat immers per definitie geen discriminatie van één van de onderliggende gronden oplevert).¹⁸³ Een en ander geldt, mutatis mutandis, ook voor Nederland nu we ook een gesloten systeem kennen dat (grotendeels) gebaseerd is op de EU-Richtlijnen.

2.1.2.5 ‘Extensies’

De bescherming tegen discriminatie mag niet restrictief worden uitgelegd.¹⁸⁴ In sommige gevallen kan een persoon worden benadeeld vanwege zijn/haar/anders *associatie* of band met een ander met beschermde persoonskenmerken. Zowel het EHRM¹⁸⁵ als het HvJ hebben geaccepteerd dat in zulke gevallen het gelijkebehandelingsrecht van toepassing is:

In *Coleman* werd een vrouw benadeeld en geïntimideerd door haar werkgever vanwege de zorg die zij leverde voor haar gehandicapte kind, terwijl ouders van niet-gehandicapte kinderen deze benadeling door de werkgever niet ondervonden. Zij kon zich beroepen op het verbod op onderscheid op grond van handicap/chronische ziekte.¹⁸⁶

De bescherming van het gelijke behandelingsrecht tegen benadeling vanwege associatie is niet beperkt tot situaties van direct onderscheid.

¹⁸¹ HvJ, 24 november 2016, C-443/15, ECLI:EU:C:2016:897 (David L. Parris v Trinity College Dublin, Higher Education Authority, Department of Public Expenditure and Reform, Department of Education and Skills).

¹⁸² Artikel 6 lid 2 Richtlijn 2000/78.

¹⁸³ Ibid, para. 81.

¹⁸⁴ HvJ, 23 april 2020, C-507/18, ECLI:EU:C:2020:289 (NH v Associazione Avvocatura per i diritti LGBTI — Rete Lenford), para. 39.

¹⁸⁵ Zie EHRM, 22 maart 2016, nr. 23682/13, ECLI:CE:ECHR:2016:0322JUD002368213 (Guberina v. Croatia), para. 79.

¹⁸⁶ HvJ, 17 juli 2008, C-303/06, ECLI:EU:C:2008:415 (S. Coleman v Attridge Law and Steve Law).

Een vrouw exploiteerde een kruidenierswinkel in een wijk voornamelijk bewoond door personen van Roma-afkomst. Een elektriciteitsbedrijf plaatste in gebieden waar (vermeend) veel fraude werd gepleegd de elektriciteitsmeters op hoge palen. De maatregel trof in de praktijk vooral wijken overwegend bewoond door Roma. De vrouw, die zelf niet identificeerde als Roma en/of woonwagenbewoner, werd hierdoor benadeeld omdat zij haar elektriciteitsmeter niet kon aflezen. Het Hof van Justitie oordeelde dat zij zich kon beroepen op het verbod op onderscheid op grond van ras/etniciteit.¹⁸⁷

Daarnaast geldt de bescherming van het gelijke behandelingsrecht ook bij discriminatie op basis van *vermeende* persoonskenmerken.

Een uitzendbureau deelde een (mogelijke) inlener van een uitzendkracht mee dat de beoogde uitzendkracht last had van (chronische) rugklachten. Daarop besloot de inlener de uitzendkracht niet langer in te lenen. De uitzendkracht in kwestie was inmiddels voor zijn aandoening behandeld waardoor hij geen beperkingen meer had. Niettemin kon hij zich op het verbod op onderscheid op grond van handicap/chronische ziekte beroepen nu hij benadeeld was vanwege zijn vermeend beperkte capaciteiten.¹⁸⁸

Een variant hierop is discriminatie op grond van verwachte of toekomstige persoonskenmerken. Klassiek voorbeeld is het niet aannemen van een vrouw vanwege (eventuele) toekomstige kindwensen of vanwege een inschatting van het ziekteverloop van een persoon met handicap/chronische ziekte.

Een universiteit legde een hoogleraar die een hersenbloeding had gehad die samenhang met een erfelijke aandoening de maatregel op dat hij bij de begeleiding van een promovendus of promovenda moest zorgen voor achtervang. Zijn collega's werd deze regel niet opgelegd. De man werd hier benadeeld vanwege de mogelijke, toekomstige ontwikkeling van zijn aandoening en kon zich op het verbod van onderscheid op grond van handicap/chronische ziekte beroepen.¹⁸⁹

Er is wel beargumenteerd dat juist bij algoritmische besluitvorming deze 'extensies' (associatief, toekomstig en vermeend) van bijzonder belang kunnen zijn omdat de data die verzameld wordt door bedrijven veelal niet direct sensitieve data zal zijn (daar dit op basis van de AVG problematisch zou zijn) maar in het bijzonder proxies waaruit deze data kan worden afgeleid.¹⁹⁰

2.1.3 Uitzondering & rechtvaardiging

Niet elk gemaakt onderscheid binnen de werkingssfeer van het gelijkebehandelingsrecht is ook verboden. Het beoordelingskader of onderscheid toegelaten is hangt af van twee elementen:

¹⁸⁷ HvJ, 16 juli 2015, C-83/14, ECLI:EU:C:2015:480 (CHEZ Razpredelenie Bulgaria AD v Komisia za zashtita ot diskriminatsia).

¹⁸⁸ CRM, 17 juli 2017, 2017-93, <https://mensenrechten.nl/nl/oordeel/2017-93>.

¹⁸⁹ HvJ, 7 juni 2019, 2019-55, <https://mensenrechten.nl/nl/oordeel/2019-55>. Zie ook: CRM, 12 augustus 2019, 2019-94, <https://mensenrechten.nl/nl/oordeel/2019-94>.

¹⁹⁰ Zie S. Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising', 35(2) Berkeley Technology Law (2020) – forthcoming (paper beschikbaar op: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3388639).

2.1.3.1 De kwalificatie van het onderscheid

Direct onderscheid is een ernstiger inbreuk op het discriminatieverbod dan indirect onderscheid en behoeft dus een sterkere rechtvaardiging. In Nederland (en in sommige EU regelgeving) is hier uiting aan gegeven met het principe – in de context van het gelijkebehandelingsrecht zoals dat o.a. is neergelegd in de AWGB – dat direct onderscheid enkel is toegelaten als er sprake is van een in de wet opgenomen uitzondering.¹⁹¹ Dit is bijvoorbeeld ‘waar het geslacht bepalend is’ (e.g. een filmrol geschreven voor een mannelijke acteur)¹⁹² en waar het gaat om zogenoemde ‘wezenlijke beroepsvereisten’ bij instellingen op godsdienstige, levensbeschouwelijke of politieke grondslag.¹⁹³ Er zijn ook uitzonderingen voor bepaalde maatregelen die proberen ondervonden nadelen/achterstand van een bepaalde groep op te heffen (ook wel ‘positive action’ genoemd):¹⁹⁴ denk daarbij aan de klassieke maatregel bij werving en selectie die stelt dat bij gelijke geschiktheid voorkeur zal worden gegeven aan een vrouw.

Indirect onderscheid is daarnaast toelaatbaar (buiten de eerdergenoemde wettelijke uitzonderingen) indien er een goede reden voor het gemaakte onderscheid is: de objectieve rechtvaardiging.¹⁹⁵ Voor een objectieve rechtvaardiging is nodig dat de maatregel die onderscheid maakt een legitiem doel dient, dat op zichzelf niet-discriminatoir van aard is. Vervolgens moet de proportionaliteit van de aangewende middelen (de gemaakte keuzes bij het vormgeven van de onderscheidmakende maatregel) worden onderzocht. Dit betreft een driedelige toets. De maatregel moet passend, noodzakelijk en evenredig zijn.¹⁹⁶

2.1.3.1.1 Passend

Passend betekent een evaluatie van de geschiktheid van de maatregel om het gestelde doel te bereiken: kan deze aan het doel bijdragen? Is er een logische link? In deze stap wordt ook geëvalueerd of de maatregel consistent (vrij van innerlijke tegenstrijdigheden) en coherent (beoordeling van de maatregel in het bredere rechtskader) is.

Voorbeeld: In *CO v Commune di Gesturi* draaide het om een regel die gepensioneerden verbood om mee te dingen in procedures voor de gunning – van overheidswege - van onderzoeks- en adviesopdrachten, tenzij zij het ging om een bestuursfunctie en zij daarvoor geen salaris vroegen. Hier is sprake van leeftijdsdiscriminatie. De ingeroepen rechtvaardigingsgrond was tweeledig: het bevorderen van werkgelegenheid (jongeren meer kansen te geven op de arbeidsmarkt) en een beperking van overheidsuitgaven (voorkomen cumulatie van pensioen en loon). Het Hof twijfelde echter zowel aan de algemene geschiktheid als de consistentie.¹⁹⁷ Ervaring was nu eenmaal nodig voor de betreffende opdrachten, en het effect op toegang tot de arbeidsmarkt zou slechts beperkt aanwezig zijn in die zin dat jongeren eventueel de baan die een meer ervaren concurrent

¹⁹¹ Zie bijvoorbeeld HvJ, 12 december 2019, C-450/18, ECLI:EU:C:2019:1075 (WA v Instituto Nacional de la Seguridad Social), para. 54. Voor de grond leeftijd geldt dit niet: daar kan elke vorm van onderscheid objectief worden gerechtvaardigd, zie Artikel 7 lid 1 sub c WGBL.

¹⁹² Zie hiervoor het Besluit Gelijke Behandeling. Zie daarnaast: HvJ, 21 mei 1985, 248/83, ECLI:EU:C:1985:214 (Commissie v Duitsland).

¹⁹³ Artikel 5 lid 2 AWGB.

¹⁹⁴ Artikel 2 lid 3 AWGB.

¹⁹⁵ Deze mogelijkheid geldt ook bij direct onderscheid op grond van leeftijd, zoals eerder vermeld.

¹⁹⁶ Voor de doeltreffende aanpassing en/of toegankelijkheid geldt een aangepaste toets die hierboven al werd vermeld. Een en ander wordt daar afgewogen in het kader van de ‘onevenredige belasting’.

¹⁹⁷ HvJ, 2 april 2020, C-670/18, ECLI:EU:C:2020:272 (CO v Commune di Gesturi), para. 47, 51.

had verlaten zouden kunnen overnemen. Maar dat vrijkomen van banen was op zijn beurt niet waarschijnlijk vanwege het incidentele en niet-permanente karakter van zulke opdrachten. Bovendien, door de optie om bestuursfuncties om niet uit te voeren zou het aannemen van een oudere bestuurder juist worden gestimuleerd (want de opdracht zou voor de uitschrijvingsautoriteit goedkoper kunnen worden uitgevoerd) ten opzichte van een jongere concurrent (die wel een salaris nodig zouden hebben).

Voorbeeld: De procedure betrof (wederom) een leeftijdseis: om als ‘(door de Duitse ziekenfondsen) gecontracteerde tandarts’ te werken mocht je maximaal 68 jaar oud zijn. Dit levert leeftijdsdiscriminatie op. De regel was onder andere bedoeld om een hoog niveau van bescherming van de volksgezondheid te verzorgen door de bekwaamheid van artsen te garanderen.¹⁹⁸ Echter, voor tandartsen die werkzaamheden verrichten buiten het verplichte zorgverzekeringsstelsel bestond geen maximumleeftijd. In die omstandigheden, zo oordeelde het Hof, werd het doel – gezien vanuit de volksgezondheid - niet coherent nagestreefd.¹⁹⁹

2.1.3.1.2 Noodzakelijk

Dit is in essentie een subsidiariteitstoets. Kan het nagestreefde doel ook met minder ingrijpende middelen (even effectief) worden bereikt?

Voorbeeld: Een technische universiteit wilde de man/vrouw verhouding verbeteren door sollicitaties op nieuwe academische functies de eerste zes maanden alleen open te stellen voor vrouwen voor een periode van vijf jaar. Alleen bij ‘uitmuntendheid’ zou in deze periode een sollicerende man een kans maken, en slechts voor maximaal 10% van de vacatures in deze periode. De regel gold universiteit breed. Het College oordeelde evenwel dat er sprake was van (direct) onderscheid op grond van geslacht. Weliswaar is er sprake van een wettelijke uitzondering voor maatregelen die pogen ondervonden nadelen verband houdende met de gronden ras of geslacht op te heffen, maar (ook) deze maatregelen dienen proportioneel te zijn. Nu de regel echter voor alle faculteiten van de universiteit gold (ook voor faculteiten waar geen of minder sprake van achterstand was) en er geen minder vergaande maatregelen waren onderzocht (bijvoorbeeld een aanstellingsbonus bij het aannemen van vrouwen) was de maatregel niet noodzakelijk om het doel te bereiken.²⁰⁰

2.1.3.1.3 Evenredig

De evenredigheidstoets is het sluitstuk van de proportionaliteitstoets: de afweging van de in het geding zijnde belangen om te bepalen of ondanks het feit dat de maatregel zowel geschikt als noodzakelijk is voor het na te streven doel, deze niettemin de belangen van de door de maatregel getroffen en onevenredig schaadt.

Voorbeeld: Een ernstig gehandicapte man verloor zijn baan en kreeg op basis van de geldende cao een overbruggingsuitkering. Deze werd ingevolge diezelfde cao echter automatisch stopgezet op het moment dat de persoon in kwestie in aanmerking kwam

¹⁹⁸ Vergelijk CRM, 13 augustus 2019, 2019-82, <https://mensenrechten.nl/en/oordeel/2019-82>.

¹⁹⁹ HvJ, 12 januari 2010, C-341/08, ECLI:EU:C:2010:4 (Domnica Petersen v Berufungsausschuss für Zahnärzte für den Bezirk Westfalen-Lippe), para. 53ff.

²⁰⁰ CRM, 2 juli 2020, 2020-53, <https://mensenrechten.nl/nl/oordeel/2020-53>.

voor een (vervroegd) pensioen. Het Duitse systeem was echter zo ingericht dat ernstig gehandicapten over het algemeen 1 tot 3 jaar eerder recht hadden op een vervroegd pensioen dan niet-gehandicapten. Zo ook in het geval van verzoeker. Hij werd hierdoor benadeeld nu zijn uitkering gunstiger was dan zijn pensioen. Het Hof van Justitie overwoog dat de ‘non-cumulatieregels’ weliswaar passend was (de overbruggingsuitkering, net als het pensioen, beoogt economische zekerheid te garanderen), maar dat de regel onvoldoende rekening hield met de specifieke belangen van gehandicapten. In het bijzonder wees het Hof op het risico dat zij door hun handicap onontkoombare financiële behoeften hebben en/of dat zij die behoeften met de jaren zien toenemen. Het ondervonden financieel nadeel als gevolg van de regel – die onafhankelijk van de wil van de persoon om door te werken in werking trad (ook als deze geen vervroegd pensioen heeft aangevraagd) – tastte de legitieme belangen van verzoeker dan ook buitensporig aan.²⁰¹

Voorbeeld: Een school weigerde een vrouw met haar assistentiehond toegang tot het schoolgebouw omdat de school in het nieuwe schooljaar een leerling had die ernstig allergisch was voor honden. Hierdoor kon de vrouw haar zoon niet naar zijn les begeleiden. De school had verschillende alternatieven onderzocht en over en weer gesprekken gevoerd, maar uiteindelijk bleek er geen oplossing mogelijk waarbij de assistentiehond zowel te allen tijde de vrouw kon vergezellen en tegelijkertijd de betrokken leerling niet werd blootgesteld aan een eventuele allergische reactie. De school liet het belang van de leerling zwaarder wegen, met een beroep op de gezondheidsexceptie vervat in Artikel 3 lid 1 sub a WGBH/CZ. Het College oordeelde dat de school terecht de toegang weigerde: het belang van het kind om zonder gezondheidsrisico's naar school te kunnen gaan woog zwaarder dan het belang van de vrouw haar zoon naar zijn les te vergezellen, gelet op de functie van de school.²⁰²

2.1.3.2 De betrokken grond.

Er is daarnaast een zekere hiërarchie tussen de verschillende gronden: onderscheid op bijvoorbeeld ras of geslacht – mits vastgesteld – zijn inherent meer verdacht dan onderscheid op bijvoorbeeld leeftijd.²⁰³ Dit uit zich in een strengere toets op de rechtvaardiging en proportionaliteit van de maatregel.

2.2 Richtsnoeren en codes of conduct

De afgelopen jaren hebben nationale en internationale bedrijven, overheden en organisaties een grote hoeveelheid aan ethische richtlijnen ontwikkeld voor AI, waarbij doorgaans een lijst met principes of waarden wordt gepresenteerd die centraal zouden moeten staan bij de ontwikkeling en het gebruik van AI. In hun systematische analyse van het landschap van AI richtlijnen identificeren Jobin et al. elf veel voorkomend principes:

²⁰¹ HvJ, 19 september 2018, C-773/18 (Surjit Singh Bedi v Bundesrepublik Deutschland, Bundesrepublik Deutschland in Prozessstandschaft für das Vereinigte Königreich von Großbritannien und Nordirland).

²⁰² CRM, 10 mei 2019, 2019-42, <https://mensenrechten.nl/en/oordeel/2019-42>.

²⁰³ Council of Europe, Guide to Article 14 of the European Convention on Human Rights and on Article 1 of Protocol No. 12 to the Convention, (Council of Europe, 2020), para. 71. Zie J. Gerards, ‘The Margin of Appreciation Doctrine, the Very Weighty Reasons Test and Grounds of Discrimination’, in M. Balboni (ed.), *The European Convention on Human Rights and the Principle of Non-discrimination*, (Editoriale Scientifica, 2017).

- rechtvaardigheid, fairness en gelijkheid;
- transparantie;
- non-malificence;
- verantwoordelijkheid en verantwoording;
- privacy;
- beneficence;
- vrijheid en autonomie;
- vertrouwen;
- duurzaamheid;
- waardigheid;
- solidariteit.²⁰⁴

De auteurs scharen non-discriminatie onder de noemer van ‘rechtvaardigheid, fairness en gelijkheid’. Deze categorie omvat echter meer dan discriminatie; diverse richtlijnen noemen ook waarden als: consistentie, inclusie, gelijkheid, gelijkwaardigheid, (non-)bias, diversiteit, pluraliteit, toegankelijkheid, omkeerbaarheid, remedie, *redress*, en betwistbaarheid. Hoewel breder in opzet dan louter het tegengaan van discriminatie, bevatten de diverse documenten met AI richtlijnen waardevolle ideeën over hoe biases te voorkomen, uitlegbaarheid te verankeren en fairness te bevorderen. Tegelijkertijd zijn deze richtlijnen door hun brede opzet vrij generiek van aard en geven ze weinig concrete handvatten voor ontwikkelaars om discriminatie tegen te gaan.²⁰⁵

Naast het scala aan AI richtlijnen is er daarom inmiddels ook een keur van codes of conduct (COC), impact assessments, standaarden en andere instrumenten ontwikkeld om de genoemde principes te operationaliseren en concreter te maken. Deze documenten vormen een belangrijke bron voor het ontwikkelen van een handreiking met AI-systeemprincipes omdat ze doorgaans zijn opgesteld in samenspraak met systeemontwikkelaars, data-analisten en anderen die direct met de regels aan de slag moeten. De ontwikkeling van de snelgroeiende reeks van instrumenten staat echter nog in de kinderschoenen en waar het nog aan ontbreekt zijn systematisch studies van hoe zij in de praktijk werken en hoe effectief ze zijn.

Er bestaan verschillende soorten aanpakken en strategieën voor het tegengaan van bias. Rovatos et al. onderscheiden vijf categorieën:

1. Statische aanpakken en software toolkits:

²⁰⁴ A. Jobin, M. Ienca & E. Vayena (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. Het merendeel van de richtlijnen uit de VS, gevolgd door de EU, de UK en Japan. Voor een ander overzicht van de diverse AI richtlijnen zie: J. Fjeld et al. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020-1). Fjeld et al. identificeren 8 veel voorkomende principes: privacy, accountability, veiligheid en beveiliging, transparantie en uitlegbaarheid, Fairness en non-discriminatie, Menselijke controle over technologie, professionele verantwoordelijkheid, menselijk waarden. Fairness is volgens Fjeld et al. een overkoepelend principe dat vaak verder uitgewerkt wordt in principes als ‘representatieve en hoogwaardige data’, ‘inclusiviteit in impact’ en ‘inclusiviteit in ontwerp’.

²⁰⁵ M. Rovatos, B. Mittelstadt & A. Koene (2019). Landscape Summary: Bias In Algorithmic Decision-Making: what is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?; E. Ntoutsis et al. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1356.

Het gaat hier om methodes en technieken om bias te detecteren of te voorkomen in data sets, in AI modellen of in de uitkomsten van modellen.²⁰⁶

2. **Discursieve raamwerken, zelfevaluatie tools en leermateriaal:**

Deze methoden zijn doorgaans minder technisch van aard en zijn bedoeld om ontwikkelaars maar ook gebruikers, beleidsmakers en andere betrokkenen te helpen om bias te herkennen, voorkomen en mitigeren. Impact assessments, vragenlijsten, evaluatie kaarten en bijsluiters zijn voorbeelden van methodes die deze categorie van strategieën.

3. **Documentatiestandaarden:**

Deze methodes richten zich op het standaardiseren van beschrijvingen van datasets en modellen. AI algoritmen maken vaak gebruik van veel verschillende soorten data of deels getrainde modellen. Door op een gestandaardiseerde manier te documenteren hoe en waarom datasets tot stand zijn gekomen en welke beslissingen er zijn gemaakt bij het trainen van modellen krijgen ontwikkelaars meer inzicht in de data en modellen waar zij mee werken en zijn zij beter in staat om bias te detecteren en te mitigeren. Voorbeelden van deze methodes zijn *data sheets* en *model cards*.

4. **Auditing:**

Het gaat hier om methodes die uitkomsten van systemen voor verschillende groepen vergelijken, gebaseerd op verschillende datasets en interacties, om na te gaan of het gebruik van de gebruikte algoritmen leidt tot discriminatie. Voorbeelden zijn enquêtes onder betrokken, A/B testing, niet-invasieve data *scraping* en *crowdsourced auditing* waar gebruikers data verzamelen door te interacteren met het systeem. De toepassing van deze methodes vindt doorgaans achteraf plaats en zegt weinig over hoe de bias in een systeem ontstaat.

5. **De ontwikkeling van technologische standaarden en certificering:**

Verschillende nationale en internationale instituties zijn momenteel bezig met de ontwikkeling van een breed palet aan standaarden voor AI, waar discriminatie, bias en fairness doorgaans ook een onderdeel van zijn. Zo werkt een werkgroep binnen ISO/IEC TR aan een rapport over "Bias in AI systems and AI aided decision making". Op het gebied van certificering zijn er diverse initiatieven die programma's ontwikkelen om duidelijk te maken of systemen zijn getest op bias en dat er maatregelen zijn genomen bias tegen te gaan, zoals de Ethics Certification Programme for Autonomous and Intelligent Systems van de IEEE SA.

Rovasots et al. erop dat de beschreven methoden veelal de nadruk leggen op het technisch deel van de problematiek waar bias en discriminatie uit voortkomen. Ze hebben weinig oog voor de culturele, organisatorisch en politieke context waarin AI-algoritmes worden ontwikkeld en gebruikt. Ook op dit gebied zijn er echter inmiddels meerdere strategieën en methodes ontwikkeld. Een zesde categorie is daarom wenselijk:

6. **Sociotechnische methodes:**

Bij deze methodes wordt er ook gekeken naar de inbedding van de ontwikkeling en het gebruik van AI-algoritme in de bredere context. Zo wijzen sommige initiatieven op het

²⁰⁶ Zie ook: A. Romei & S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5), 582-638.

belang van diverse teams, inspraak van betrokkenen en aandacht voor de machtsverhoudingen en -structuren in het sociotechnisch systeem (b.v. wie beslist over welk algoritme of welke optimalisatie functie worden gebruikt).²⁰⁷

De zes genoemde categorieën kunnen overlappen en de diverse methodes, aanpakken en strategieën kunnen elkaar ook aanvullen. Zo kan een documentatiestandaard bijvoorbeeld ook vragen stellen die reflectie stimuleren op de diversiteit van het team dat de dataset samenstelde.

In de eerste categorie, **statische methodes en software tools**, maken Rovatsos et al., zoals veel anderen, en een onderscheid tussen de verschillende stadia in ontwikkeling van een AI-systeem:²⁰⁸

- **pre-processing gericht op data:** de methodes die vallen onder deze strategie richten zich met name op het samenstellen van evenwichtige datasets en het minimaliseren van bias in de dataset, zoals *instance class modification*, *instance selection*, en *instance weighting*,
- **in-processing gericht op algoritmes:** hier gaat het om methodes die gericht zijn op het aanpassen van het algoritme op zodanige wijze dat het minimaliseren van discriminerend uitkomsten expliciet wordt meegenomen als doel bij het trainen van een model. Methodes die ontwikkelaars daarvoor kunnen gebruiken zijn bijvoorbeeld *classification model adaption*, *regularisation/loss function s.t. constraints*, *latent fair classes*. Een complicerende factor hierbij is dat er voor dit soort methodes vaak beschermde persoonsgegevens nodig hebben om bias te mitigeren.
- **post-processing gericht op ML modellen:** dit type methode is gericht op het beperken van bias na het trainen van het classificatiemodel. Het kan hier gaan om *white-box* methodes die interne aanpassingen maken aan het model, of om *black-box* methodes die de voorspellingen van een model aanpassen. Voorbeelden van methodes zijn *confidence/probability score corrections*, *promoting demoting boundary decisions*, *wrapping a fair classifier on top of a black-box baselearner*.

Zoals gezegd bevinden de ontwikkelingen nog in een beginfase en bestaan er nog geen eenduidige breed gedragen richtlijnen over welke methodes het beste gebruikt kunnen worden en onder welke condities. Om meer inzicht te geven in de verschillende soorten methodes en strategieën wordt hieronder op basis van een quickscan van de literatuur (breed opgevat: wetenschappelijk, bedrijfsinformatie, brochures, presentaties) een aantal relevante vindingen en aanknopingspunten voor de te ontwikkelen handreiking in kaart gebracht. Dit overzicht is uitgesplitst per type organisatie die het desbetreffende instrument heeft ontwikkeld: private sector (2.4.1), overheid (2.4.2), Non-Governmental Organisations (NGO's) (2.4.3), wetenschap (2.4.4) en professionele organisaties en overige (2.4.5). De focus van deze quickscan ligt op instrumenten specifiek ontwikkeld om bias en discriminatie te detecteren en te mitigeren. Tot slot wordt wat algemener gekeken naar de richtsnoeren en aanbevelingen die in de Nederlandse context zijn gedaan (2.4.6).

²⁰⁷ M. Kasy & R. Abebe (2020). "Fairness, Equality, and Power in Algorithmic Decision-Making". Working Paper.

²⁰⁸ Zie ook: S. A. Friedler et al. (2019). A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the conference on fairness, accountability, and transparency (pp. 329-338). E. Ntoutsi et al. (2020). Bias in Data-driven AI Systems--An Introductory Survey. arXiv preprint arXiv:2001.09762.

2.2.1 Private sector

Veel bedrijven in het tech domein ontwikkelen instrumenten om discriminatie en andere onwenselijke gevolgen van AI gebruik tegen te gaan. Grote spelers zoals IBM, Microsoft en Google vertrekken doorgaans vanuit bepaalde AI principes.²⁰⁹ Deze worden dan gebruikt en verder toegespitst om de ontwikkeling van specifieke AI applicaties (zoals bijvoorbeeld gezichtsherkenningstechnologie) te sturen. Hoewel alle bedrijven natuurlijk hun eigen kleuring en invulling geven aan de principes waarop ze hun AI werk baseren, valt toch vooral op hoeveel overeenstemming er is. Met name als het gaat om het tegengaan van discriminatie, wordt door al deze bedrijven *fairness* en *uitlegbaarheid* (of in het verlengde daarvan, transparantie en verantwoording) als kernwaarden gezien.

Principes benoemen is één ding, deze in de praktijk vormgeven is een tweede. Ook voor de praktische implementatie zijn er tal van initiatieven. Zo stelt Microsoft een *Inclusive design* methodologie beschikbaar om ontwikkelaars te trainen en begeleiden in het ontwerpen voor inclusiviteit.²¹⁰ De methodologie omvat een palet aan met name socio-technisch methodes bedoeld voor ontwikkelaars met onder andere leermateriaal over bias in AI en inclusief ontwerpen, *activity cards* en *best practices*.

Google heeft fairness-indicatoren vrijgegeven die zijn ingebed in hun ML-diagnose *What If*-tool. Met deze pre-processing software tool kunnen ontwikkelaars verschillende definities van fairness uitproberen in hun dataset.²¹¹ Daarnaast heeft Google ook meerdere initiatieven lopen waarin het tegengaan van discriminatie centraal staat. Voor ontwikkelaars biedt het onder andere een crash-course fairness, een video over unconscious bias²¹² en een webpagina met resources voor fairness in ML.²¹³ De website adviseert onder meer om meerdere beoordelingsmaatstaven te gebruiken, veel te testen, de ruwe data te onderzoeken, de beperkingen van een model goed te begrijpen en het systeem ook na implementatie te blijven monitoren en aan te passen. Als onderdeel van thinkwithgoogle.com doet het bedrijf een aantal aanbevelingen aan professionals in marketing over data fairness: wees transparant; *test, tune and test again*; ga opzoek naar andere perspectieven (neem mensen aan met verschillende achtergronden aan, nodig het publiek uit om kennis te delen), en stel vragen.²¹⁴

Ook IBM stelt zijn AI Fairness 360 (AIF360) toolkit vrijelijk ter beschikking. Dit is een uitgebreide open source toolkit met diverse pre-, in- en postprocessing beoordelingsmaatstaven om te controleren op ongewenste bias in datasets en ML-modellen, en met state-of-the-art algoritmes om dergelijke bias te verminderen.²¹⁵ Verder omvat de toolkit ook verschillende tutorials, uitleg over de verschillende instrumenten en definities van fairness, artikelen en suggesties voor verdere verdieping. IBM heeft ook een complementaire AI Explainability 360

²⁰⁹ Zie bijvoorbeeld: Microsoft. Microsoft AI principles. <https://www.microsoft.com/en-gb/ai/responsible-ai?activetab=pivot1%3aprimar6>; bezocht 16 juli 2020, IBM (2019). Everyday ethics for Artificial Intelligence. <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>; bezocht 16 juli 2020. hai, S. (June 7, 2018). AI at Google: our principles. <https://www.blog.google/technology/ai/ai-principles/>.

²¹⁰ Microsoft. Inclusive Design. <https://www.microsoft.com/design/inclusive/>.

²¹¹ D. Weinberger. Playing with fairness. <https://pair-code.github.io/what-if-tool/ai-fairness.html>;

²¹² <https://www.youtube.com/watch?v=pHT-ImFXP>

²¹³ <https://ai.google/responsibilities/responsible-ai-practices/>

²¹⁴ <https://www.thinkwithgoogle.com/feature/ml-fairness-for-marketers/#next-steps>

²¹⁵ IBM (2018) AI Fairness 360 Open Source Toolkit.

http://aif360.mybluemix.net/?_ga=2.178706980.1238303720.1594898118-1456494440.1591891979&cm_mc_uid=07399206762615765859038&cm_mc_sid_50200000=43118981594898118410; bezocht op 16 juli.

open source toolkit om inzicht te geven in hoe ML modellen tot voorspellingen komen. Deze toolkit bestaat uit acht algoritmes voor inzichtelijke ML en beoordelingsmaatstaven voor explainability.²¹⁶

Ten slotte is het nog goed op te merken dat naast het formuleren van principes en het ontwikkelen van praktische tools, het ook cruciaal is om dit alles te verankeren in de organisatie. Al deze initiatieven zijn immers vrijwillig en kunnen, als het om wat voor reden dan ook niet meer zo goed uitkomt, eenvoudig aan de kant worden geschoven. Het is daarom van cruciaal belang dat al deze initiatieven voorzien zijn van organisatorische handhavingsmechanismen om het risico op “ethics washing” te voorkomen.²¹⁷ Onderzoek wijst uit dat er op dat vlak werk aan de winkel is (Hagendorff 2020).²¹⁸ Van de meer dan 160 AI richtlijnen die werden verzameld, beschikten er slechts tien over de juiste handhavingsmechanismen.²¹⁹

Kortom, enerzijds is het veelbelovend dat er zoveel strategieën worden ontwikkeld, maar de huidige veelheid aan strategieën en het gebrek aan effectieve handhavingsmechanismen maakt het ook moeilijk om te beslissen welke het beste past. Hoewel er enkele factoren zijn die kunnen wijzen op het succes van deze initiatieven (bijv. betrokkenheid bij de wet, specificiteit, bereik, afdwingbaarheid, iteratie en follow-up) is er nog geen uitgebreid vergelijkend onderzoek gedaan om de effectiviteit van een van deze benaderingen vast te stellen.²²⁰ Voor de ontwikkeling van de AI systeemprincipes is het belangrijk om met deze beperkingen en uitdagingen rekening te houden.

2.2.2 Overheden

Wereldwijd wordt vanuit overheden getracht sturing te geven aan de ontwikkeling van AI. Sommige van deze beleidsdocumenten en richtsnoeren hebben als doel om reguleringsprincipes te vertalen naar de AI context, andere zijn meer bedoeld als een handvat voor AI actoren om beginselen in de praktijk toe te passen.²²¹

2.2.2.1 Verenigd Koninkrijk

Het *Data Ethics Framework* (2018),²²² gepubliceerd door het *Department for Digital, Culture, Media & Sport* van het Verenigd Koninkrijk, benoemt 7 gerelateerde principes om verantwoordelijk datagebruik te faciliteren. Zij hebben daarbij een *Data Ethics Workbook*²²³ ontwikkeld waarbij organisaties een vragenlijst kunnen invullen om ervoor te zorgen dat de applicatie op een goede manier wordt ontwikkeld en ingezet. Bij deze code of

²¹⁶ IBM (no date) AI Fairness Explainability 360 Open Source Toolkit. <https://aix360.mybluemix.net>, bezacht 8 augustus.

²¹⁷ B. Wagner (2018). "Ethics as an escape from regulation: From ethics-washing to ethics shopping." *Being profiling*. Cogitas ergo sum: 84-90.

²¹⁸ T. Hagendorff (2020). "The ethics of AI ethics: An evaluation of guidelines." *Minds and Machines*:1-22.

²¹⁹ <https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/>.

²²⁰ D. Schiff et al. (2020). "What's Next for AI Ethics, Policy, and Governance? A Global Overview." *Proceedings of the AAAI/A Conference on AI, Ethics, and Society*.

²²¹ Janssen verwijst o.a. nog naar assessment tools uit Canada en Denemarken. Zie: H. L. Janssen (2020). *An approach for a fundamental rights impact assessment to automated decision-making*, *International Data Privacy Law*.

²²²

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737137/Data_Ethics_Framework.pdf

²²³ <https://www.gov.uk/government/publications/data-ethics-workbook/data-ethics-workbook>

conduct ligt de nadruk vooral op privacy en gegevensbescherming, maar de volgende vragen zijn relevant voor de te ontwikkelen handreiking:

- Could you clearly explain why you need to use this data to members of the public?
- Does this use of data interfere with the rights of individuals?
- If yes, is there a less intrusive way of achieving the objective?
- Is there a fair balance between the rights of individuals and the interests of the community?
- How interpretable are the outputs of your work?
- How are you explaining how approaches were designed in plain English to other practitioners, policy makers and if appropriate, the public?
- What is the quality of the model outputs, and how does this stack up against the project objectives?
- If using data about people, is it possible that a data science technique is basing analysis on proxies for protected variables which could lead to a discriminatory policy decision?

De Information Commissioner's Office (ICO) heeft ook een draft guideline ter consultatie beschikbaar.²²⁴ Daarin staan drie methodes om discriminatie in AI tegen te gaan:

- Anti-classification is where a model is fair if it excludes protected characteristics when making a classification or prediction. Some anticlassification approaches also try to identify and exclude proxies for protected characteristics (eg attendance at a single-sex school). This can be impractical as removing all possible proxies may leave very few predictively useful features. Also, it is often hard to know whether a particular variable (or combination of variables) is a proxy for a protected characteristic without further data collection and analysis.
- Outcome / error parity compares how members of different protected groups are treated by the model. With outcome parity, a model is fair if it gives equal numbers of positive or negative outcomes to different groups. With error parity, a model is fair if it gives equal numbers of errors to different groups. Error parity can be broken down into parity of false positives or false negatives (see section 2.3 on statistical accuracy for more details). Equal calibration – calibration measures how closely the model's estimation of the likelihood of something happening matches the actual frequency of the event happening. According to 'equal calibration' a model is fair if it is equally calibrated between members of different protected groups. For instance, if a classification model sorts loan applicants into those with a low, medium, or high chance of repayment, there should be equal proportions of male and female applicants who actually repay within each risk category. This doesn't mean there should be equal proportions of men and women across different risk categories. For instance, if women have actually had higher repayment rates than men, there may be more women than men in the low risk category.

2.2.2.2 EU

Binnen de EU heeft de High level expert group of AI trustworthy AI in 2019, na een uitgebreid consultatieproces, richtlijnen (de Guidelines) gepubliceerd die momenteel verder worden uitgewerkt en geoperationaliseerd door middel van verder consultatie en pilots. Op basis van de fundamentele rechten en ethische principes in de Guidelines, heeft de HLEG een set van zeven kern vereisten geformuleerd. Nummer vijf is *Diversity, non-discrimination and fairness*. Deze

²²⁴ <https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf>

vereiste werkt de HLEG verder uit in drie normen:

- **Vermijd unfair bias** door bijvoorbeeld toezichtprocessen te ontwikkelen om het doel van het systeem, de beperkingen, vereisten en beslissingen te analyseren op een transparante manier. Maar ook door meer mensen vanuit verschillende achtergronden, culturen, en disciplines aan te nemen.
- **Toegankelijkheid en universeel ontwerp** zodat iedereen de producten kan gebruiken, inclusief mensen met een handicap. Universeel-ontwerp principes moeten worden gebruikt om zoveel mogelijke gebruikers te kunnen bedienen.
- **Participatie van betrokkenen** tijdens en na het ontwikkelproces door bijvoorbeeld werknemers te betrekken bij het ontwikkelen van een systeem.²²⁵

Als onderdeel van piloting fase heeft de HLEG ook een Evaluatie lijst opgesteld voor partijen die de Guidelines willen gebruiken bij hun ontwikkelproject. De lijst bestaat uit een aantal vragen voor elk van de zeven vereisten. Bij vereiste vijf zijn dat vragen over de drie normen. Zo is een van de vragen: “Did you ensure an adequate working definition of “fairness” that you apply in designing this AI system?”. De vervolgvraag is: “Is your definition commonly used? Did you consider other definitions before choosing this one?”.

2.2.2.3 Verenigde Staten

Meer algemene AI-principes zijn ook geïntroduceerd door de Amerikaanse overheid.²²⁶ Tien principes om AI gebruik door de overheid in goede banen te leiden zijn opgenomen in een factsheet, waarbij één principe specifiek ingaat op fairness en non-discriminatie.

“Agencies should consider in a transparent manner the impacts that AI applications may have on discrimination. AI applications have the potential of reducing present-day discrimination caused by human subjectivity. At the same time, applications can, in some instances, introduce real-world bias that produces discriminatory outcomes or decisions that undermine public trust and confidence in AI. When considering regulations or non-regulatory approaches related to AI applications, agencies should consider, in accordance with law, issues of fairness and non-discrimination with respect to outcomes and decisions produced by the AI application at issue, as well as whether the AI application at issue may reduce levels of unlawful, unfair, or otherwise unintended discrimination as compared to existing processes.”

Een uitleg hoe dit in de praktijk vorm te geven ontbreekt echter.

2.2.3 NGOs

Wereldwijd zijn ook niet-gouvernementele organisatie op de wagen gesprongen om bij te dragen aan een verantwoordelijke omgang met AI. Door de band genomen richten zij zich op het ontwikkelen van praktische instrumenten om organisaties te helpen bij het verantwoord ontwikkelen en inzetten van AI

2.2.3.1 ECP

²²⁵ <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Diversity>

²²⁶ Memorandum for the heads of executive departments and agencies <https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>; bezocht op 17 juli 2020.

Zo is in Nederland door het ECP een AI Impact Assessment ontwikkeld (AIIA).²²⁷ Impact assessments zijn instrumenten die ingezet kunnen worden om in kaart te brengen wat de mogelijke juridische, ethische en maatschappelijke gevolgen zijn bij het ontwikkelen en gebruiken van een AI gedreven applicatie. De AIIA bestaat uit 8 stappen: Bepaal de noodzaak voor het doen van een AIIA (1), beschrijf de AI-toepassing (2), beschrijf de baten van de AI-toepassing (3), is het doel en de wijze waarop het doel wordt bereikt ethisch en juridisch verantwoord? (4), is de toepassing betrouwbaar, veilig en transparant? (5), afweging en beoordeling (6), vastlegging en verantwoording (7), evalueer periodiek (8). Deze AIIA wordt gecombineerd met een uitgebreide Gedragscode Artificiële intelligentie.

Voor de te ontwikkelen handreiking AI systeemprincipes is stap vier van de AIIA een belangrijk onderdeel (is het doel en de wijze waarop het doel wordt bereikt ethisch en juridisch verantwoord?). Het vertrekpunt is hierbij altijd het juridisch kader waaraan voldaan dient te worden, maar dit kader kan onvolledig of ontoereikend zijn voor een grondige afweging. Om vast te kunnen stellen of meer inzet nodig is (technisch, organisatorisch of anderszins) om discriminatie (maar ook andere negatieve gevolgen) tegen te gaan moeten drie zaken in kaart worden gebracht volgens de AIIA:

1. Welke actoren zijn betrokken bij en/of worden geraakt door mijn toepassing van AI?
2. Welke waarden en belangen spelen een rol in de context van mijn toepassing van AI?
3. Zijn deze waarden en belangen geconcretiseerd in wet- en regelgeving?

2.2.3.2 AI Now Instituut

Een andere relevante impact assessment is de Algorithmic Impact Assessment (AIA)²²⁸ van het AI Now Institute. De AIA kent vijf stappen en is, anders dan de AIIA die breed inzetbaar is, specifiek gericht op publieke actoren die geautomatiseerde besluitvoering en AI willen gebruiken. Voor de te ontwikkelen handreiking is het interessant dat de AIA met name evalueert wat de mogelijke impact is op fairness, rechtvaardigheid en bias voor gemeenschappen die worden geraakt door de AI. Relevant is ook het belang dat gehecht wordt aan het betrekken van externe reviewers (zowel professionals als ook de gemeenschappen op welke de AI tool gericht is) om doorheen de tijd de impact van de AI toepassing te volgen. Ten slotte wordt ook het belang onderstreept om *due process* mechanismen te ontwikkelen voor individuen en groepen die AI gedreven besluiten of unfair, biased of op een andere wijze schadelijke AI systemen willen aanvechten.

2.2.3.3 CPAIS

Het Collaborations Between People and AI systems framework (CPAIS)²²⁹ heeft dan weer een vragenlijst ontwikkeld om specifiek de samenwerking tussen mensen en AI toepassingen in goede banen te leiden. De vragenlijst ziet vooral toe op de AI-in-use, een belangrijk, en vaak over het hoofd gezien, deel van de AI life cycle waar ook discriminatie kan plaatsvinden, bijvoorbeeld door het optreden van de eerder genoemde *perception bias* of *technology bias*. Belangrijke vragen zijn dan bijvoorbeeld:

²²⁷ <https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf>

²²⁸ <https://ainowinstitute.org/aiareport2018.pdf>

²²⁹ <https://www.partnershiponai.org/wp-content/uploads/2019/09/CPAIS-Framework-and-Case-Studies-9-23.pdf>

- Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
- Are the human and the AI system's goals aligned?
- How much agency does the human have? The AI system? [None, limited, moderate, high, full]
- How significant are the consequences should the AI fail to perform as designed/expected? What are those consequences? [Low, moderate, high]
- How significant are the benefits of the AI to the users should it perform as designed/expected? What are those benefits? [Low, moderate, high]
- What are the potential consequences and benefits of the outcome of the collaboration?
- What might be the broader impacts of the human-AI collaboration?
- Who is the main party or individual assessing the nature and effectiveness of the human-AI collaboration?
- Are assessments of the human-AI collaboration's outcome subjective or objective?
- Can the AI system communicate its confidence levels to a human?
- How does the AI system communicate its decision-making process and inputs to that decision-making process to the human?

2.2.3.4 AI Global

AI Global is een non-profitorganisatie die zich inzet voor een wereld met “betrouwbare, veilige en eerlijke technologie”.²³⁰ Ze zetten in op enerzijds het produceren van gebruiksvriendelijke toolkits voor verantwoordelijk AI ontwikkeling en gebruik. Deze toolkits komen tot stand in nauwe samenwerking met mensen in het veld. Deze samenwerking stelt AI Global anderzijds ook in staat om overheden en instanties die werken met AI te voorzien van onderbouwde adviezen. AI Global heeft tot nu toe drie tools opgeleverd:

- *Responsible Community AI portal*: een open platform dat rapporten, standaarden, modellen, overheidsbeleid, open datasets en open-source software combineert om leden te helpen beter door het AI-landschap te navigeren en rechtstreeks contact te leggen met de experts die deze waardevolle bronnen creëren.
- *Responsible AI design assistant*: een open, online beoordeling om leden te helpen te anticiperen op problemen en hun AI-systeem toekomstbestendig te maken. Deze tool brengt onderzoek en best practices uit de branche samen om ontwerpers, ontwikkelaars en producteigenaren te helpen de belangrijkste AI-uitdagingen in gedachten te houden, waaronder gegevensrechten en gebruik, privacy, beveiliging, uitlegbaarheid, eerlijkheid, bias en robuustheid.
- *Responsible AI check*: een onafhankelijk certificeringsmerk om consumenten te helpen gemakkelijk vertrouwde AI-services en -tools te identificeren. Het certificeringsmerk is gebaseerd op een AI Trust Index en wordt continu afgestemd op principes, best practices en standaarden voor verantwoord ontwerp, implementatie en gebruik van AI.

2.2.3.5 Partnership on AI (PAI)

Het Partnership on AI (PAI) is een Multi stakeholder organisatie die academici, onderzoekers, maatschappelijke organisaties, bedrijven die AI-technologie bouwen en gebruiken samenbrengt

²³⁰ <https://ai-global.org/about/>; bezocht 17 juli 2020.

om de impact van AI beter te begrijpen.²³¹ Het partnerschap is opgericht om *best practices* op het gebied van AI-technologieën te bestuderen en te formuleren, om het begrip van AI bij het publiek te vergroten en om te dienen als een open platform voor discussie en betrokkenheid over AI en zijn invloeden op mens en samenleving. “fair, transparant, en accountable AI” is één van de zes thema’s waarop zij werken.

Een relevant PAI project voor de ontwikkeling van de systeemprincipes is ABOUT ML (Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles).²³² De focus van dit voortdurend project ligt op het inzichtelijk maken van de AI black box door in te zetten op het documenteren van ML systems. Het achterliggend idee is dat betere documentatie het mogelijk maakt de juiste vraag op het juiste moment in het AI-ontwikkelingsproces te stellen, waardoor AI teams sneller potentiële problemen kunnen identificeren en passende mitigerende maatregelen kunnen nemen.

2.2.3.6 World Economic Forum

In het Whitepaper van het World Economic Forum ‘How to Prevent Discriminatory Outcomes in Machine Learning’²³³ worden vier centrale principes geformuleerd ten aanzien van het voorkomen van bias in ML:

- Active Inclusion: The development and design of ML applications must actively seek a diversity of input, especially of the norms and values of specific populations affected by the output of AI systems.
- Fairness: People involved in conceptualizing, developing, and implementing machine learning systems should consider which definition of fairness best applies to their context and application, and prioritize it in the architecture of the machine learning system and its evaluation metrics.
- Right to Understanding: Involvement of ML systems in decision-making that affects individual rights must be disclosed, and the systems must be able to provide an explanation of their decision-making that is understandable to end users and reviewable by a competent human authority. Where this is impossible and rights are at stake, leaders in the design, deployment and regulation of ML technology must question whether or not it should be used.
- Access to Redress: Leaders, designers and developers of ML systems are responsible for identifying the potential negative human rights impacts of their systems. They must make visible avenues for redress for those affected by disparate impacts, and establish processes for the timely redress of any discriminatory outputs.

Ook staan er drie best practices in voor organisaties:

1. Identifying human rights risks linked to business operations. We propose that common standards for assessing the adequacy of training data and its potential bias be established and adopted, through a multistakeholder approach.

²³¹ <https://www.partnershiponai.org/about-ml/>.

²³² Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles (ABOUT ML) <https://www.partnershiponai.org/wp-content/uploads/2019/07/ABOUT-ML-v0-Draft-Chapter-1.pdf>.

²³³

http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

2. Taking effective action to prevent and mitigate risks. We propose that companies work on concrete ways to enhance company governance, establishing or augmenting existing mechanisms and models for ethical compliance.
3. Being transparent about efforts to identify, prevent, and mitigate human rights risks. We propose that companies monitor their machine learning applications and report findings, working with certified third-party auditing bodies in ways analogous to industries such as rare mineral extraction. Large multinational companies should set an example by taking the lead. Results of audits should be made public, together with responses from the company.

2.2.4 Wetenschap

Ook in de academische wereld wordt er gewerkt aan middelen om discriminatie in de context van AI tegen te gaan, vaak als onderdeel van een overkoepelend doel om AI systemen op een verantwoorde wijze te ontwikkelen en toe te passen.

2.2.4.1 Inclusive Design Toolkit

Onderzoekers aan Cambridge University hebben een Inclusive Design Toolkit ontwikkeld. Inclusief ontwerp richt zich op de diversiteit van mensen en de impact hiervan op ontwerpbeslissingen. Die diversiteit gaat verder dan alleen maar de vaardigheden die mensen bezitten, maar heeft ook betrekking op de dagelijkse werkelijkheid waarin zij vertoeven, hun achtergrond, ambities, geslacht, etc.²³⁴ Wat interessant is aan deze toolkit is dat het niet louter om datagedreven ontwerpen hoeft te gaan en het dus ook niet louter op data focust als oorzaak van discriminatie. Aangezien AI systemen uiteindelijk deel uitmaken van een interface is ook een bredere kijk op discriminatie in de AI context leerzaam.

2.2.4.2 EAML

Een andere ontwikkeling die positief kan bijdragen aan het tegengaan van discriminatie door AI is de inzet van expert advies bij Machine learning (EAML).²³⁵ Het betrekken van experts uit het domein waarvoor men AI ontwikkelt kan er onder meer aan bijdragen verkeerd gecodeerde variabelen te ontdekken. EAML presteert bovendien goed op zowel accuratesse als interpreteerbaarheid (de regels waarop het model is gebaseerd kunnen van tevoren onderzocht worden). Dit draagt bij aan het vertrouwen in ML.

2.2.4.3 Verificatie mechanismes

Een recent rapport over Trustworthy AI Development heeft als doel om op basis van de vele hierboven beschreven principes ook mechanismes te benoemen die het mogelijk maken om de claims van AI ontwikkelaars ook afdoende te verifiëren.²³⁶ Ze maken daarbij een onderscheid tussen institutionele (1), softwarematige (2) en hardwarematige (3) mechanismen. Op het oog relevante mechanismen voor de te ontwikkelen handleiding zijn:

²³⁴ <http://www.inclusivedesigntoolkit.com/whatis/whatis.html>.

²³⁵ E. D. Gennatas et al. (2020). "Expert-augmented machine learning." *Proceedings of the National Academy of Sciences* 117, no. 9 4571-4577.

²³⁶ M. Brundage et al. (2020). "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *arXiv preprint arXiv:2004.07213*.

- De inzet van *audits* door onafhankelijke partijen (1)
- Bias en veiligheid *bounties* (1)
- Uitlegbaarheid en documentatie (2)
- *Compute support* voor universitaire onderzoekers (om claims over grootschalige AI-systemen te kunnen evalueren) (3)

2.2.4.4 Datasheets for Data Sets

Om onbedoelde biases en misbruik van datasets te voorkomen, zijn er de *Datasheets for Data Sets* ontwikkeld.²³⁷ Het achterliggende idee is dat voor elke dataset wordt beschreven hoe die tot stand is gekomen en wat de sterke en zwakke punten van de dataset zijn. Dit wordt bewerkstelligd door een document bij te houden welke is gebaseerd op zeven categorieën vragen:

- Motivatie voor de creatie van de dataset
- Compositie van de dataset
- Hoe is de data verzameld
- Waar kan de dataset wel/niet voor gebruikt worden
- Reflectie over mogelijke distributie van de dataset
- Onderhoud van de dataset
- Impact en uitdagingen

2.2.4.5 Factsheet voor conformiteitsverklaringen van AI leveranciers

Conformiteitsverklaringen zijn (vaak niet-juridisch verplichte) documenten die leveranciers gebruiken om inzichtelijk te maken hoe een product tot stand is gekomen, hoe het is getest, wat de te verwachten performance is, etc.²³⁸ Een factsheet is ontwikkeld toegespitst op AI leveranciers en te gebruiken door consumenten/burgers. Een aantal van de items die deel uitmaken van de factsheet en relevant zijn voor de te ontwikkelen systeemprincipes:

Waarvoor is de dienst bedoeld?

- Op welke datasets is de dienst getest?
- Kent u mogelijke voorbeelden van vooringenomenheid, ethische kwesties of andere veiligheidsrisico's als resultaat van het gebruik van de dienst?
- Zijn de outputs verklaarbaar en/of interpreteerbaar?
- Voor elke dataset die door de service wordt gebruikt: werd de dataset gecontroleerd op bias? Welke inspanningen zijn gemaakt om ervoor te zorgen dat het eerlijk en representatief is?
- Wordt er biasdetectie en -herstel uitgevoerd?
- Wat zijn de verwachte prestaties op ongeziene gegevens of gegevens met verschillende distributies?

2.2.4.6 Modelkaarten

Modelkaarten zijn korte documenten die worden afgegeven bij getrainde ML modellen en bieden een benchmarkevaluatie onder verschillende omstandigheden, zoals tussen

²³⁷ T. Gebru (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

²³⁸ M. Hind et al. (2018). Increasing Trust in AI Services through Supplier's Declarations of Conformity. <https://arxiv.org/abs/1808.07261>

verschillende culturele, demografische of fenotypische groepen (bijv. ras, geografische locatie, geslacht, Fitzpatrick-huidtype) en intersectionele groepen (bijv. leeftijd en ras, of geslacht en Fitzpatrick-huidtype) die relevant zijn voor de beoogde toepassingsdomeinen.²³⁹ Modelkaarten geven ook de context weer waarin modellen moeten worden gebruikt, details van de prestatie-evaluatieprocedures en andere relevante informatie. De categorieën die worden meegenomen in de modelkaarten zijn:

- Basisinfo over het model
- Bedoeld gebruik
- Welke factoren worden meegenomen in het model
- Metrics
- Evaluatie data
- Training data
- Quantitatieve analyse

2.2.4.7 Data statements voor Natural Language Processing

Door middel van data statements kan het NLPveld beginnen met het aanpakken van kritische wetenschappelijke en ethische kwesties die het gevolg zijn van het gebruik van gegevens van bepaalde populaties bij de ontwikkeling van tech voor andere populaties.²⁴⁰ De data statements moeten met name een manier zijn om bias en discriminatie in NLP tegen te gaan. Categorieën die worden meegenomen in de Data Statements zijn:

- Curation rationale (welke teksten zijn meegenomen en welke niet, met welk doel? Etc.)
- Taalvariatie
- Sprekersdemografie
- Annotatiedemografie
- Spreeksituatie
- Tekstkaracteristieken
- Opnamekwaliteit
- Overige
- Herkomst appendix (voor samengestelde datasets)

In hun analyse van verschillende fairness interventie methods, benadrukken Friedler et al. dat als er meerdere manieren zijn om een data set te bewerken om meer training data te krijgen voor het algoritme, dan moeten er ook meer performance beoordelingsmaatstaven zijn voor meer dan een van de mogelijke keuzes.²⁴¹ Bovendien stellen zij dat als algoritmes vergeleken worden met elkaar, dat er dan zorggedragen moet worden dat ze vergeleken worden op dezelfde voorbereikte data. Tot slot wijzen zij op het belang van het verantwoorden van training instabiliteit. Het is niet genoeg om performance te laten zien op een enkele training-test split.

²³⁹ M. Mitchell (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM. <https://arxiv.org/abs/1810.03993>

²⁴⁰ E. M. Bender & B. Friedman (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604. <https://aclweb.org/anthology/papers/Q/Q18/Q18-1041/>

²⁴¹ S. A. Friedler (2019). A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019. 329–338. <https://doi.org/10.1145/3287560.3287589>

De auteurs adviseren om het succes en de stabiliteit van algoritme te baseren op een beperkt aantal randomized training-test splits.

2.2.5 Professionele organisaties en overige

Vanzelfsprekend zijn ook professionele organisaties gericht op de AI professional zelf aan zet om ervoor te zorgen dat in de verschillende betrokken beroepsgroepen er richtlijnen en andere instrumenten worden ontwikkeld om de professional bij te staan in het maken van goede keuzes. Een belangrijk middel daartoe zijn *codes of conducts*. Een *code of conduct* (CoC) is een document waarin organisaties –zoals bedrijven of beroepsverenigingen–, richtlijnen voor verantwoordelijk gedrag vastleggen. Zo een document maakt het mogelijk om te expliciteren wat de normen en waarden zijn van een organisatie, het kan de accountability en professionaliteit van een organisatie vergroten en bijdragen aan het vertrouwen van burgers en klanten in de organisatie en de geleverde AI-diensten.²⁴²

2.2.5.1 Professional Data Scientist

De *Oxford-Munich Code of Conduct Professional Data Scientist* - een gezamenlijk initiatief van de Duitse energieleverancier E.ON en Oxford University - is een goed voorbeeld van een COC voor data scientists en AI professionals. Deze CoC is tot stand gekomen na verschillende, internationale rondetafel-bijeenkomsten met experts en mensen in het veld en kent 7 punten die gedetailleerd zijn uitgewerkt: lawfulness (1), competence (2), dealing with data (3), algorithms and models (4), transparency, objectivity and truth (5), working alone and with others (6), (extra) upcoming ethical challenges (7).²⁴³

Op verschillende punten worden strategieën aangedragen die discriminatie tegen moeten gaan. Onder (1) wordt naast privacy -en gegevensbescherming expliciet verwezen naar de verantwoordelijkheid van de data scientist geen attributen op te nemen die in strijd zijn met non-discriminatiewetgeving. Onder (3) worden handvatten geboden om biases tegen te gaan, zoals het werken met *internal log-books* waarin onder andere wordt bijgehouden of er potentieel een bias aanwezig is in de verzamelde data, hoe de data wordt opgeschoond en gebruikt en welke *features* mogelijk kunnen fungeren als surrogaat voor attributen (bijvoorbeeld: geslacht, afkomst, religie) die een impact kunnen hebben op de mensenrechten. Onder (4) wordt verwezen naar de verantwoordelijkheid van de data scientists om verschillende technische biases tegen te gaan in de ontwikkeling van algoritme en modellen door diverse in- en postprocessing verantwoordelijkheden te benoemen, zoals het onderzoeken van de data en afwegingen maken tussen accuracy en explainability. Maar de code wijst ontwikkelaars ook op hun verantwoordelijkheid om zicht te houden op het grotere plaatje voorbij de technische maatstaven en te begrijpen hoe een model in de praktijk gaat werken. Ook wijst de auteur op het belang van auditing bij het gebruik van voor-getrainde modellen van derden. Onder (5) wordt transparantie expliciet als een verplichting voor de data scientist genoemd. Hier wordt onder andere aangegeven dat ervoor gekozen moet worden transparantie zo ruim mogelijk op te vatten (begrensd door wat juridisch is toegestaan) en dat er alleen objectieve statements mogen worden opgenomen van data scientists in levensamenvattingen. Er wordt nadrukkelijk gesteld dat de kwaliteit van de resultaten transparant moet zijn en hiervoor gebruik moet gemaakt worden van een standaard framework (gebaseerd op best practices in het veld).

²⁴² I. van de Poel (2011). *Ethics, technology, and engineering: An introduction*. John Wiley & Sons.

²⁴³ <http://www.code-of-ethics.org/code-of-conduct/#conduct-34>

2.2.5.2 CPAIS

Het *Collaborations Between People and AI systems framework* (CPAIS) levert een vragenlijst om de samenwerking tussen personen en AI toepassingen in goede banen te leiden.²⁴⁴ De vragenlijst die daarvoor is ontwikkeld ziet vooral toe op de *AI-in-use*, een belangrijk, en vaak over het hoofd gezien, deel van de *AI life cycle* waar ook discriminatie kan plaatsvinden, bijvoorbeeld door het optreden van de eerder genoemde perception bias of technology bias. Belangrijke vragen zijn dan bijvoorbeeld:

- Is the AI fixed once deployed or evolving over time via model updates/continual interaction?
- Are the human and the AI system's goals aligned?
- How significant are the consequences should the AI fail to perform as designed/expected?
- Who is the main party assessing the nature and effectiveness of the human-AI collaboration?
- Can the AI system communicate its confidence levels to a human?
- How does the AI system communicate its decision-making process and inputs to the human?

Een voorbeeld van een certificatieprogramma ontwikkeld door een internationale samenwerking tussen verschillende partijen is de *Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)*. De ECPAIS is een initiatief van IEEE-SA, in samenwerking met onder andere het Finse Ministerie van Financien, de Espoo, Wenen en New York, UN organisaties zoals Unicef en High-level Panel on Digital Cooperation, en industriepartners zoals EY en Accenture. Het doel van ECPAIS is “to create specifications for certification and marking processes that improve transparency and accountability, and help to reduce algorithmic bias in AI systems”.²⁴⁵ Het is echter vooralsnog onduidelijk wat deze specificaties inhouden.

2.2.5.3 P7000

The IEEE SA heeft ook een project geïnitieerd over algorithmic bias (P7003) onder haar *Ethically aligned design program* (P7000).²⁴⁶ Het doel van de P7003 standaard is om ontwikkelaars een *development framework* te geven om bias te voorkomen. Het framework zal onder andere bestaan uit een set richtlijnen over hoe systemen te ontwikkelen volgens een duidelijk methodologie, hoe stakeholders te betrekken, hoe doelen voor het gebruik van het systeem te rechtvaardigen en hoe de onderliggende principes van het systeem te valideren. Daarnaast zal het framework ook handvatten bieden om met bias om te gaan, zoals hulp bij het kiezen van methoden, benchmark procedures, of richtlijnen voor het communiceren over de mogelijkheden en beperkingen van het systeem. Het document zal bestaan uit drie delen: 1) fundamentele secties, 2) concrete adviezen en 3) use cases. De fundamentele secties zijn bedoeld om meer inzicht te geven in wat bias is en zullen elementen bevatten zoals ‘Taxonomy of Bias’, ‘Legal frameworks related to Bias’, ‘Psychology of Bias’ en ‘Cultural context of Bias’. De concrete adviezen zullen over algoritmische systeem ontwerp en implementatie gaan en zullen elementen bevatten zoals: ‘Algorithmic system design stages’, ‘Person categorizations and identifying of affected groups’, ‘Representativeness and balance of

²⁴⁴ <https://www.partnershiponai.org/wp-content/uploads/2019/09/CPAIS-Framework-and-Case-Studies-9-23.pdf>

²⁴⁵ <https://standards.ieee.org/industry-connections/ecpais.html>

²⁴⁶ <https://standards.ieee.org/project/7003.html> en <https://ethicsinaction.ieee.org>

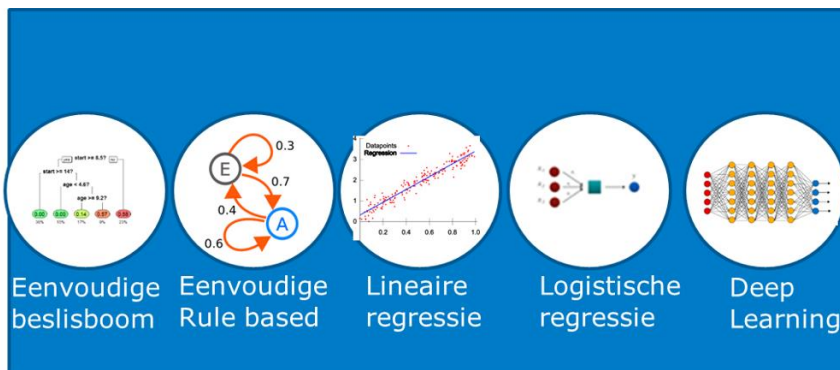
testing/training/validation data’, ‘System outcomes evaluation’, ‘Evaluation of algorithmic processing’, ‘Assessment of resilience against external biasing manipulation’, ‘Assessment of scope limits for safe system usage’ and ‘Transparent documentation’.²⁴⁷

2.2.6 Nederlandse context

Ook binnen de Nederlandse context is een aantal relevante richtlijnen verschenen. Het gaat te ver om die allemaal in detail te behandelen. In plaats daarvan zullen drie kaders uitgebreid worden besproken.

2.2.6.1 Ministerie van Justitie en Veiligheid

Het Ministerie van Justitie en Veiligheid heeft een Richtlijn voor het toepassen van algoritmes door overheden uitgegeven.²⁴⁸ De Richtlijn maakt een onderscheid tussen verschillende type algoritmes. Deze kunnen variëren van een eenvoudige beslisboom met een beperkt aantal variabelen tot complexe en zelflerende algoritmes, zoals zogenaamde ‘*machine learning*’ of ‘*deep learning*’ algoritmes. Deze laatste twee varianten schaaft de Richtlijn onder de categorie kunstmatige intelligentie.



De Richtlijn geeft op zeven punten meer duiding:

1. Bewustzijn risico's: Specifieke aandachtspunten die hierbij worden genoemd:
 - a. Test het algoritme op basis van test cases of scenario's en evolueer test cases periodiek en elke keer als de software verandert om te voorkomen dat nieuwe fouten ontstaan dan wel functionaliteit onbedoeld wordt aangepast.
 - b. Houd rekening met het feit dat een toename van de gegevens in de tijd van invloed kan zijn op de uitkomsten van een algoritme en corrigeer daar zo nodig op.
 - c. Er moeten binnen de grenzen van de wetgeving controlemechanismen worden opgebouwd die specifiek toetsen of er geen sprake is van discriminatie of stigmatisering.
 - d. Onderzoek de kwaliteit van databronnen en breng in kaart welke beperkingen het gebruik van een databron kent, bijvoorbeeld doordat er rechten zijn gevestigd op de databron of op het algoritme en analysemethode. Een ander voorbeeld is dat een databron waarvan de kwaliteit van de informatie slechts globaal of veranderlijk is, minder geschikt kan zijn voor een werkwijze die leidt tot het doen van uitspraken over individuen, maar een dergelijke bron kan wel geschikt zijn voor het beschrijven

²⁴⁷ A. Koene, L. Dowthwaite & S. Seth (2018). IEEE P7003™ standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the International Workshop on Software Fairness* (pp. 38-41).

²⁴⁸ <https://zoek.officielebekendmakingen.nl/blg-904102.pdf>

- van groepen. Datakwaliteit is essentieel voor het kunnen uitvoeren van een gedegen analyse en brongegevens bevatten eigenlijk altijd biases en fouten. Beschrijf daarom het doel van de analyse en hoe je omgaat met de fouten in de datasets en uitkomsten.
- e. Maak een bewuste keuze voor data-analyse technieken. Het is bijvoorbeeld niet zondermeer nodig kunstmatige intelligentie methoden in te zetten op data. Vaak zijn min of meer traditionele analysemethoden geschikt om de kwaliteit van bronnen te onderzoeken of om patronen te vinden. Belangrijk is ook of je wilt werken met vooraf bedachte hypothesen die je wilt toetsen d.m.v. data-analyse, of dat je zonder hypothese te werk wilt gaan. In dat laatste geval is het in het algemeen lastiger om een werkwijze te legitimeren.
 - f. Hanteer een standaard methodiek van werken, bijvoorbeeld CRISP DM (Cross Industry Standard Process for Data Mining). Deze methodiek is toepasbaar bij data mining (techniek om patronen in datasets te ontdekken). Wanneer je een hypothese wil toetsen aan de hand van data-analyse is deze methodiek minder van toepassing.
 - g. Hanteer gelet op de aard van gegevensverwerking juridische en beleidsmatige toetsingskaders en gebruik instrumenten als een gegevensbeschermingseffectbeoordeling om risico's voor de privacyrechten van betrokkenen zoveel mogelijk te beperken.
 - h. Ga na of het algoritme per se van causaliteit moet uitgaan. Dat zal in het algemeen gelden voor algoritmes die gebruikt worden voor besluitvorming. Dat impliceert dat in dat geval het gebruik van deep learning minder voor de hand ligt, omdat die techniek in toenemende mate gebruik maakt van correlaties, d.w.z. statistische verbanden.
2. Uitlegbaarheid: Daarbij is het volgens de Richtlijn van belang een onderscheid te maken tussen “technische transparantie” en “uitlegbaarheid”. Onder technische transparantie verstaan we inzicht in de algoritmische methode die wordt toegepast (beslisboom, neurale netwerk), de broncode, hoe het algoritme is getraind, alsook de data, invoervariabelen, parameters en drempelwaarden die worden gebruikt etc. Bij uitlegbaarheid gaat het om het in begrijpelijke taal kunnen uitleggen van de uitkomsten van data-analyses en hoe deze tot stand zijn gekomen. Specifieke maatregelen die hierbij worden genoemd zijn:
 - a. Organiseer de code in modules welke separaat en gecombineerd kunnen worden geëvalueerd.
 - b. Test deze modules op correcte functionaliteit zowel afzonderlijk als in combinatie.
 - c. Leg de gehanteerde analysemethode uit en meet de nauwkeurigheid.
 - d. Leg de input gegevens (brondata/datasets) vast die gebruikt worden en gebruik daarbij enkel relevante data. Documenteer en leg dit vast.
 - e. Beschrijf en onderzoek de kwaliteit van de gebruikte databron(nen).
 - f. Leg de aannames die gehanteerd zijn vast.
 3. Gegevensherkenning: Voorbeelden van handelingen die uit dit principe volgen zijn bijvoorbeeld:
 - a. Leg de trainingsgegevens of andere gebruikte informatie om te komen tot parametrisering vast, zodat het mogelijk is resultaten te reproduceren.
 - b. Maak, indien mogelijk, analyses met betrekking tot de gevolgen die een andere keuze van parameterisering of inzet van trainingsdata (meer of minder, volgorde van aanbieden van trainingsdata) heeft op de resultaten.
 4. Auditeerbaarheid: Als specifieke aandachtspunten worden hierbij genoemd:
 - a. Werk niet met confidentiële algoritmes, maar met open algoritmes, welke toegankelijk zijn voor toezichthouders ter controle en bij voorkeur ook voor experts en burgers. Dat impliceert het volgende:

- i. het algoritme dient niet-confidentieel te zijn;
- ii. het algoritme dient gedocumenteerd te zijn;
- iii. gebruik zoveel mogelijk algoritmes en analysemethoden die wetenschappelijk gevalideerd zijn;
- iv. gebruik indien mogelijk algoritmes die reeds open source zijn of stel deze als open source beschikbaar.

Er kunnen redenen zijn om van bovenstaande richtlijnen op basis van een juiste onderbouwing af te wijken.

- b. Onderbouw keuzes, zoals de keuze voor specifieke algoritmes en gebruikte data.
 - c. Noteer waarnemingen, zoals afwijkingen in de gegevens of onverwachte/onverklaarbare resultaten.
 - d. Gebruik eenvoudige methoden boven complexe methoden daar waar mogelijk. Dit komt ten goede aan de uitlegbaarheid, auditeerbaarheid en beperking van risico's.
 - e. Verifieer zowel voor statistische analysemethoden als complexere methoden de uitkomsten gebaseerd op de specifieke input.
 - f. Lever een gedetailleerde omschrijving van het algoritme en de werking ervan, samen met een machine-controleerbare validatie dat de code overeenkomt met de specificatie.
 - g. Zorg voor reproduceerbaarheid.
- 5. Verantwoording
 - 6. Validatie
 - 7. Toetsbaarheid

2.2.6.2 VNG

VNG's Principes voor Digitale Samenleving is algemener van aard en geeft vier specifieke uitgangspunten:²⁴⁹

1. Van maatschappelijke waarde: Dataverzameling en -gebruik in de openbare ruimte moet ten dienste staan van het maatschappelijk belang en bijdragen aan de leefbaarheid van dorpen en steden.
2. Rechten over data geregeld: Data is open, inzichtelijk en gedeeld, tenzij wet- en regelgeving, veiligheidsrisico's of beschikkingsrechten op de data dit beperken. Een individu heeft beschikkingsrecht op data over hem of haar en beslist of dit gedeeld mag worden met anderen en wat er verder mee gebeurt, tenzij wet- en regelgeving dit beperken.
3. Toegankelijke en veilige digitale infrastructuur: De digitale infrastructuur voor dataverzameling en -(her)gebruik is voor iedereen goed beschikbaar en toegankelijk. De gebruikte technologie is veilig, inzichtelijk, koppelbaar en kent "open interfaces", "open protocollen
4. Partijen verbonden: Marktpartijen, instellingen, overheden en inwoners werken samen waar dat voor hen wenselijk, respectievelijk nodig is. Zij zijn allen leveranciers en/of gebruikers van netwerkvoorzieningen, de apparatuur, de connectiviteit en de 'slimme' diensten.
5. Transparantie centraal: Marktpartijen, instellingen, overheden en inwoners zijn zoveel mogelijk transparant over apparatuur en technologie in de openbare ruimte. Gemeenten stimuleren hen om dit te doen.

²⁴⁹ <https://vng.nl/sites/default/files/2019-11/09a-bijlage-principes-voor-de-digitale-samenleving.pdf>

2.2.6.3 Ethisch kader data-gedreven werken Verzekeraars

Verwijzend naar High-Level Expert Group on Artificial Intelligence heft het Verbond van Verzekeraars een zeer gedetailleerd raamwerk met vereisten en subvereisten neergelegd, dat hier integraal is overgenomen.²⁵⁰

Vereiste voor Verantwoorde AI	Subvereiste	Norm voor verzekeraars
Menselijke autonomie en controle	Inzet van AI	1. Voordat wij data gedreven toepassingen inzetten, voeren wij (al dan niet in het kader van een PARP proces) een adequate compliance beoordeling uit, waarbij wij een bewuste keuze maken met betrekking tot geconstateerde risico's in vergelijking tot meer traditionele technieken en processen. 2. Bij gebruik van data gedreven toepassingen zoals chatbots zullen wij waar nodig vermelden dat de klant met een systeem van doen heeft en niet met een mens, om verwarring of onduidelijkheid hierover te voorkomen.
Technische robuustheid en veiligheid	Cybersecurity	3. Wij zorgen dat ten aanzien van data gedreven toepassingen (inclusief databaseer) passende beveiligingsmaatregelen zijn genomen. 4. Wij zorgen dat data gedreven toepassingen technisch veilig en robuust zijn en alleen op basis van duidelijke kaders en onder toezicht 'zelflerend opereren'.
	Fall-back en algemene veiligheid	5. Indien een data gedreven toepassing niet (langer) technisch veilig of robuust wordt geacht, zullen wij zo spoedig mogelijk maatregelen treffen om ervoor te zorgen dat de toepassing wel voldoet.
	Betrouwbaarheid reproduceerbaarheid	6. Verzekeraars monitoren of gebruikte data gedreven systemen in overeenstemming met vooraf gestelde doelen, doelstellingen en beoogde toepassingen werken.
	Kwaliteit en integriteit van data	7. Wij zorgen voor adequate kwaliteit (waaronder evaluatie van de datakwaliteitscriteria: volledigheid, juistheid, tijdigheid, adequaatheid en representativiteit) van (trainings)data die

²⁵⁰ <https://www.verzekeraars.nl/media/7541/ethisch-kader.pdf>

		gebruikt wordt voor datagedreven toepassingen. 8. Bij gebruik van data gedreven toepassingen maken wij een weloverwogen keuze om wel of geen biometrische gegevens, data gegenereerd uit ‘affective computing’, social media data, online (be)zoekgegevens, locatie- en IoT-data te gebruiken en zullen klanten indien gewenst daarover transparant informeren.
	Toegang tot gegevens	9. Wij zorgen voor verantwoord beheer van data en waarborging van goede data governance.
Privacy en data governance	Respect voor privacy en databescherming	10. Bij gebruik van persoonsgegevens voor data gedreven toepassingen werken wij conform de Algemene Verordening Gegevensbescherming (AVG) en de Uitvoeringswet AVG (UAVG) en de Gedragscode Verwerking Persoonsgegevens Verzekeraars. 11. Voorafgaand aan de inkoop, ontwikkeling en/of ingebruikname van data gedreven toepassingen voeren wij waar dit noodzakelijk blijkt een gegevensbeschermingseffectbeoordeling (een DPIA) uit. 12. Wij kiezen voor data gedreven systemen die zo min mogelijk potentieel gevoelige data of persoonsgegevens verwerken (dataminimalisatie) en/of waarbij de mogelijkheid bestaat om de privacy te vergroten via bijvoorbeeld encryptie, pseudonimisering/ anonimisering of aggregatie. 13. Wij zorgen voor gedegen bescherming van (trainings)data tegen aantasting, vervuiling of hacking
	Menselijke controle	14. Wij zorgen voor adequate training van medewerkers die met data gedreven toepassingen werken, met name ter voorkoming van ‘confirmation bias’ (voorkeur voor bevestiging) en met het oog op behoud van menselijke autonomie
	Menselijk toezicht	15. Het gebruik van data gedreven toepassingen in de praktijk vindt altijd plaats onder adequaat menselijk toezicht en menselijke verantwoordelijkheid,

		bijvoorbeeld door waar nodig AI te hertrainen.
		16. Wij zullen nieuwe technieken eerst testen in een vertrouwde setting, om te vergelijken of foutmarges en andere risico's toenemen ten opzichte van alternatieve methoden en processen.
Transparantie		17. Voordat wij data gedreven systemen inzetten bedenken wij hoe we zo goed mogelijk uitleg kunnen geven aan klanten over de uitkomsten van de toepassing.
		18. Bij de inzet van data gedreven toepassingen zal altijd een beroep gedaan kunnen worden op menselijke tussenkomst en uitleg verkregen kunnen worden door klanten omtrent de uitkomsten bij een toepassing.
Diversiteit, nondiscriminatie en rechtvaardigheid	Voorkomen onrechtvaardige bias	19. Wanneer inbreuk op grondrechten, waaronder ongerechtvaardigde discriminatoire bias in data gedreven toepassingen niet vermeden of uitgesloten kan worden, zullen wij een toepassing niet inzetten
	Toegankelijkheid en inclusief ontwerp	20. Bij de keuze voor gebruik van data gedreven systemen hebben wij oog voor diversiteit en inclusiviteit, in het bijzonder voor mensen die een risico lopen op uitsluiting of benadeling vanwege bijzondere behoeften en/of een beperking.
Maatschappelijk welzijn	Sociale gevolgen	21. Wij zullen de gevolgen van de inzet van datagedreven besluitvorming voor groepen klanten intern monitoren.
	Samenleving en democratie	22. Wij streven ernaar om zo veel mogelijk klanten verzekeraar te houden en zullen klanten die moeilijker of onverzekerbaar dreigen te worden informeren over manieren om risico's te verlagen of alternatieve manieren om risico's af te dekken.
Verantwoording	Controleerbaarheid	23. Wij zorgen voor een intern controle- en verantwoordingsmechanisme voor het gebruik van AI systemen en de gebruikte databronnen. 24. Wij bevorderen de kennis van onze bestuurders en interne toezichthouders ten aanzien van data gedreven toepassingen.

		25. Wij zorgen voor gedegen interne communicatie over het gebruik van data gedreven systemen
	Minimalisering en verslaglegging negatieve gevolgen	26. Wij voeren voor alle data gedreven toepassingen een risico- en effectbeoordeling uit ten aanzien van direct belanghebbenden. 27. Wij bevorderen de deskundigheid van onze medewerkers die werkzaam zijn op het gebied van verantwoording en controle van data gedreven systemen via een programma van educatie. 28. Wij zorgen voor een open cultuur in ons bedrijf waarin medewerkers worden aangemoedigd om ethische afwegingen te maken en een gedegen systeem waarbij (potentiele) negatieve gevolgen van het gebruik van een data gedreven toepassing kunnen worden gemeld en adequaat worden afgehandeld.
	Documentatie van afwegingen	29. Wij zullen gemaakte keuzes met betrekking tot de inzet van datagedreven besluitvorming in ons interne beleid vastleggen, waarbij de doorslaggevende factoren die hebben geleid tot de gemaakte keuzes ook worden vastgelegd
	Klachten	30. Wij informeren klanten over de mogelijkheden om klachten in verband met het gebruik van data gedreven toepassingen kenbaar te maken, allereerst bij het eigen bedrijf en daarna bij aangewezen geschilbeslechtsers.

2.3 Statistische standaarden en methodiek

De WRR heeft er in haar rapport over Big Data op gewezen dat Big Data globaal in drie fasen kan worden verdeeld: verzamelen, analyse en gebruik. Het stelt dat met name de eerste en de laatste fase zijn ingekaderd door bijvoorbeeld het privacy- en gegevensbeschermingsrecht en het discriminatieverbod en recht op gelijke behandeling, maar dat de tweede fase soms buiten sluit blijft, omdat hier niet noodzakelijkerwijs persoonsgegevens worden verwerkt (bijvoorbeeld als de verwerking geschiedt op geaggregeerd niveau en er naar algemene verbanden wordt gezocht). Daarom verwijst de Raad onder meer naar de noodzaak om aansluiting te zoeken bij de diverse principes voor statistiek, aangezien Big Data, Artificial Intelligence en Machine Learning gebruik maken van statistische correlaties en methodiek. Hieronder volgt een korte bespreking van een aantal van de belangrijkste standaarden die ook in Nederland gelden.

De AVG vereist dat statistische beginselen en procedures in acht moeten worden genomen. ‘In order to ensure fair and transparent processing in respect of the data subject, taking into account

the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised []'.²⁵¹

Hierbij kan worden verwezen naar artikel 338 van het Verdrag betreffende de werking van de Europese Unie dat stelt: '1. Without prejudice to Article 5 of the Protocol on the Statute of the European System of Central Banks and of the European Central Bank, the European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall adopt measures for the production of statistics where necessary for the performance of the activities of the Union. 2. The production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality; it shall not entail excessive burdens on economic operators.'²⁵² Deze beginselen zijn onder meer uitgewerkt in de verordening betreffende de Europese statistiek, waarvan artikel 2 bepaalt dat de volgende beginselen in acht moeten worden genomen:²⁵³

- Professional independence: statistics must be developed, produced and disseminated in an independent manner, particularly as regards the selection of techniques, definitions, methodologies and sources to be used, and the timing and content of all forms of dissemination, free from any pressures from political or interest groups or from Community or national authorities, without prejudice to institutional settings, such as Community or national institutional or budgetary provisions or definitions of statistical needs;
- Impartiality: must be developed, produced and disseminated in a neutral manner, and that all users must be given equal treatment;
- Objectivity: statistics must be developed, produced and disseminated in a systematic, reliable and unbiased manner; it implies the use of professional and ethical standards, and that the policies and practices followed are transparent to users and survey respondents;
- Reliability: statistics must measure as faithfully, accurately and consistently as possible the reality that they are designed to represent and implying that scientific criteria are used for the selection of sources, methods and procedures
- Statistical confidentiality: the protection of confidential data related to single statistical units which are obtained directly for statistical purposes or indirectly from administrative or other sources and implying the prohibition of use for non-statistical purposes of the data obtained and of their unlawful disclosure;
- Cost effectiveness: the costs of producing statistics must be in proportion to the importance of the results and the benefits sought, that resources must be optimally used and the response burden minimised. The information requested shall, where possible, be readily extractable from available records or sources.

²⁵¹ Recital 71 GDPR.

²⁵² <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT&from=EN>

²⁵³ Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities (Text with relevance for the EEA and for Switzerland).

Bovendien wordt in artikel 12 van de verordening bepaald dat de volgende criteria in acht moeten worden genomen om de statistische kwaliteit te waarborgen:

- Relevance: the degree to which statistics meet current and potential needs of the users;
- Accuracy: the closeness of estimates to the unknown true values;
- Timeliness: the period between the availability of the information and the event or phenomenon it describes;
- Punctuality: the delay between the date of the release of the data and the target date (the date by which the data should have been delivered);
- Accessibility and Clarity: the conditions and modalities by which users can obtain, use and interpret data;
- Comparability: the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas, sectoral domains or over time;
- Coherence: the adequacy of the data to be reliably combined in different ways and for various uses.

Daarnaast kan worden verwezen naar de Algemene Vergadering van de Verenigde Naties, die 10 basisprincipes van de statistica heeft aangenomen.²⁵⁴ Verwijzend naar de cruciale rol van officiële statistische informatie bij beleidsbeslissing en benadrukkend dat het essentiële vertrouwen van het publiek in de integriteit van officiële statistische systemen en het vertrouwen in statistieken in grote mate afhangt van de eerbiediging van de fundamentele waarden en beginselen die de basis vormen van elke samenleving die zichzelf wil begrijpen en de rechten van haar leden wil respecteren, en in dit verband dat professionele onafhankelijkheid en verantwoordingsplicht van statistische bureaus van cruciaal belang zijn, worden de volgende tien beginselen nader toegelicht:²⁵⁵

1. Utility: Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.
2. Trust: To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.
3. Scientific reliability: To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.
4. Educational role: The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.
5. Quality, timeliness, costs and burden-sharing: Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

²⁵⁴ Zie ook: <https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf>

²⁵⁵ Resolution adopted by the General Assembly on 29 January 2014 [*without reference to a Main Committee (A/68/L.36 and Add.1)*] 68/261. Fundamental Principles of Official Statistics

6. Confidentiality: Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.
7. Transparency: The laws, regulations and measures under which the statistical systems operate are to be made public.
8. Coordination: Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.
9. Consistency: The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.
10. Cooperation: Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

Als laatste inspiratiebron kan worden verwezen naar de Praktijkcode Europese Statistieken, die in 2011 door het Comité voor het Europees statistisch systeem is goedgekeurd en waarin, voor zover relevant, de volgende beginselen zijn opgenomen (die hieronder worden geparafraseerd):

256

1. Professional environment:
 - a. *Independence*: Independence from political and other external interference in developing, producing and disseminating statistics is specified in law and assured for other
 - b. *Access*: Access to policy authorities and administrative public bodies.
 - c. *Competence*: The capacity of the head and employees of the organizations is beyond dispute and other arguments than their competence should play no role in their appointment.
 - d. *Transparency*: There is openness on the statistical work programs.
 - e. *Clarity*: Statistical releases are clearly distinguished and issued separately from political/policy statements.
2. Mandate and recourses:
 - a. *Mandate*: Authorities should have a mandate to gathered and process data for statistical analysis.
 - b. *Recourses*: Staff, financial, and computing resources, adequate both in magnitude and in quality, are available to meet current statistical needs.
3. Quality oversight: Authorities should monitor the quality of their statistical analysis, evaluate their work and the procedures in place to guarantee quality and external experts are consulted when appropriate.
4. Privacy: the data are kept safely and confidentially, both organizationally and technically.
5. Objectivity:
 - a. *Objective compilation*: Statistics are compiled on an objective basis determined by statistical considerations.
 - b. *Choices based on statistical considerations*: Choices of sources and statistical methods as well as decisions about the dissemination of statistics are informed by statistical considerations.
 - c. *Correction of errors*: Errors discovered in published statistics are corrected at the earliest possible date and publicized.

²⁵⁶ The European Statistics Code of Practice,
 <http://ec.europa.eu/eurostat/documents/64157/4392716/Revised_CoP_Nov_2017.pdf>.

- d. *Transparency*: Information on the methods and procedures used is publicly available.
 - e. *Notice of changes*: Advance notice is given on major revisions or changes in methodologies
 - f. *Objective and non-partisan*: Statistical releases and statements made in press conferences are objective and non-partisan.
6. Quality:
- a. *Consistency*: Procedures are in place to ensure that standard concepts, definitions and classifications are consistently applied throughout the statistical authority.
 - b. *Evaluation*: The business register and the frame for population surveys are regularly evaluated and adjusted if necessary in order to ensure high quality.
 - c. *Concordance*: Detailed concordance exists between national classifications systems and the corresponding European systems.
 - d. *Relevant expertise*: Graduates in the relevant academic disciplines are recruited.
 - e. *Relevant training*: Statistical authorities implement a policy of continuous vocational training for their staff.
 - f. *Cooperation with scientific community*: Co-operation with the scientific community is organized to improve methodology, the effectiveness of the methods implemented and to promote better tools when feasible.
7. Validation
- a. *Prior testing*: In the case of statistical surveys, questionnaires are systematically tested prior to the data collection.
 - b. *Reviewing*: Survey designs, sample selections and estimation methods are well based and regularly reviewed and revised as required
 - c. *Monitoring*: Data collection, data entry, and coding are routinely monitored and revised as required.
 - d. *Editing*: Appropriate editing and imputation methods are used and regularly reviewed, revised or updated as required.
 - e. *Designing*: Statistical authorities are involved in the design of administrative data in order to make administrative data more suitable for statistical purposes.
8. Reporting burden:
- a. *Proportionality*: Reporting burden is proportionate to the needs of the users and is not excessive for respondents.
 - b. *Necessity*: The range and detail of the demands is limited to what is absolutely necessary.
 - c. *Equality*: The reporting burden is spread as widely as possible over survey populations.
 - d. *Availability*: The information sought from businesses is, as far as possible, readily available from their accounts and electronic means are used where possible to facilitate its return.
 - e. *No duplicity*: Administrative sources are used whenever possible to avoid duplicating requests for information.
 - f. *Generality*: Data sharing within statistical authorities is generalized in order to avoid multiplication of surveys.
 - g. *Linkability*: Statistical authorities promote measures that enable the linking of data sources in order to reduce reporting burden.
9. Reality and comparability:

- a. *Validation*: Source data, intermediate results and statistical outputs are regularly assessed and validated.
 - b. *Documentation*: Sampling errors and non-sampling errors are measured and systematically documented according to the European standards.
 - c. *Analyzed*: Revisions are regularly analyzed in order to improve statistical processes.
 - d. *Coherent*: Statistics are internally coherent and consistent (i.e. arithmetic and accounting identities observed).
 - e. *Comparable*: Statistics are comparable over a reasonable period of time.
 - f. *Standardification*: Statistics are compiled on the basis of common standards with respect to scope, definitions, units and classifications in the different surveys and sources.
 - g. *Comparability*: Statistics from the different sources and of different periodicity are compared and reconciled.
 - h. *Exchange*: Cross-national comparability of the data is ensured through periodical exchanges between the statistical systems.
10. Accountability:
- a. *Metadata*: Statistics and the corresponding metadata are presented, and archived, in a form that facilitates proper interpretation and meaningful comparisons.
 - b. *Dissemination*: Dissemination services use modern information and communication technology and, if appropriate, traditional hard copy.
 - c. *Transparency*: Custom-designed analyses are provided when feasible and the public is informed.
 - d. *Microdata*: Access to microdata is allowed for research purposes and is subject to specific rules or protocols.
 - e. *Metadata*: Metadata are documented according to standardized metadata systems.
 - f. *Information*: Users are kept informed about the methodology of statistical processes including the use of administrative. Users are kept informed about the quality of statistical outputs with respect to the quality criteria.

2.4 Privacy en gegevensbescherming

Ten aanzien van het recht op gegevensbescherming is van belang dat de data die worden verwerkt voor AI en ML meestal op een gegeven moment in het proces wel als een persoonsgegeven zijn te kwalificeren, zodat onder meer de Algemene Verordening Gegevensbescherming van toepassing is. Daarbij moet er op worden gewezen dat er bij discriminatie een aantal punten en spanningsvelden een rol spelen.

- Legitiem: Gegevensverwerking moet legitiem zijn. Dat wil onder meer zeggen dat er een legitieme verwerkingsgrondslag dient te zijn. Voor overheidsinstanties is daarbij de standaard dat er een wettelijke bevoegdheid is en dat de verwerking een publiek belang dient. Voor private organisaties geldt dat zij zich kunnen beroepen op de toestemming van het datasubject of op een belang dat het belang van het datasubject overstijgt. Dat overstijgende belang kan echter alleen worden ingeroepen ten aanzien van gewone persoonsgegevens en niet voor bijzondere persoonsgegevens (zie hieronder verder over bijzondere persoonsgegevens).²⁵⁷

²⁵⁷ Artikel 6-10 AVG.

- Fair: Daarbij stelt de AVG dat gegevensverwerking ‘fair’ (behoorlijk) moet zijn. Alhoewel dit een tamelijk open begrip is, sluit het aan bij de in hoofdstuk 1 van dit rapport genoemde nadruk op fairness. Het gegevensbeschermingsrecht legt deze open norm op aan het gehele proces van gegevensverwerking, van het verzamelen en het opslaan tot aan het analyseren van de data en het gebruik van profielen voor automatische besluitvorming.²⁵⁸
- Doel en doelbinding: In hoofdstuk 1 kwam reeds contextualiteit ter sprake. Enerzijds is de contextualiteit van data van belang voor het goed kunnen begrijpen en duiden van uitkomsten van data-analyse; anderzijds is het voor het vertrouwen van burgers van belang dat data binnen een bepaalde context blijven. Deze nadruk ligt ook besloten in het principe van doel- en doelbinding als vervat in de AVG, waaruit volgt dat data die voor het ene, specifieke doel zijn verzameld, in principe slechts voor dat doel mogen worden verwerkt. Een uitzondering op het doelbindingsprincipe geldt echter als het gaat om gegevensverwerking voor statistische doeleinden. Hiermee wordt primair bedoeld op statistische verwerkingen door bijvoorbeeld het CBS. In hoeverre hier ook binnen de AI-context een beroep op kan worden gedaan is niet eenduidig vast te stellen.²⁵⁹
- Data minimalisatie en opslagbeperking: De AVG stelt globaal dat slechts die data mogen worden verzameld die noodzakelijk zijn voor het specifieke doel waarvoor ze worden verzameld en dat ze weer moeten worden verwijderd zodra het doel is verwezenlijkt. Dit kan botsen met AI-projecten die open-ended zijn geformuleerd en waarvan de uiteindelijke uitkomsten niet van tevoren duidelijk zijn. Ook is er een spanningsveld met de praktijk waarin zoveel mogelijk data worden verzameld om een algoritme te trainen.²⁶⁰
- Datakwaliteit: Interessant is dat de AVG ook vereist dat de data kwalitatief op orde zijn. Dit is een plicht van de persoon of organisatie die de data verwerkt, niet slechts een recht dat een datasubject (degene waarover de gegevens gaan) kan inroepen. Dit betekent onder meer dat als de data worden verzameld, er voor zorggedragen moet worden dat die correct zijn, dat als ze worden geanalyseerd ze up to date zijn en dat als ze worden toegepast in besluitvormingstrajecten, het gevonden patroon inderdaad van toepassing moet zijn op de persoon in kwestie. Daarbij heeft het datasubject ook het recht om aanvullende gegevens over hem aan te dragen, die het mogelijke beeld van hem zouden kunnen doen laten kantelen. Belangrijk is dat het principe van datakwaliteit met zich kan brengen dat data langer worden opgeslagen en meer gegevens moeten worden verzameld dan uit het dataminimalisatieprincipe zou volgen, bijvoorbeeld om te kunnen controleren of een AI-systeem nog op de meest recente en correcte data werkt of om metadata van gegevensverzamelingsinitiatieven te verzamelen.²⁶¹
- Transparantie: Een beperking van het dataminimalisatieprincipe kan ook volgen uit de plicht van de organisatie die de gegevens verwerkt om transparant te zijn en het recht van de burger om nadere informatie te vragen over de gegevensverwerking. Dit bleek onder meer uit een zaak van een Nederlands ingezetene die voor het Hof van Justitie kwam. De Nederlander had in een brief aan het College van Burgermeester en Wethouders gevraagd om opgave van alle verstrekkingen van hem betreffende informatie uit de gemeentelijke basisadministratie aan derden in de twee jaren

²⁵⁸ Artikel 5 AVG.

²⁵⁹ Artikel 5 AVG.

²⁶⁰ Artikel 5 AVG.

²⁶¹ B. van der Sloot, ‘From Data Minimization to Data Minimumization’, In: B. Custers, T. Calders, B. Schermer & T. Zarsky (eds.), ‘Discrimination and Privacy in the Information Society’, Springer, Heidelberg 2013, p. 273-287.

voorafgaand aan zijn verzoek. Het College wees dit verzoek slechts gedeeltelijk toe aangezien de opgevraagde gegevens die ouder waren dan één jaar voorafgaand aan zijn verzoek, automatisch waren verwijderd. Het Hof oordeelde hieromtrent: ‘Het staat aan de lidstaten, een termijn voor de bewaring van deze informatie en een daarop afgestemde toegang daartoe vast te stellen die een juist evenwicht vormen tussen, enerzijds, het belang van de betrokkene om zijn persoonlijke levenssfeer te beschermen, met name door middel van de bij richtlijn 95/46 [de voorloper van de AVG] voorziene mogelijkheden om de voor de verwerking verantwoordelijke in te schakelen en zich tot de rechter te wenden, en, anderzijds, de last die de verplichting tot bewaring van die informatie inhoudt voor de voor de verwerking verantwoordelijke. Een regeling waarbij informatie over de ontvangers of de categorieën ontvangers van de gegevens en over de inhoud van de verstrekte gegevens slechts één jaar wordt bewaard en de toegang tot die informatie dienovereenkomstig wordt beperkt, terwijl de basisgegevens veel langer worden bewaard, vormt geen juist evenwicht tussen het belang en de verplichting in kwestie, tenzij wordt aangetoond dat een langere bewaring van deze informatie een buitensporige last zou inhouden voor de voor de verwerking verantwoordelijke. Het staat aan de nationale rechter, de nodige controles te verrichten.’²⁶² Naast het recht op informatie over wie welke gegevens verwerkt en waarom heeft het datasubject ook het recht op informatie over het bestaan van geautomatiseerde besluitvorming en profilering en nuttige informatie over de onderliggende logica, alsmede het belang en de verwachte gevolgen van die verwerking voor het datasubject.²⁶³

- Bijzondere persoonsgegevens: Er geldt een speciaal regime voor het verwerken van bijzondere of gevoelige persoonsgegevens, waaronder valt persoonsgegevens waaruit ras of etnische afkomst, politieke opvattingen, religieuze of levensbeschouwelijke overtuigingen, of het lidmaatschap van een vakbond blijken en de verwerking van genetische gegevens, biometrische gegevens met het oog op de unieke identificatie van een persoon, of gegevens over gezondheid, of gegevens met betrekking tot iemands seksueel gedrag of seksuele gerichtheid. De verwerking van dit type gegevens is in principe verboden, tenzij er een uitzondering voor geldt. Dat kan bijvoorbeeld zijn vanwege publieke belangen omtrent veiligheid, volksgezondheid en archivering. In hoeverre hier ook onder valt het verwerken van bijzondere persoonsgegevens ten einde het gevaar voor discriminatie te verminderen is niet eenduidig te zeggen.²⁶⁴
- Geautomatiseerde besluitvorming: geheel geautomatiseerde besluitvorming en profiling die juridische gevolgen heeft of mensen in aanmerkelijke mate treft is verboden, tenzij er een wettelijke grondslag is of toestemming is verkregen. Het doel van deze bepaling is dat de organisatie die aan profiling doet controleert of het algemene beeld ook daadwerkelijk van toepassing is in het concrete geval waarin het profiel wordt toegepast.²⁶⁵

De Autoriteit Persoonsgegevens geeft in zijn richtsnoer op een aantal punten meer duidelijkheid ten aanzien van de toepasselijkheid van de AVG op AI.²⁶⁶

- Zo stelt de AP dat alhoewel een algoritme beoordelen op behoorlijkheid (fairness) op zichzelf complex is, dit wel degelijk mogelijk is met de juiste middelen en kennis. ‘De vraag of een algoritme of een uitkomst van een algoritme ‘fair’ of behoorlijk is, is

²⁶² Hof van Justitie, M. E. E. Rijkeboer v. Nederland, zaak C-553/07, 12 december 2007, ECLI:EU:C:2009:293.

²⁶³ Artikel 13 AVG.

²⁶⁴ Artikel 9 AVG.

²⁶⁵ Artikel 22 AVG.

²⁶⁶ https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/toezicht_op_ai_en_algoritmes.pdf

immers sterk verweven met de omstandigheden van het geval, maar daarnaast ook met mogelijk subjectieve opvattingen over rechtvaardigheid. Een verwerkingsverantwoordelijke zal zelf actief moeten verantwoorden en motiveren waarom een algoritme fair is en het gebruik van het gekozen algoritme niet leidt tot onbehoorlijke uitkomsten. Desalniettemin is het actief bevorderen van transparantie over algoritmes belangrijk voor onze democratische rechtsstaat. Zo gezien is transparantie niet alleen van belang voor betrokkenen zelf, maar ook voor andere partijen, zoals de media, belangenorganisaties in het maatschappelijk veld en andere toezichthouders.’

- Daarnaast wijst de AP op tal van verantwoordingsplichten die in de AVG staan genoemd, die interne en externe auditing van algoritmen en algoritmische besluitvorming mogelijk maken, zoals de plicht om een register bij te houden van dataverwerkingen, data protection impact assessments te doen bij risicovolle gegevensverwerkingen en het implementeren van data protection by design standaarden.

Een studie van Sator et al. uitgevoerd voor het Europees Parlement concludeert dat AI niet in strijd hoeft te zijn met de AVG.²⁶⁷ De studie lijkt een flexibele interpretatie van de AVG voor te staan. Nu en dan is het rapport meer voorschrijvend dan beschrijvend: ‘The fundamental data protection principles – especially purpose limitation and minimisation – should be interpreted in such a way that they do not exclude the use of personal data for machine learning purposes. They should not preclude the creation of training sets and the construction of algorithmic models, whenever the resulting AI systems are socially beneficial and compliant with data protection rights. The use of personal data in a training set, for the purpose of learning general correlations and connection, should be distinguished from their use for individual profiling, which is about making assessments about individuals. The inference of new personal data, as it is done in profiling, should be considered as creation of new personal data, when providing an input for making assessments and decisions. The same should apply to the re-identification of anonymous or pseudonymous data.’

De European Data Protection Supervisor, waarin alle nationale handhavende organisaties ten aanzien van het gegevensbeschermingsrecht zijn verenigd, lijkt wat kritischer en heeft naar aanleiding van het AI-raamwerk van de Europese Commissie een opinie uitgegeven waarin het stelt dat een eventueel nieuw raamwerk voor AI aan de volgende voorwaarden moet voldoen:

- applies both to EU Member States and to EU institutions, offices, bodies and agencies;
- is designed to protect from any negative impact, not only on individuals, but also on communities and society as a whole;
- proposes a more robust and nuanced risk classification scheme, ensuring any significant potential harm posed by AI applications is matched by appropriate mitigating measures;
- includes an impact assessment clearly defining the regulatory gaps that it intends to fill.
- avoids overlap of different supervisory authorities and includes a cooperation mechanism.
- Regarding remote biometric identification, the EDPS supports the idea of a moratorium on the deployment, in the EU, of automated recognition in public spaces of human features, not only of faces but also of gait, fingerprints, DNA, voice, keystrokes and other biometric or behavioural signals, so that an informed and democratic debate can take place and until the moment when the EU and Member States have all the

²⁶⁷ [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU\(2020\)641530_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf)

appropriate safeguards, including a comprehensive legal framework in place to guarantee the proportionality of the respective technologies and systems for the specific use case.²⁶⁸

Zuiderveen Borgesius heeft voor de Raad van Europa de principes ten aanzien van discriminatie en AI in kaart gebracht.²⁶⁹ Ten aanzien van het gegevensbeschermingsrecht kom hij tot een aantal punten van aandacht:

- First, there is a compliance and enforcement deficit. Data Protection Authorities have limited resources. And many Data Protection Authorities do not have the power to impose serious. Previously, many organisations did not take compliance with data protection law seriously. It appears that compliance improved with the arrival of the GDPR, but it is too early to tell.
- Second, parts of algorithmic processes are outside the scope of data protection law. Data protection law only applies when personal data are processed. It does not apply to predictive models because they do not relate to identifiable persons. For example, a predictive model that says "80% of the people living in postal code F-67075 pay their bills late" is not a personal datum, as the model does not refer to an individual.
- Third, data protection law uses many open and abstract norms, rather than black-and-white rules. Data protection law must use open norms, because its provisions apply in many different situations, in the private and the public sector. This regulatory approach, an omnibus approach, has many advantages. For instance, the open norms do not have to be adapted each time when a new technology is developed. But one disadvantage is that the open norms can be difficult to apply.
- Fourth, data protection law has strict rules on "special categories" of, such as data regarding racial origin or revealing health status. Those rules create challenges for assessing and mitigating discrimination. Many of the methods to tackle discrimination in AI systems implicitly assume that organisations hold these sensitive data – yet to meet data protection law, many organisations may not be holding them. Tension remains between respecting data protection law and collecting sensitive data to fight discrimination.
- Fifth, even where explanations of AI decisions might be legally required by the GDPR or Convention 108, it is often difficult to explain the logic behind a decision, when an AI system, analysing large amounts of data, arrives at that decision. And in some cases, it is not clear how much an explanation would help people, especially insofar as it places the burden on them to understand the decision and its appropriateness.

Het recht op privacy, zoals onder meer vervat in artikel 8 van het Europees Verdrag voor de Rechten van de Mens geeft nog een aantal aanvullende standaarden. Daarbij moet worden bedacht dat zelfs als er geen ‘persoonsgegevens’ worden gebruikt, het privacyrecht wel van toepassing op besluiten die aan iemands persoonlijke levenssfeer raken. Het privacyrecht kent geen verbod op bijvoorbeeld automatische besluitvorming, maar onderkent wel een aantal procedurele rechten van burgers, met name als overheidsinstanties besluiten nemen. Al in 1987, in een aantal uitspraken die het Hof deed op dezelfde dag,²⁷⁰ maakte het Hof duidelijk dat het niet alleen keek naar materiële inperkingen op het recht op privacy, maar ook op de wijze waarop dergelijke inbreuken tot stand zijn gekomen. In die zaken stelde het Hof, nog voor toe

²⁶⁸ https://edps.europa.eu/sites/edp/files/publication/20-06-19_opinion_ai_white_paper_en.pdf

²⁶⁹ <https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73>

²⁷⁰ EHRM, W. v. The United Kingdom, no. 9749/82, 8 July 1987.

te komen aan de beoordeling van de klacht, dat de ‘background to the instant case is constituted by certain judicial or local authority decisions regarding the applicant’s child S. The Court finds it important to emphasise at the outset that the present judgment is not concerned with the merits of those decisions; this issue was not raised by the applicant before the Commission and did not form part of the application which it declared admissible. Since the Commission’s admissibility decision delimits the compass of the case brought before the Court, the latter is not in the circumstances competent to examine or comment on the justification for such matters as the taking into public care or the adoption of the child or the restriction or termination of the applicant’s access to him.’²⁷¹

In die zaken vond het Hof een schending van artikel 6 EVRM, het recht op een eerlijk proces, maar beoordeelde het EHRM primair of er een schending was van het recht op privacy. Het Hof stelde daarbij: ‘It is true that Article 8 contains no explicit procedural requirements, but this is not conclusive of the matter. The local authority’s decision-making process clearly cannot be devoid of influence on the substance of the decision, notably by ensuring that it is based on the relevant considerations and is not one-sided and, hence, neither is nor appears to be arbitrary. Accordingly, the Court is entitled to have regard to that process to determine whether it has been conducted in a manner that, in all the circumstances, is fair and affords due respect to the interests protected by Article 8. Moreover, the Court observes that the English courts can examine, on an application for judicial review of a decision of a local authority, the question whether it has acted fairly in the exercise of a legal power. The relevant considerations to be weighed by a local authority in reaching decisions on children in its care must perforce include the views and interests of the natural parents. The decision-making process must therefore, in the Court’s view, be such as to secure that their views and interests are made known to and duly taken into account by the local authority and that they are able to exercise in due time any remedies available to them.’²⁷²

Sindsdien heeft het EHRM tal van voorwaarden gesteld aan overheidsbeslissingen die aan de persoonlijke levenssfeer van burgers raken:

- Burgers moeten worden geïnformeerd dat er een beslissing over hen wordt genomen;
- Zij moeten toegang hebben tot alle relevante informatie;
- De beslissing dient op een onafhankelijke wijze tot stand te komen;
- Burgers hebben het recht om betrokken te worden bij dergelijke besluiten en gehoord te worden in hun visie;
- Burgers hebben het recht beslissingen aan te vechten;
- Beslissingen moeten tijdig worden genomen;
- Burgers hebben het recht op bijstand en vertegenwoordiging;
- De beslissingen moeten begrijpelijk en fair zijn;
- De procedure om tot die beslissing te komen moet fair en adequaat zijn.²⁷³

²⁷¹ EHRM, B. v. the United Kingdom, no. 9840/82, 08 July 1987, § 58.

²⁷² EHRM, B. v. The United Kingdom, no. 9840/82, 8 July 1987, § 63-64.

²⁷³ Zie verder: B. van der Sloot, 'Decisional privacy 2.0: the procedural requirements implicit in Article 8 ECHR and its potential impact on profiling', International Data Privacy Law, Volume 7, Issue 3, 1 August 2017.

3. Evaluatie bestaande standaarden en principes

Op basis van stappen 1 en 2 is een eerste overzicht van principes en standaarden gemaakt (paragraaf 3.1). Deze is vervolgens voorgelegd aan drie verschillende testgroepen, met verschillende achtergronden en expertises (paragraaf 3.2). Op basis daarvan is vervolgens aanvullend onderzoek gedaan naar anti-discriminatiejurisprudentie op Europees niveau en relevante buitenlandse jurisprudentie (paragraaf 3.3).

3.1 Overzicht standaarden

Op basis van stappen 1 en 2 is een eerste overzicht van principes en standaarden gemaakt, dat hieronder is opgenomen. Daarbij zijn drie keuzes van belang.

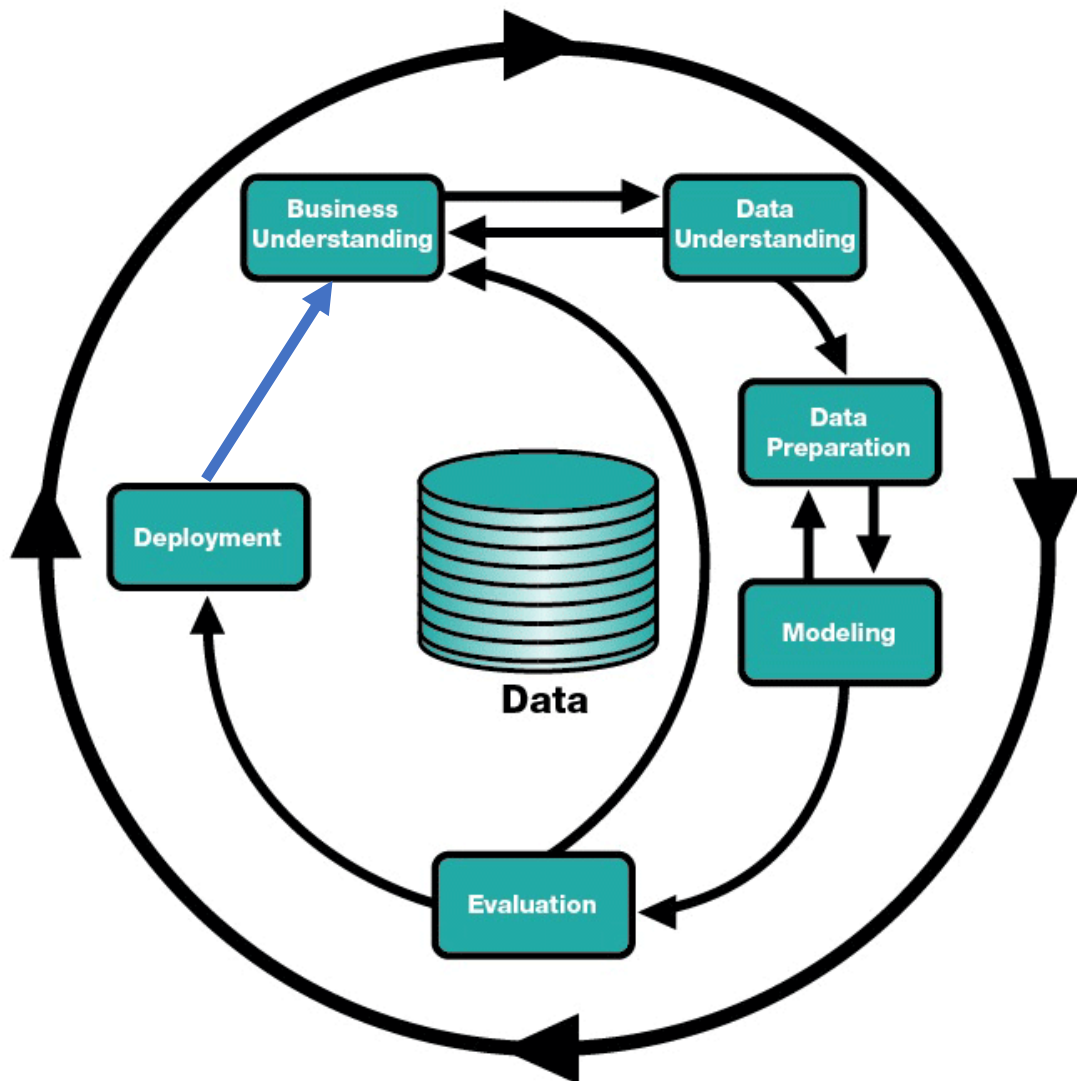
Allereerst is het probleem dat anti-discriminatierecht of gelijkebehandelingsjurisprudentie geen duidelijk handvatten geeft over procesinrichting, over hoe een besluit tot stand moet komen. Als een besluit eenmaal is genomen, dan wordt *ex post* getoetst of het besluit direct of indirect op een van de verboden gronden is genomen, of het besluit onevenredige effecten heeft op bepaalde groepen in de samenleving en als dat zo is, of daar een objectieve rechtvaardiging voor is. De handreiking ziet voor een groot deel op de *ex ante* fase van een besluit, het zorgen dat een door AI aangedreven of gestuurd besluitvormingsproces zo neutraal en non-discriminatoir mogelijk is. Alhoewel sommige van de juridische standaarden kunnen worden geëxtrapoleerd naar eerdere fasen, levert dit onvoldoende materiaal op om tot een volwaardige handreiking te komen. Daarom is in deze handreiking ook gekeken naar principes uit vakgebieden die traditioneel meer zien op de procesinrichting van besluitvormingssystemen, zoals de AI-literatuur, statistische beginselen en het gegevensbeschermingsrecht.

Ten tweede is gekozen om de principes onder te verdelen in juridische beginselen, technische beginselen en organisatorische beginselen. Daarbij moet uiteraard worden opgemerkt dat die type standaarden niet altijd eenduidig zijn te categoriseren. In de Algemene Verordening Gegevensbescherming staan immers tal van principes die primair organisatorisch van aard zijn; in de technische literatuur wordt een aantal juridische principes meegenomen en uitgewerkt; etc. Ook bestaat er de nodige overlap tussen de verschillende principes. Toch is er gekozen voor deze driedeling, waarbij de organisatorische principes zien op de procesinrichting (wie zit er in het team, hoe worden beslissingen gedocumenteerd, etc.), technische principes zien op de inrichting van het AI-systeem als zodanig (welke fairness standaard wordt gekozen, welke bias wordt als aanvaardbaar geacht, etc.) en juridische principes, die zien op de evaluatie van het systeem (waarom worden er bepaalde gegevens verzameld, zijn de gegevens wel echt nodig voor dat doel, etc.).

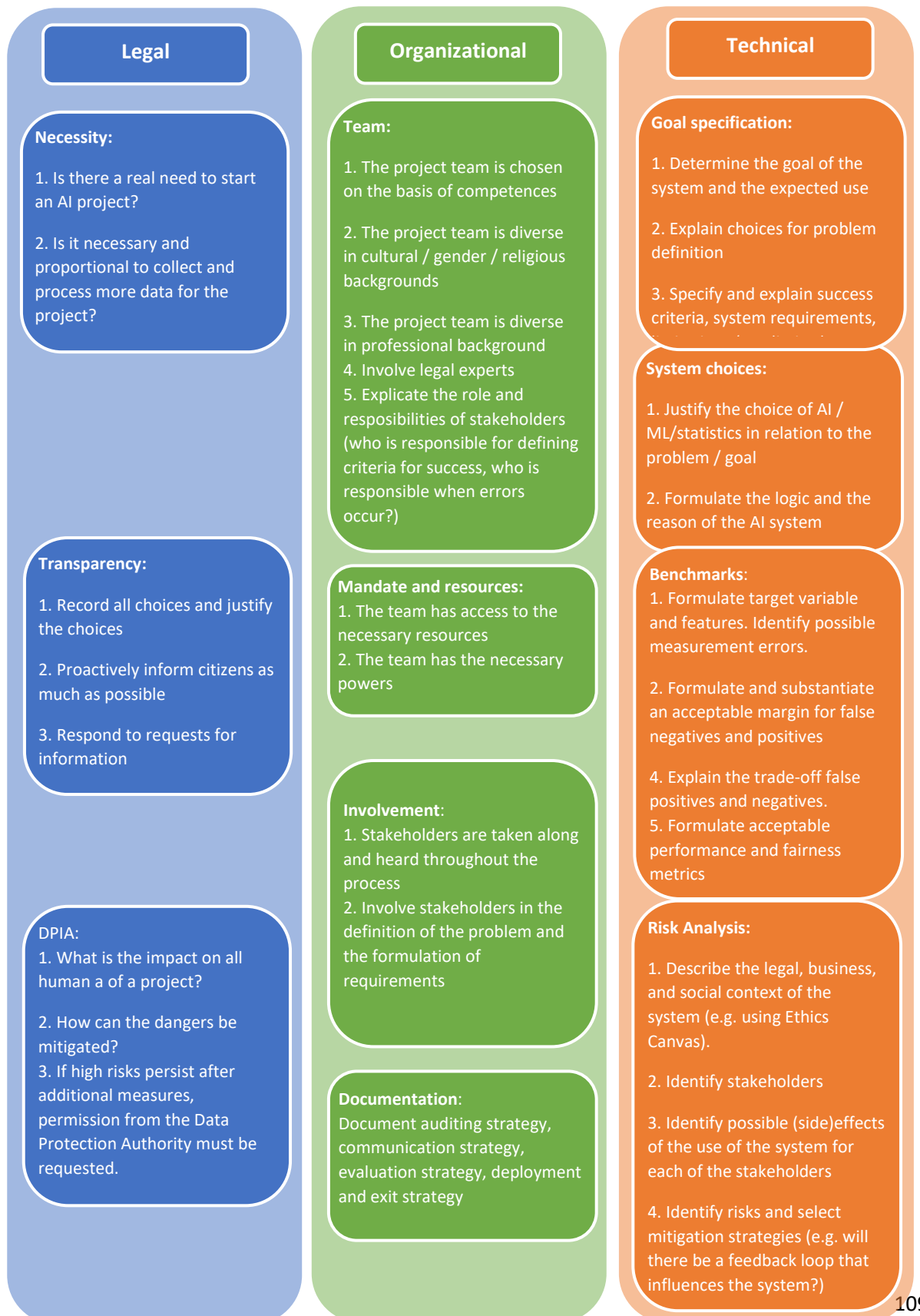
Ten derde is besloten om voor de indeling aan te sluiten bij de visualisering in de Cross-industrie standaardproces voor datamining ([CRISP-DM](#)). Dit model is goed bekend bij AI-experts en geeft gelijk een indeling per fase van het proces. De concrete juridische, organisatorische en technische handvatten zullen per fase in het proces worden geclusterd, om zo een stappenplan te hebben voor het bouwen, ontwikkelen en gebruiken van een AI-systeem. Daarbij moet wel worden bedacht dat het in de realiteit geen lineair proces betreft, maar eerder een Processie van Echternach is, waarbij steeds twee stappen vooruit wordt gezet en dan weer eentje achterwaarts. Met name bij de laatste twee fasen zijn in handreiking aanpassingen gedaan op dit model, om meer ruimte te geven voor permanente evaluatie.

Draft overview standards and principles

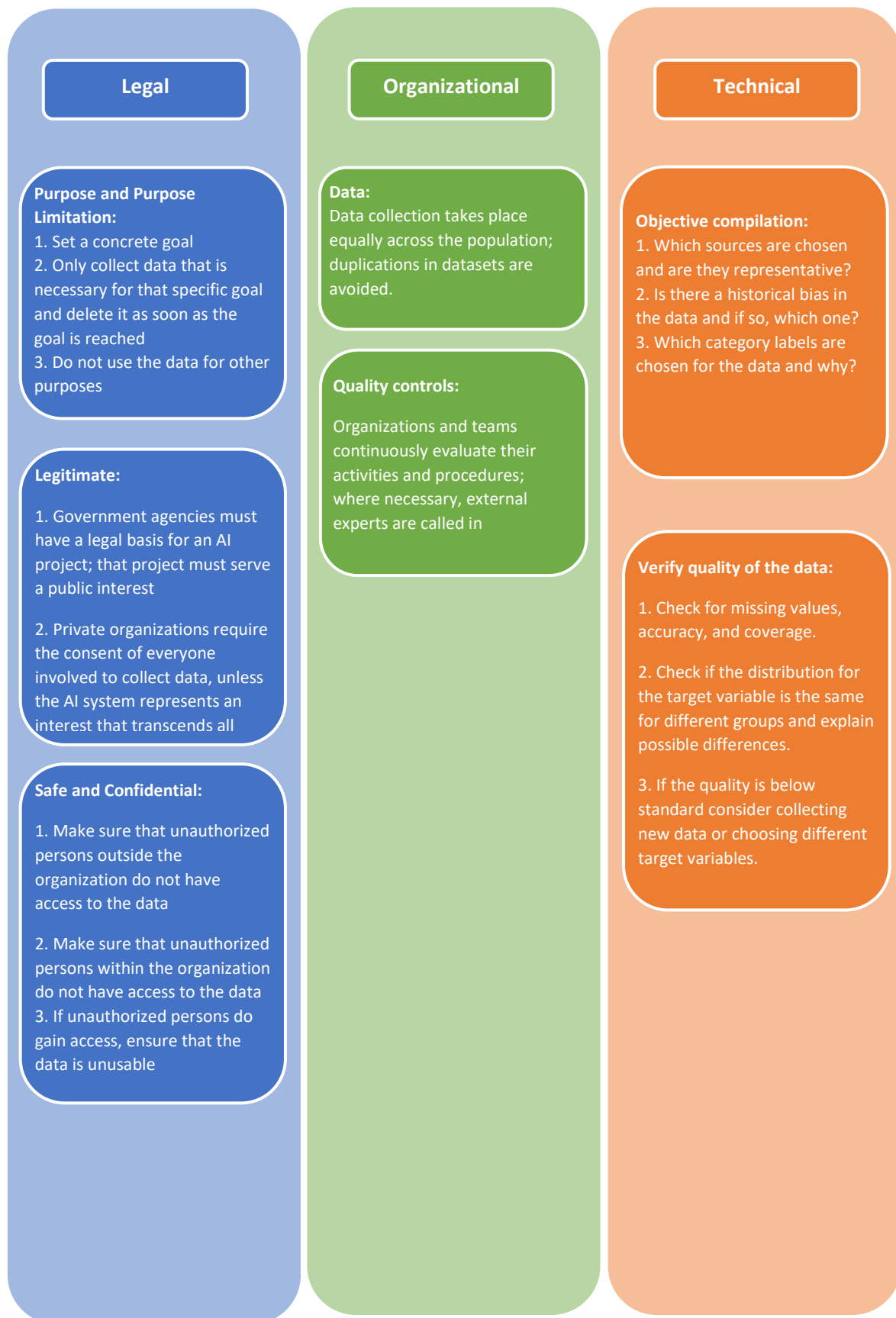
The overview below is based on the Cross-industry standard process for data-mining (CRISP-DM) model. This model is widely used as a process model by data mining experts. We have used this model to structure the standards and principles we identified in our literature review and further organized them along three dimensions: legal, organizational and technical.



(1) Problem definition: identifying the problem / the goal and that translation to AI/ML tasks



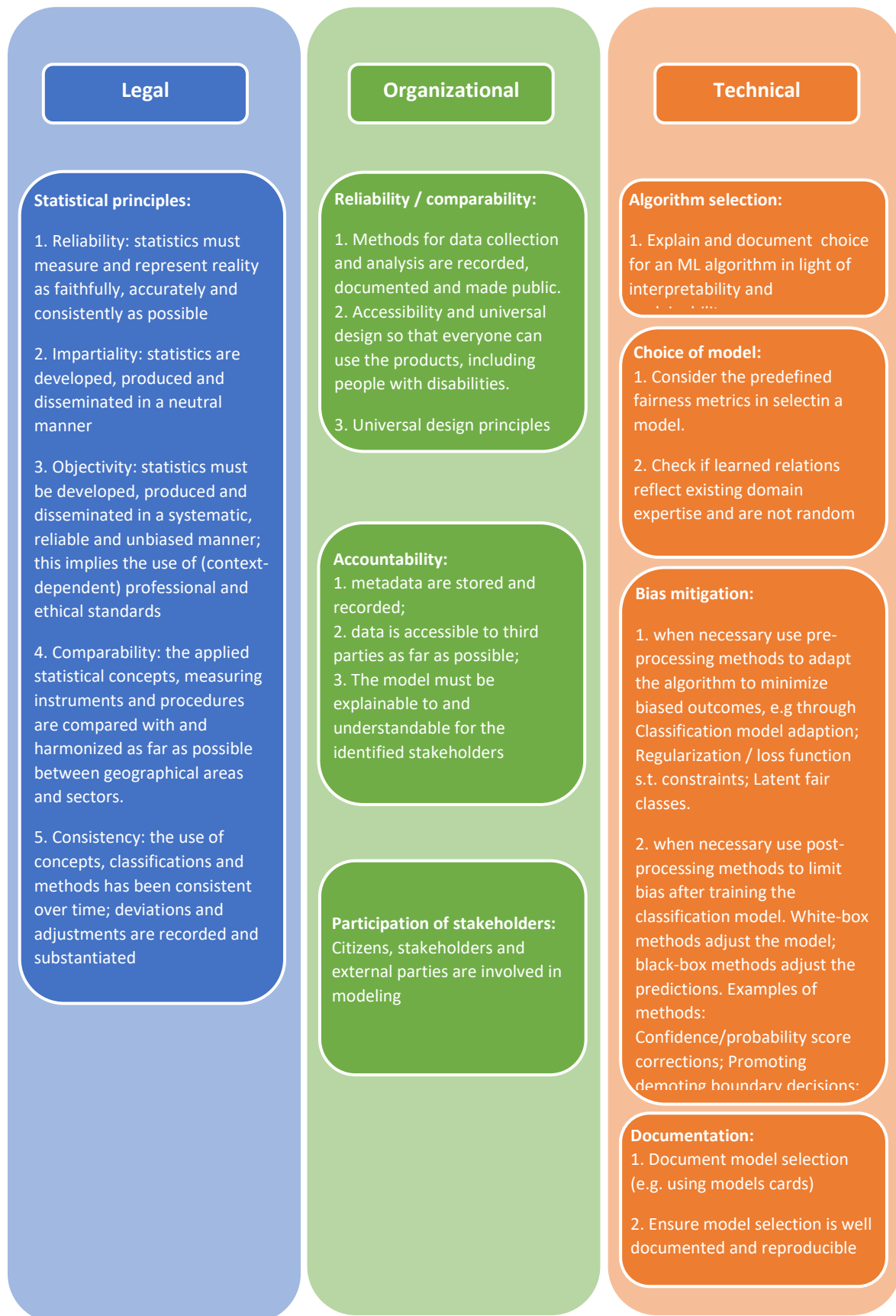
(2) Initial data collection and storage



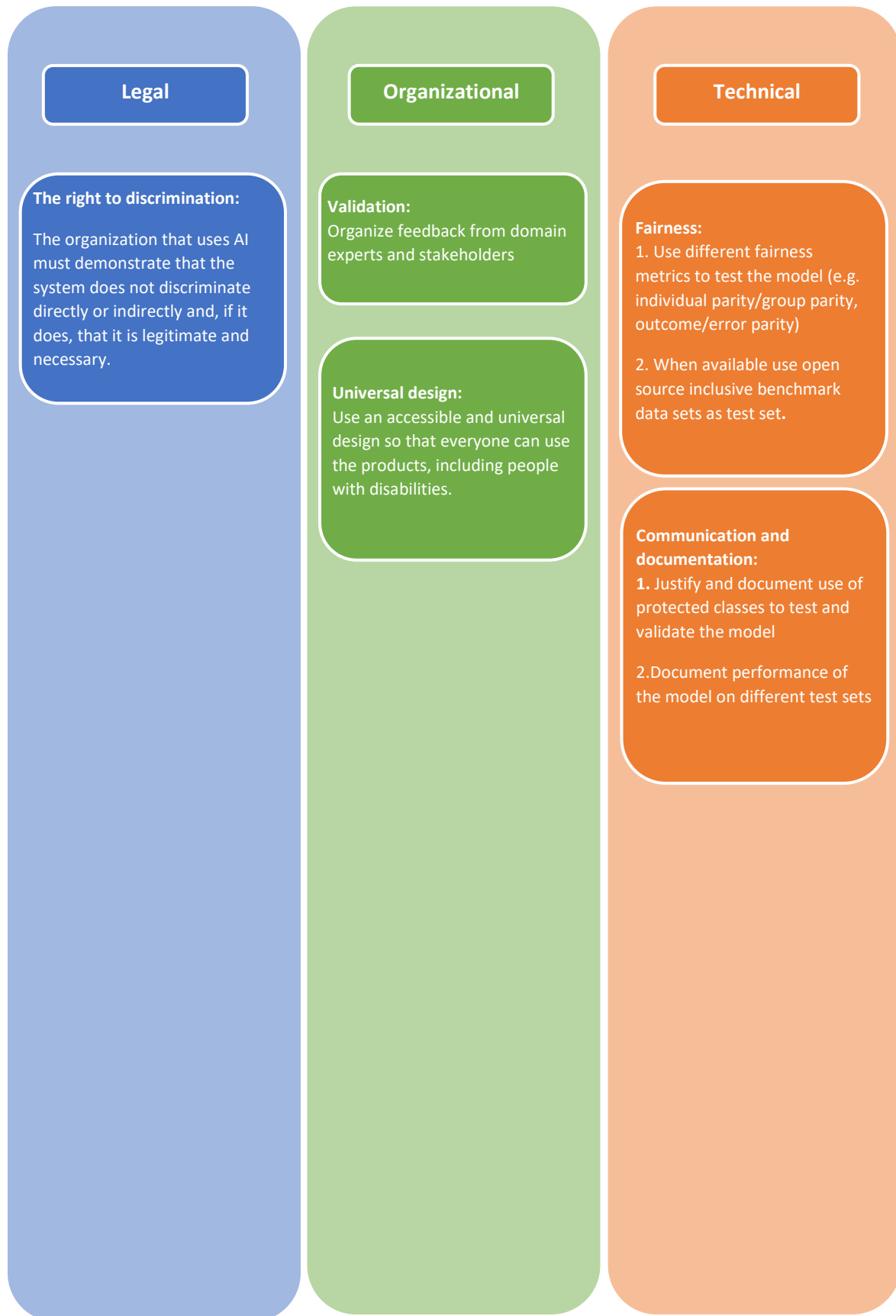
(3) Data analysis and preparation



(4) Modeling



(5) Evaluation (evaluate the selected model based on the success criteria established in step 1 and a “test set”)



(6) Deployment (start using the model)



3.2 Testgroepen

Dit overzicht, plus de samenvatting van hoofdstukken 0, 1 en 2, is vervolgens voorgelegd aan drie verschillende testgroepen. Testgroep 1 bestond uit juristen (Marlies van Eck, Janneke Gerards, Jan-Peter Loof en Frederik Zuiderveen Borgesius), testgroep 2 uit ‘techneuten’ (Harry van Geijn, Seda Gürses, Violetta Misheva en Yugesh Raghoenath) en testgroep 3 bestond uit een combinatie van deelnemers met verschillende achtergronden (Francien Desechne, Sennay Ghebreab, Ronald Leenes en Monique Steijns). Hierdoor werd vanuit de juridische en de technische kant disciplinaire feedback verkregen en werd ook gekeken hoe er vanuit een interdisciplinair perspectief naar deze principes werd gekeken. De uitgebreide verslagen van deze testgroepen zijn in de bijlages opgenomen. Hieronder staan de hoofdlijnen.

3.2.1 Testgroep 1

Beperking ambities: Het is vrijwel onmogelijk om anti-discriminatiejurisprudentie in een helder model te vatten, zowel omdat het anti-discriminatierecht zoveel verschillende afwegingen en keuzes vergt, als omdat de juridische factoren die moeten worden meegenomen complex zijn, als omdat de invulling van deze factoren en uitgangspunten context-afhankelijk is.

Hou het zacht/open: ‘techneuten’ willen graag weten wat wel en wat niet mag. Dat kan een dergelijke handreiking niet bieden, want dat zegt het recht niet zo helder. Bovendien hebben duidelijke principes het nadeel dat eromheen kan worden gewerkt; is bijvoorbeeld het principe dat er geen besluiten op basis van factoren A, B en C mogen worden genomen, dan kan met gemak naar proxies worden gezocht.

Bewustwording: Het gaat dus om het duidelijk maken van de kernuitgangspunten van het recht, zodat die tussen de oren zitten. ‘Techneuten’ kunnen dan zelf vanuit die uitgangspunten nadenken over de inrichting van hun systeem, of om hulp vragen, als zij vrezen voor mogelijke problemen. Als de ‘techneuten’ de basisprincipes snappen, dan is daar al heel veel mee gewonnen.

Red flags: Maak een stappenplan en zet daar ook procedurele stappen in. Bij discriminatiegevaaren kan er ook worden gewerkt met red flags; als A of B aan de hand is, schakel dan een jurist in voor een second opinion, of iets dergelijks.

Ex-ante en ex-post: Een handreiking zal zowel moeten inzetten op ex-ante zorgvuldigheidsprincipes, als op een ex-post toets op discriminatie. Daarbij is het van belang dat de uitkomsten van de ex-post toets weer terug worden gegeven aan het systeem. Eindig dus met een goede evaluatieronde.

Domeinkennis: AI-systeembouwers moeten altijd ook kennis hebben van het domein waarin het systeem wordt ingezet; anders zegt de feedback die aan een systeem wordt gegeven niets en is het gevaar dat data verkeerd worden geïnterpreteerd. Daarbij speelt ook dat bij bepaalde domeinen extra vereisten spelen, zoals binnen de strafrechtketen en het medische domein.

3.2.2 Testgroep 2

Organisatie: De meeste problemen en meeste oplossingen zijn te vinden in het organisatorische gedeelte van het overzicht.

Diversiteit: Het team dat een AI-systeem uitwerkt en test moet divers zijn, zowel qua professionele als qua persoonlijke achtergrond. Teams moeten samen worden gesteld met een brede variëteit aan taken en expertises. Het is belangrijk hier ook stakeholders bij te betrekken, het liefst bij alle fasen van het proces.

Noodzakelijkheid: De noodzakelijkheidsvraag (is dit systeem überhaupt nodig) moet centraal staan. Ook de noodzaak van data voor een bepaald systeem. Problemen ontstaan doordat meer data dan strikt noodzakelijk aan het systeem worden toegevoegd en dat daaruit weer allerhanden kruisbestuivingen en proxies kunnen worden afgeleid. Dat gevaar is minder aanwezig als het strikt blijft bij de data die nodig zijn voor het directe doel.

Documentatie: Documentatie van alle vragen en stappen is essentieel, omdat het vaak een iteratief proces betreft. Wel moet duidelijk worden gemaakt wat transparantie en documentatie precies inhouden, op wie de plicht rust om hiervoor zorg te dragen en aan wie transparantie moet worden geboden.

Datakwaliteit: Belangrijk is dat als data van derde partijen worden verkregen, goed in kaart wordt gebracht hoe die data zijn verzameld, waartoe, etc. Moeilijkheid gaat ontstaan als er moet worden gecontroleerd op juistheid van de data. Dat is een belangrijk aspect.

Continue proces: AI-systemen blijven leren, dus is het belangrijk om te blijven testen en te evalueren.

3.2.3 Testgroep 3

Beperking ambities: Als de verwachting is ‘vermijd alle bias’ dan heeft het project geen zin. Het gaat erom hoe je er mee omgaat. Het zou wel duidelijk moeten maken wat de beperkingen van een bepaald design zijn. Er zijn verschillende juridische noties van fairness; de keuze die je maakt voor formalisering creëert nieuwe bias die niet per se schadelijk is. Maar het is belangrijk om de impact van de keuzes die je maakt te kennen.

Iteratief proces: Er moet zich een voortdurend iteratief proces voltrekken, tussen de verschillende fasen, maar ook tussen de hoog-over principes en de praktische toepassing daarvan in concrete casuïstiek, en tussen de juridische, organisatorische en technische principes, die ook met elkaar samenhangen.

Handreiking nooit af: Een statische handreiking slaat de plank mis, want AI-systemen, anti-discriminatiejurisprudentie en ook de interactie daartussen verandert steeds.

Bewustwording: Niet alleen is bewustwording belangrijk ten aanzien van anti-discriminatiejurisprudentie en van mogelijke gevaren van AI-systemen in dat kader, maar ook van het feit dat de wereld al gebiased/discriminatoire is.

Intersectionaliteit: Het is belangrijk om zo veel mogelijk disciplines bij elkaar te brengen, ook omdat uiteindelijk alles met alles samenhangt. Daarbij moet zeker ook plaats zijn voor specifieke domeinkennis en een consultatieproces met stakeholders.

Operationalisering handreiking: Je begint in de praktijk niet met juridische vragen maar met een probleem dat opgelost moet worden met een AI-systeem, dan pas ga je over andere zaken nadenken. Projectdefinitie fase start vanuit de organisatie of business, afhankelijk van de context, maar daar moet een probleemstelling komen: dit is het probleem en we willen het oplossen met dit systeem.

3.3 Aanvullend onderzoek

Mede op basis van de input van de drie testgroepen is gezocht naar aanvullende anti-discriminatieprincipes. Daarbij is gekozen naar de jurisprudentie van het Europees Hof voor de Rechten van de Mens (paragraaf 3.3.1), het EU Hof van Justitie (paragraaf 3.3.2) en een selectie aan rechtszaken over AI-gedreven systemen (paragraaf 3.3.3). Daarbij is allereerst van belang dat op basis van het in hoofdstuk 2 beschrevene het volgende kan worden vastgesteld:

1. Er is geen uniforme reikwijdte van de verschillende gelijke behandelingsbeginselen.
Onderscheid overheidshandelen (brede, niet-limitatieve beschermde gronden) en relaties in het economisch verkeer (limitatief opgesomde gronden op specifieke terreinen).
2. Er zijn een veelheid aan discriminatieconcepten.
Direct, indirect, en daarvan de varianten associatief, vermeend, toekomstig... Daarnaast nog specifieke elementen zoals intimidatie, doeltreffende aanpassing en toegankelijkheid.
3. Het type discriminatie en de grond bepalen hoe streng getoetst wordt en welke uitzonderingen en/of rechtvaardigingen toelaatbaar zijn. Deze toetsing is bovendien contextgebonden (e.g. rechtvaardigingsgronden bestaan niet noodzakelijkerwijs objectief maar worden door partijen aangedragen).

Het is de vraag of een self-cleaning algoritme volledig mogelijk is, gegeven deze omstandigheden. Het is recent betoogd door Wachter et al. dat de contextgebonden analyse die in de Europese gelijkebehandelingsstoets plaatsvindt zich hier lastig voor leent. Als alternatief beargumenteren de auteurs dat je het vaststellen van het onderscheid zou moeten scheiden van de toetsing ervan: het zwaartepunt zou moeten liggen op het herkennen van (voor het gelijkebehandelingsrecht relevante) discriminatie.²⁷⁴

Zij stellen daarvoor zelf de conditional demographic disparity-test voor, dat er als volgt uitziet:

²⁷⁴ S. Wachter, B. Mittelstadt & C. Russell (2020). 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI', https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547922.

$$A_R = \frac{\text{No. of protected people in advantaged group with attributes R}}{\text{total no. of people in advantaged group with attributes R}}$$

$$D_R = \frac{\text{No. of protected people in disadvantaged group with attributes R}}{\text{Total no. of people in disadvantaged group with attributes R}}$$

We can write the set of tests as the following formula:

Conditional Demographic Disparity	$D_R > A_R$ for any choice of attributes R

Het lijkt kansrijk te onderzoeken of er een mogelijkheid is om een algoritme een zelf-toets te laten uitvoeren om *a prima facie* discriminatie vast te stellen (stap 1+2), dit te loggen (zodat deze ook voor anderen dan de eigenaar-gebruiker beschikbaar is en blijft bij een eventueel onderzoek) en daar de eigenaar-gebruiker van het algoritme op te attenderen. Dit zou vervolgens gecombineerd kunnen worden met een self-reporting obligation.

3.3.1 EHRM

Wat volgt is een bespreking van de non-discriminatie jurisprudentie van het Europees Hof voor de Rechten van de Mens (EHRM) en een poging tot vertaalslag naar AI-systeem principes. Ter herinnering halen we artikel 14 van het Europees Verdrag voor de Rechten van de Mens (EVRM) erbij om de verboden gronden van discriminatie te identificeren en wordt een overzicht gegeven van de toets die het Hof uitvoert om tot haar oordeel te komen of een handeling al dan niet discriminerend is. Daarna bespreken we jurisprudentie waaruit blijkt dat een vuistregel voor non-discriminatie zeer moeilijk, zo niet onmogelijk, te bekomen is. Of er al dan niet sprake is van discriminatie hangt, zo blijkt, bijna altijd af van de omstandigheden van het geval. Een vertaalslag naar ‘een AI-systeem’, zonder meer, blijft dan ook uit. Om dit duidelijk te maken gebruiken we als leidraad de manieren waarop een overheid een verschil in behandeling kan rechtvaardigen. Deze manieren zijn: aantonen dat de verdachte grond geen relevante factor is in de beslissing; aantonen dat de verdachte grond relevant is om in overweging te nemen; of aantonen dat de maatschappelijke meerwaarde zwaarder doorweegt dan het nadeel ondervonden door de ongelijke behandeling.²⁷⁵ De vertaalslag die we wel kunnen maken is bijna altijd afgelijnd door de omstandigheden van het geval.

Artikel 14 van het Europees Verdrag voor de Rechten van de Mens

Het verbod op discriminatie wordt in artikel 14 als volgt uitgelegd:

“Het genot van de rechten en vrijheden die in dit Verdrag zijn vermeld, moet worden verzekerd zonder enig onderscheid op welke grond ook, zoals geslacht, ras, kleur, taal, godsdienst,

²⁷⁵ J. Gerards (2017). The Margin of Appreciation Doctrine, the Very Weighty Reasons Test and Grounds of Discrimination. M. Balboni (ed) (2017). The principle of discrimination and the European Convention of Human Rights, Editoriale Scientifica, p. 19.

politieke of andere mening, nationale of maatschappelijke afkomst, het behoren tot een nationale minderheid, vermogen, geboorte of andere status.”²⁷⁶

Een schending van artikel 14 slechts kan voorkomen wanneer dit wordt samengenomen met een ander door het EVRM beschermde voorziening. Discriminatie moet dus altijd worden vastgesteld in relatie tot een in het verdrag beschermd recht. Het Hof legt dit als volgt uit:

“Article 14 only complements the other substantive provisions of the Convention and its Protocols. It has no independent existence, since it has effect solely in relation to the “rights and freedoms” safeguarded by those provisions. Although the application of Article 14 does not presuppose a breach of one or more of such provisions, and to this extent it is autonomous, there can be no room for its application unless the facts of the case fall within the ambit of one or more of the latter.”²⁷⁷

Wanneer bijvoorbeeld iemand meent het recht op vrijheid van vergadering en vereniging (artikel 11 van het EVRM) te zijn ontzegd geweest op basis van zijn of haar seksuele geaardheid kan de persoon in kwestie een zaak aanhangig maken voor het Hof.

Toets

De toets die het Hof uitvoert om te oordelen of een handeling al dan niet discriminerend is omschrijft ze zelf als volgt:

“According to the Court’s established case-law, in order for an issue to arise under Article 14 there must be a difference in treatment of persons in relevantly similar situations. Such a difference of treatment is discriminatory if it has no objective and reasonable justification; in other words, if it does not pursue a legitimate aim or if there is not a reasonable relationship of proportionality between the means employed and the aim sought to be realised. The Court reiterates that the Contracting States enjoy a margin of appreciation in assessing whether and to what extent differences in otherwise similar situations justify a difference in treatment”²⁷⁸

De toets kan men dus samenvatten als volgt:

- Eerst wordt een verschil in behandeling geconstateerd
- Vervolgens wordt nagegaan of de verschillend behandelde partijen zich in een vergelijkbare situatie bevinden
- Daarna kijkt het Hof of er voor dit verschil in behandeling een redelijke en objectieve rechtvaardiging is
 - o Dit doet ze door na te gaan of het beoogde doel legitiem is
 - o En door te kijken of de handeling proportioneel is tot dat doel

Het Hof maakt ook melding van het feit dat de lidstaten een bepaalde mate van discretionaire bevoegdheid hebben om te bepalen of een verschil in behandeling al dan niet gerechtvaardigd is.

a. Margin of appreciation

²⁷⁶ Verdrag tot Bescherming van de Rechten van de Mens en de Fundamentele Vrijheden Rome, 4.XI.1950.

²⁷⁷ EHRM, 2 maart 2010, nr. 13102/02, ECLI:CE:ECHR:2010:0302JUD001310202 (Kozak v. Poland), para. 82.

²⁷⁸ EHRM, 20 juni 2017, nr. 67667/09, 44092/12, 56717/12, ECLI:CE:ECHR:2017:0620JUD006766709 (Bayev and Others v. Russia), para. 88.

Er wordt een onderscheid gemaakt tussen *narrow*, *certain* en *wide margin of appreciation*. Het zal voor een Staat gemakkelijker zijn om een verschil in behandeling te rechtvaardigen in zaken waar het een grote mate van discretionaire bevoegdheid geniet.²⁷⁹ Een grote mate van discretionaire bevoegdheid vinden we veelal terug in zaken betreffende economische of sociale aangelegenheden.²⁸⁰ Daar de nationale hoven beter geplaatst zijn dan het Hof om in deze aangelegenheden de noden van de maatschappij in te schatten en dus een betere afweging kunnen maken van een verdragsrecht tegen andere zaken van nationaal publiek belang (subsidiariteitsbeginsel)²⁸¹, accepteert het Hof doorgaans deze uitspraken.²⁸² Enkel wanneer het Hof van oordeel is dat een uitspraak “*manifestly without reasonable foundation*” is, zal zij oordelen dat de lidstaat de haar toegekende *margin of appreciation* niet gerespecteerd heeft. De doctrine van de *margin of appreciation* is dus een systeem dat de minimumstandaarden van het verdrag waarborgt met oog voor de verschillen in nationale contexten tussen de lidstaten.²⁸³

Naast subsidiariteit hebben ook volgende factoren een invloed op de marge die het Hof aan de lidstaten toekent:

- Het al dan niet bestaan van consensus tussen de lidstaten²⁸⁴
- De aard en het belang van het verdragsrecht in kwestie (waarbij de normen en waarden van de democratische samenleving, menselijke waardigheid en persoonlijke autonomie het zwaarst zullen doorwegen)²⁸⁵

Voorts zal het Hof ook kijken naar de aanwezigheid van procedurele waarborgen om te oordelen of een lidstaat haar beoordelingsmarge geschonden heeft.²⁸⁶

Deze consensus zal dus, onder andere, bepalend zijn voor de mate waarin een lidstaat marge heeft om een verschil in behandeling te rechtvaardigen. Maar ook de culturele, morele en religieuze gevoeligheid van de materie zal hierbij een belangrijke factor zijn. Zo zal er weinig beoordelingsmarge zijn in niet-gevoelige zaken waarover een Europese consensus bestaat en net meer marge bij gevoelige zaken, ook al bestaat er daarrond een Europese consensus.²⁸⁷ Inderdaad, “*this consensus cannot be a decisive factor in the Court’s examination of whether the impugned prohibition ... struck a fair balance between the conflicting rights and interests, notwithstanding an evolutive interpretation of the Convention*”.²⁸⁸

²⁷⁹ J. Gerards (2017). The Margin of Appreciation Doctrine, the Very Weighty Reasons Test and Grounds of Discrimination. M. Balboni (ed) (2017). The principle of discrimination and the European Convention of Human Rights, Editoriale Scientifica, p. 5.

²⁸⁰ EHRM, 5 december 2017, nr. 57101/10, ECLI:CE:ECHR:2017:1205JUD005710110 (Ribac v. Slovenia), para. 53.

²⁸¹ E. Frantziou (2014). The Margin of Appreciation Doctrine in European Human Rights Law, ULC Policy Briefing, p. 1 – 2.

²⁸² EHRM, 15 september 2016, nr. 44818/11, ECLI:CE:ECHR:2016:0915JUD004481811 (British Gurkha Welfare Society and Others v. The United Kingdom), para. 62

²⁸³ Ibid., p. 3.

²⁸⁴ EHRM, 28 november 1984, nr. 8777/79, ECLI:CE:ECHR:1984:1128JUD00877779 (Rasmussen v. Denmark), para. 40.

²⁸⁵ Ibid, p. 7 – 8.

²⁸⁶ EHRM, 13 november 2007, nr. 57325/00, ECLI:CE:ECHR:2007:1113JUD005732500 (D.H. and Others v. The Czech Republic), para. 206.

²⁸⁷ K. Lemmens (2018). The Margin of Appreciation in the ECtHR’s Case Law – A European Version of the Levels of Scrutiny Doctrine?, European Journal of Law Reform, (20) 2-3, p. 92.

²⁸⁸ EHRM, 16 december 2010, nr. 25579/05, ECLI:CE:ECHR:2010:1216JUD002557905 (A, B and C v. Ireland), para. 237.

Voor de zogeheten ‘verdachte gronden’ zal een staat zeer weinig discretionaire bevoegdheid hebben, voor andere gronden dan weer meer.²⁸⁹

Vertaalslag naar AI

Bij het rechtvaardigen van een ongelijke behandeling van mensen die tot één der verdachte categorieën behoren komt het er voor de lidstaat op aan om aan te tonen dat de verdachte grond geen relevante factor is in de beslissing; of dat de verdachte grond relevant is om in overweging te nemen; of dat de maatschappelijke meerwaarde zwaarder doorweegt dan het nadeel ondervonden door de ongelijke behandeling. In wat volgt bespreken we per categorie een aantal zaken en maken we een vertaalslag naar hoe je deze beslissing kan automatiseren.

a. de verdachte grond is geen relevante factor is in de beslissing

Abdulaziz, Cabales en Balkandali v. The United Kingdom: De aanvragers deden beroep op artikel 8 van het EVRM dat het recht op eerbiediging van privé-, familie- en gezinsleven beschermt en claimden een ongelijke behandeling in de zin van artikel 14 op basis van geslacht. In casu werd een maatregel betwist die het voor een in het Verenigd Koninkrijk gevestigde man gemakkelijker maakte om zijn buitenlandse partner te laten overkomen en daar te vestigen dan voor een vrouw. Het Verenigd Koninkrijk wou, in tijden van hoge werkloosheid, de binnenlandse arbeidsmarkt beschermen door enkel vrouwen toe te laten op het grondgebied, daar het statistisch gezien minder waarschijnlijk was dat deze vrouwen werk zouden zoeken en dus de binnenlandse arbeidsmarkt in gevaar zouden brengen. Hoewel het Hof van mening was dat het beschermen van de arbeidsmarkt een legitiem doel was, acht het Hof gendergelijkheid één der belangrijkste doelen van de lidstaten waarvoor bijgevolg zwaarwichtige redenen moeten worden voorgedragen om zulk onderscheid te verantwoorden, wat in deze zaak niet het geval was.²⁹⁰ Het Hof geeft daarbij geen voorbeeld van zwaarwichtige redenen die zulk verschil wel zouden kunnen rechtvaardigen.

Stel nu dat je met het oog op het beschermen van de arbeidsmarkt een AI-systeem wil ontwikkelen dat de migratiestroom zal reguleren. In de ontwikkeling daarvan, en dus in de beslissing om iemand wel of niet toe te laten tot het grondgebied, mag geslacht niet in rekening worden gebracht.

National Union of Belgian Police v. Belgium: In een zaak waarbij een Belgische gemeentelijke politievakbond door de Belgische overheid niet werd uitgenodigd voor het consultatieproces aangaande tewerkstelling in de publieke sector oordeelde het Hof dat het doel dat met deze discriminatie werd nagestreefd legitiem was. Waar de vakbond in kwestie enkel gemeentelijke politieambtenaren vertegenwoordigde, vertegenwoordigden de vakbonden die wel werden uitgenodigd zowel gemeentelijke als provinciale politieambtenaren. De Belgische overheid motiveerde deze keuze door te argumenteren dat het daarmee een situatie van “vakbondsanarchie” wenste te vermijden en achtte deze uitsluiting noodzakelijk om een coherent en gebalanceerd

²⁸⁹ O. M. Arnardóttir (2014). The Differences That Make a Difference: Recent Developments on the Discrimination Grounds and the Margin of Appreciation Under Article 14 of the European Convention on Human Rights (ECHR) 14 Human Rights Law Review 647-670, p. 651.

²⁹⁰ EHRM, 28 mei 1985, nr. 9214/80 9473/81 9474/81, ECLI:CE:ECHR:1985:0528JUD000921480 (Abdulaziz Cabales and Balkandali v. The United Kingdom), para. 78.

ambtenarenbeleid te garanderen. Het Hof oordeelde dat deze uitsluiting op basis van representativiteit van de vakbond gedoogd werd en het nagestreefde doel dus legitiem was.²⁹¹

Swedish Engine Driver's Union v. Sweden: Ook hier was de representativiteit van een vakbond een beslissende factor om al dan niet uitgenodigd te worden voor de ondertekening van een collectieve arbeidsovereenkomst. Het legitieme doel dat hier werd nagestreefd en dat door het Hof werd bevestigd is dat de instantie bevoegd voor de collectieve arbeidsovereenkomst er als algemene regel op nahoudt het aantal onderhandelingspartners te willen beperken teneinde gemakkelijker tot een concreet resultaat te komen en haar inspanningen niet verloren te laten gaan.²⁹²

Indien een overheid de selectie van coalitiepartners wil automatiseren is de representativiteit van die partners een toegestane grond op basis waarvan het systeem een onderscheid mag maken.

D.H. and Others v. The Czech Republic: In deze zaak oordeelde het Hof dat de manier waarop kinderen werden toegewezen aan scholen die normaliter bestemd zijn voor kinderen met beperkte vaardigheden het non-discriminatiebeginsel schendt daar de regeling, hoewel neutraal geformuleerd, een onevenredig hoog aantal Roma kinderen trof in vergelijking met de rest van de bevolking. De toewijzing gebeurde immers door middel van psychologische testen die de intellectuele capaciteiten van de kinderen zou moeten inschatten. Te gevolge getuigenissen van verschillende tussenkomende derde partijen besloot het Hof dat deze tests op zijn minst het risico in zich hielden bevooroordeeld te zijn en werden geanalyseerd zonder inachtneming van de bijzondere kenmerken van de Roma kinderen in kwestie. Het Hof volgt daarmee de redenering van de Staat dat de psychologische tests een rechtvaardiging zijn voor het verschil in behandeling niet.²⁹³ Evenmin was het Hof akkoord met het argument dat de ouders van de kinderen uiteindelijk het beslissende en doorslaggevende woord hadden over de plaatsing van de kinderen in de scholen omdat duidelijk was dat er geen sprake was van een degelijk geïnformeerde beslissing.²⁹⁴ Voorts redeneert het Hof dat als een laatste beslissend woord ervoor zou kunnen zorgen dat iemand instemt met een verschil in behandeling, ook al is dat verschil in behandeling discriminerend, dit zou neerkomen op een afstand doen van het recht niet gediscrimineerd te worden en benadrukt dat dit geenszins mogelijk is voor eender welk verdragsrecht.

Indien een overheid de toewijzing van kinderen aan scholen wil automatiseren is het uiteraard niet toegestaan om hierin een onderscheid te maken op basis van ras of etniciteit. Wat wel toegestaan is, zoals uit bovenstaand voorbeeld blijkt, is om de toewijzing te laten afhangen van de uitkomst van psychologische testen. De procedurele waarborgen waaraan zulke testen moeten voldoen zijn de volgende:

²⁹¹ EHRM, 27 oktober 1975, nr. 4464/70, ECLI:CE:ECHR:1975:1027JUD000446470 (National Union of Belgian Police v. Belgium), para. 48.

²⁹² EHRM, 6 februari 1976, nr. 5614/72, ECLI:CE:ECHR:1976:0206JUD000561472 (Swedish Engine Driver's Union v. Sweden), para. 46.

²⁹³ EHRM, 13 november 2007, nr. 57325/00, ECLI:CE:ECHR:2007:1113JUD005732500 (D.H. and Others v. The Czech Republic), para. 201.

²⁹⁴ Ibid, para. 202 – 203.

- Er moet een zekere mate van consensus zijn over de validiteit van psychologische tests om de intellectuele capaciteiten van studenten in te schatten
- Psychologische tests moeten gebenchmarkt zijn op de gehele bevolking, niet enkel op een deel van de bevolking (of psychologische tests moeten inclusief opgesteld worden)

Deze waarborgen zouden er dan voor moeten zorgen dat bepaalde kwetsbare groepen in de samenleving door een algemeen geformuleerde regel niet indirect gediscrimineerd worden.

Deaconu v. Romania: Het Hof achtte de beslissing om een verschil in schadevergoeding voor immateriële schade toe te kennen op basis van leeftijd onvoldoende gefundeerd daar deze beslissing werd genomen op basis van één getuigenis, zonder de benadeelde personen in kwestie te horen. Er wordt verder ook vermeld dat de aanwezigheid van een experten rapport en psychologische evaluatie zulke beslissing zouden hebben kunnen rechtvaardigen, maar dat het gezien de afwezigheid ervan om een arbitraire bepaling ging over de leeftijd waarop iemand immateriële schade kan oplopen door het verlies van een naaste.²⁹⁵

Hieruit kunnen we afleiden dat niet leeftijd, maar wel de uitkomst van een psychologische evaluatie en een experten rapport toegestane gronden zouden kunnen zijn op basis waarvan men mag discrimineren om te bepalen of iemand al dan niet immateriële schade opgelopen heeft.

Cînta v. Romania en S.S. v. Slovenia: Waar uit het voorgedragen bewijsmateriaal in S.S. v. Slovenia duidelijk was dat de aanklager in kwestie te gevolge haar mentale toestand niet in staat was om voor haar kind te zorgen, was dat in Cînta v. Romania niet duidelijk daar de staat naliel de specifieke omstandigheden van de zaak na te gaan en enkel steunde op de mentale toestand van de aanklager.²⁹⁶

Hieruit blijkt het belang om naar de omstandigheden van het geval te kijken. Waar beide zaken gaan over de zorg van een kind door een persoon die geestesziek is, was het in de eerste zaak duidelijk dat de persoon in kwestie daar niet toe in staat was. In de tweede zaak was dat echter allerm minst duidelijk. Beide gevallen gelijk beoordelen op grond van het feit dat beide geestesziek zijn zou dus niet rechtvaardig zijn. Voor de hypothetische automatisering van het proces om te bepalen of iemand voor zijn/haar/... kind kan zorgen is de opname van 'geestesziekte' in de berekening dus niet toegestaan. De parameter 'kan voor zijn/haar/... kind zorgen' zou hier dan wel correct zijn.

In Leonov v. Russia zien we dat een rechterlijke beslissing om een kind bij de moeder te laten wonen, ten nadele van de vader, gerechtvaardigd werd door het feit dat het kind slechts 3 jaar oud was en altijd bij de moeder gewoond had. Het kind op die jonge leeftijd uit zijn omgeving trekken zou ernstige psychologische gevolgen met zich kunnen meedragen, zelfs al waren de levensomstandigheden bij de vader gunstiger en had hij geen strafblad.²⁹⁷

²⁹⁵ EHRM, 29 januari 2019, nr. 66299/12, ECLI:CE:ECHR:2019:0129JUD006629912 (Deaconu v. Romania), para. 35 - 39

²⁹⁶ EHRM, 18 februari 2020, nr. 3891/19, ECLI:CE:ECHR:2020:0218JUD000389119 (Cînta v. Romania), para. 67 - 73

²⁹⁷ EHRM, 10 april 2018, nr. 77180/11, ECLI:CE:ECHR:2018:0410JUD007718011 (Leonov v. Russia), para. 88

De toewijzing van het hoederecht mag dus niet gebeuren op basis van geslacht maar wel op basis van de geprojecteerde psychologische gevolgen voor het kind. In bovenstaande voorbeelden kan men dus de verdachte grond weglaten uit de berekening om niet discriminerend te zijn. In wat volgt zien we dat de verdachte gronden soms echter wel in de berekening dienen te worden opgenomen.

b. de verdachte grond is relevant om in overweging te nemen

Hieronder volgen voorbeelden waar de verdachte grond wel relevant is om mee te nemen in de berekening van het AI-systeem.

Stoian v. Romania: Het Hof oordeelde dat de lidstaat voorzag in redelijke accommodatie en toewijzing van middelen om aan de educatieve noden van kinderen met een beperking tegemoet te komen.²⁹⁸

Enver Şahin v. Turkey: Het Hof maakte duidelijk dat het de verantwoordelijkheid is van een lidstaat om ervoor te zorgen dat ieder zijn of haar recht op onderwijs op gelijke voet zou moeten kunnen uitoefenen. In casu verzaakte de lidstaat aan haar verantwoordelijkheid ieder dit recht te verzekeren daar het onmogelijk was voor iemand die aan dwarlaesie²⁹⁹ lijdt zelfstandig de universiteitsgebouwen te betreden. De voorgestelde oplossing om een assistent aan te stellen die de persoon in kwestie zou helpen met toegang tot de gebouwen was niet voldoende om dit nadeel in balans te brengen. Het aanpassen van de universiteitsinfrastructuur aan de noden van de student had dat wel geweest. Het recht op onderwijs zonder enig onderscheid houdt dus, onder andere, in dat iedereen dit recht op een zelfstandige wijze moet kunnen uitoefenen.³⁰⁰

Stel dat scholen de toewijzing van hun budget tot op zekere hoogte willen automatiseren is het hier toegestaan om de parameter ‘kinderen met een beperking’ of ‘iemand die lijdt aan dwarslaesie’ in de berekening mee te nemen.

In Alexandru v. Romania meende de aanklager gediscrimineerd te worden op grond van zijn geslacht daar vrouwelijke gevangenen met kinderen tot één jaar oud een uitstel van strafuitvoering konden aanvragen waar mannelijke gevangenen in dezelfde situatie dat niet konden. Het legitieme doel dat met deze regel werd nagestreefd was om de belangen van pasgeborenen te beschermen. Door rekening te houden met de specifieke persoonlijke situaties, waaronder de zwangerschap van vrouwelijke gevangenen en de periode voorafgaand aan de eerste verjaardag van de baby, met name gelet op de speciale banden die tijdens die periode bestaan tussen de moeder en het kind, werd het verschil in behandeling gerechtvaardigd.³⁰¹

²⁹⁸ EHRM, 25 juni 2019, nr. 289/14, ECLI:CE:ECHR:2019:0625JUD000028914 (Stoian v. Romania), para. 110

²⁹⁹ Letsel van het ruggenmerg of van de in het wervelkanaal lopende zenuwbanen met een tijdelijk of permanent verlies van de motorische, sensorische of autonome functie tot gevolg. Bron: <https://www.uzleuven.be/nl/dwarslaesie>

³⁰⁰ EHRM, 30 januari 2018, nr. 23065/12, ECLI:CE:ECHR:2018:0130JUD002306512 (Enver Şahin v. Turkey), para. 57 - 69

³⁰¹ EHRM, 3 oktober 2017, nr. 16986/12, ECLI:CE:ECHR:2017:1003JUD001698612 (Alexandru Enache v. Romania), para. 76

Ook hier zien we een voorval waarin een verdachte grond, in casu geslacht, wel relevant is om mee te nemen in de berekening voor de hypothetische automatisering van de aanvragen voor strafuitstel daar enkel vrouwen drager van een kind kunnen zijn.

c. de maatschappelijke meerwaarde weegt zwaarder door dan het nadeel ondervonden door de ongelijke behandeling

In Dakir v. Belgium was de aanklager van mening dat de algemene bewoording van een statuut dat het volledig verbergen van het aangezicht verbiedt indirect discriminerend was ten aanzien van moslima's die een volledig bedekkende hoofddoek willen dragen in het openbaar als uiting van dier geloof. Ook hier is sprake van discriminatie in de zin van artikel 14 in afwezigheid van een objectieve en redelijke rechtvaardiging bij wijze van een legitiem doel. Gezien het gebrek aan consensus tussen de lidstaten over het al dan niet verbieden van volledige gezichtsbedekking kent het Hof hier aan de lidstaten een zeer grote mate van discretionaire bevoegdheid toe.³⁰² Het legitiem doel dat hier door de maatregel werd nagestreefd was het beschermen van een vorm van interactie tussen individuen die volgens de overheid essentieel was voor het functioneren van een democratische samenleving.³⁰³

Stel dat je als overheid een zwarte lijst wil opstellen van mensen wier gedrag incompatibel is met de democratische samenleving. Het opnemen van de parameter 'draagt hoofddoek die gezicht volledig bedekt' is hier toegestaan, ook al is dat indirect discriminerend ten aanzien van moslima's, daar dit een vorm van interactie die essentieel is voor het functioneren van een democratische samenleving verhindert. Het opnemen van de parameter 'moslima' is uiteraard niet toegestaan. Hier zien we dus dat het maatschappelijke meerwaarde zwaarder doorweegt dan het nadeel ondervonden door de ongelijke behandeling.

Conclusie

Door de manier waarop een lidstaat een verschil in behandeling kan verantwoorden als leidraad te nemen wordt duidelijk dat een algemene vertaling van non-discriminatie principes naar concreet toe te passen AI-systeem principes niet mogelijk is omdat discriminatiegronden soms wel relevant zijn om in overweging te nemen en soms niet. De exacte voorwaarden die maken dat een bepaalde discriminatiegrond dan wel, dan niet in overweging dient genomen te worden zijn eveneens moeilijk vast te stellen. In andere gevallen dient het belang van non-discriminatie dan weer afgewogen te worden tegen andere belangen. Men zal dus altijd naar de omstandigheden van het geval moeten kijken om uit te maken of iets discriminerend is of niet.

Janneke Gerards – Rechterlijke toetsing aan het gelijkheidsbeginsel³⁰⁴

Analoog aan de poging een vertaalslag van jurisprudentie naar AI-systeem principes te maken zocht Janneke Gerards in 2002 in haar proefschrift naar een vergelijkbaar algemeen toetsingskader. In wat volgt wordt een kort overzicht gegeven van dat toetsingskader waaruit wederom blijkt dat de omstandigheden van het geval veelal bepalend zullen zijn voor de uitkomst van de zaak.

³⁰² EHRM, 11 juli 2017, nr. 4619/12, ECLI:CE:ECHR:2017:0711JUD000461912 (Dakir v. Belgium), para. 59

³⁰³ Ibid, para. 56

³⁰⁴ J.H. Gerards (2002). Rechterlijke toetsing aan het gelijkheidsbeginsel – Een rechtsvergelijkend onderzoek naar een algemeen toetsingsmodel, Sdu Uitgevers.

Daar het ontbreken van duidelijke, algemene regels omtrent gelijke behandeling in vele rechtssystemen de voorspelbaarheid en objectiviteit van een rechterlijke beslissingen in de weg staat, nam Janneke Gerards in haar proefschrift het initiatief hiervoor een algemeen toetsingsmodel te ontwikkelen. Dit om de problemen van ongelijke behandeling in gelijke situaties en het sluipen van vooroordelen en persoonlijke meningen in de rechterlijke beslissing aan te pakken. Zodoende wordt de uitkomst van iedere zaak meer voorspelbaar en ontstaat er meer rechtszekerheid. Dit model dient als leidraad bij de rechterlijke redenering om het algemene, ongeschreven principe van gelijkheid meer praktische houvast te geven en gaat uit van het idee dat de gelijkheidstoets altijd in 3 stappen gebeurt, zijnde: de vaststelling dat een zaak ongelijke behandeling betreft; de bevinding dat een verschil in behandeling al dan niet gerechtvaardigd is; en tot slot het oordeel om het verschil in behandeling te compenseren via een schadevergoeding, een teniet verklaring van de rechtsregel die het verschil in behandeling bewerkstelligt of een annulering van de beslissing die een verschil in behandeling tot gevolg had. Het model is gebaseerd op een analyse van rechtspraak van het Europees Hof voor de Rechten van de Mens, het Europees Gerechtshof, het Amerikaans Hooggerechtshof en verschillende Nederlandse rechterlijke instanties.

Intensiteit van de gelijkheidstoets

Alvorens over te gaan tot deze drie stappen dient de intensiteit van de test te worden vastgesteld. Deze intensiteit zal consistent zijn doorheen de drie bovenvermelde stappen. Er wordt een onderscheid gemaakt tussen een marginale, een neutrale en een intensieve test.

Bij een marginale test oefent het Hof de grootste mate van terughoudendheid uit. Deze wordt uitgevoerd in zaken waarbij de vermeende ongelijke behandeling geen serieuze gevolgen heeft door bijvoorbeeld het bestaan van alternatieven die het effect van de behandeling teniet doen, in zaken waarbij het beslissingsorgaan een grote mate van discretionaire bevoegdheid heeft, in zaken waarbij het beslissingsorgaan in een betere positie is dan de rechtbank om de noodzakelijkheid van een maatregel in te schatten, of in zaken waar de beslissing voor ongelijke behandeling genomen werd in een noodsituatie.

Aan de andere kant van het spectrum vinden we een intensieve test. Deze wordt uitgevoerd wanneer er gronden zijn aan te nemen dat het besluitvormingsproces dat leidde tot de beslissingen werd beïnvloed door inherent discriminatoire of bevooroordeelde overwegingen. Dit zal veelal het geval zijn voor de zogeheten ‘verdachte gronden’. De maatschappelijke irrelevantie en de onveranderlijkheid van het persoonlijk kenmerk zullen indicatief zijn voor het kwalificeren van een grond als verdacht. Ook wanneer er gronden zijn aan te nemen dat het verschil in behandeling zou voortvloeien uit minachting voor een sociaal achtergestelde of politiek niet-significante groep, of wanneer de fundamentele rechten van een bepaalde groep geschonden werden, zal men doorgaans voor een intensieve test kiezen.

Wanneer geen van beide bovenvermelde factoren de keuze voor een marginale, dan wel intensieve test, noodzakelijk maken, wordt er een neutrale test uitgevoerd.

Stap 1: Vaststelling van de ongelijke behandeling

Uit het onderzoek blijkt dat de hoven kiezen voor ofwel de vergelijkingstoets ofwel de benadeelde behandelingstoets om een verschil in behandeling vast te stellen. Daar de vergelijkingstoets enkele kenmerken in zich draagt die de voorspelbaarheid en rechtszekerheid van de uitkomst in het gedrang brengen, argumenteert Gerards dat de benadeelde

behandelingstoets een gepaster instrument is om een vaststelling van ongelijke behandeling te doen. Eénmaal een direct, dan wel indirect benadeelde behandeling is vastgesteld kan men overgaan tot het onderzoek of deze behandeling gerechtvaardigd is. Als er geen benadeelde behandeling wordt vastgesteld kan het Hof de claim verwerpen.

Stap 2: Rechtvaardiging voor de ongelijke behandeling

Het oordeel of een ongelijke behandeling al dan niet gerechtvaardigd is komt voort uit de beoordeling van volgende criteria. Voor elk criteria is de beoordeling verschillend naargelang de intensiteit van de toets.

1) Legitiem doel

Het vertrekpunt hier is het doel dat wordt voorgebracht door de verweerder, enkel in zaken waarbij de Hof een intensieve toets noodzakelijk acht zal het uit eigen beweging nagaan of er andere redenen zouden kunnen zijn voor de ongelijke behandeling. Daarna gaat het Hof de legitimiteit van dat doel na door te kijken naar de consistentie ervan met nationaal en internationaal recht, de verwezenlijking van dat doel zonder dat daarbij de bevoegdheid van het beslissingsorgaan overschreden werd, of het doel op zich discriminatoir was, en of dat doel een oprechte noodzaak van de verweerder bekwaam.

In gevallen waar meerdere doelen worden nagestreefd en waarvan er één niet legitiem bevonden wordt zal de beoordeling van de overige doelen beslissend zijn in het al dan niet kwalificeren van een behandeling als ongelijk.

2) Verhouding tussen het doel en de middelen

In deze stap zal worden nagegaan of het gemaakte onderscheid op basis waarvan een ongelijke behandeling voortkwam met voldoende precisie bekomen werd. Wanneer hierbij, door de makers van het onderscheid, gebruik werd gemaakt van *proxies*, is het noodzakelijk dat deze voldoende inclusief zijn voor het bereiken van het beoogde doel om over- of onder-inclusiviteit te vermijden. Bij over-inclusiviteit omvat de gebruikte proxy een te brede groep waardoor een te groot aantal mensen dan nodig voor het beoogde doel nadelig behandeld wordt. Onder-inclusiviteit zal tot gevolg hebben dat een bepaalde groep mensen geen voordelige behandeling geniet terwijl dat eigenlijk wel het geval had moeten zijn. De mate waarin het Hof onder- of over-inclusiviteit accepteert is wederom afhankelijk van de intensiteit van de toets. Bij een marginale toets zal er meer ruimte zijn voor *misfit* dan bij een intensieve toets. Ontoelaatbare mate van *misfit* resulteren onherroepelijk in de bevinding dat het gelijkheidsbeginsel geschonden werd.

3) Effectiviteit van de ongelijke behandeling om het legitieme doel te bereiken

Wanneer het Hof heeft vastgesteld dat het onderscheid waaruit een ongelijke behandeling voortkwam met voldoende precisie bekomen werd, zal het overgaan tot de beoordeling of dit onderscheid geschikt is. Dit wil zeggen dat het Hof zal overgaan tot de beoordeling of het gemaakte onderscheid effectief is het bereiken van het beoogde, legitieme doel. Ook hierbij is de intensiteit van de toets bepalend. Bij een marginale toets wordt de effectiviteit verondersteld, bij een intensieve toets zal de verweerder veelal concrete en controleerbare feiten naar voren moeten dragen waaruit de effectiviteit van het onderscheid objectief kan worden vastgesteld.

4) Noodzakelijkheid of subsidiariteit

Vervolgens gaat het Hof over tot de noodzakelijkheids- dan wel de subsidiariteitstoets. Naast de omstandigheden van de zaak zal ook hier de intensiteit van de toets bepalend zijn in de keuze van het Hof voor één der twee toetsen. In zaken waarbij het Hof een marginale of neutrale toets voldoende acht zal het een noodzakelijkheidstoets uitvoeren, bij een intensieve toets zal het Hof ook nagaan of er andere, minder drastische of ingrijpende maatregelen voor handen waren die hetzelfde doel bereikten.

5) Proportionaliteit

Tot slot gaat het Hof over tot de proportionaliteitstoets waarin het eerst nagaat welke belangen er langs beide kanten spelen. Vervolgens gaat het Hof deze belangen wegen. In deze weging zal het Hof zaken in acht nemen zoals de ernst van de geledede schade door de ongelijke behandeling of het aanwezig zijn van schending van een fundamenteel recht. Daarna zal het Hof beoordelen of er in de beslissing of rechtsregel die leidde tot de ongelijke behandeling voldoende aandacht besteed is geweest aan een redelijke afweging tussen de verschillende belangen. Ook hier is de intensiteit van de toets bepalend in de zin dat het Hof bij een marginale toets slechts zal oordelen of de belangenafweging onredelijk en arbitrair. Bij een intensieve toets zal het Hof echter nagaan of de belangen van de verweerder voldoende zwaar en overtuigend zijn om een ongelijke behandeling, en dus een schending van de legitieme belangen van een persoon of een groep van personen, te rechtvaardigen.

3.3.2 EU HvJ

Het beginsel van non-discriminatie vindt zijn grondslag in artikel 18 VWEU. Het Europese non-discriminatiebeginsel behoort ook tot een van de pijlers van het Europese recht en de rechtspraak van het Hof leert dat daaraan een brede uitleg gegeven dient te worden. Het Hof maakt aan de hand van zijn uitspraken duidelijk dat het vergaande consequenties kan hebben. Het begrip indirecte discriminatie, vindt zijn oorsprong in de jurisprudentie van het Hof en heeft later zijn weg gevonden in de wetgeving rondom discriminatie.

(In)directe werking

Sinds de jaren '70 van de vorige eeuw heeft het Hof van Justitie buigt het Hof zich over zaken aangaande discriminatie. In een van de eerste uitspraken van het Hof is de gelijke behandeling tussen mannen en vrouwen inzake beloning als algemeen beginsel van gemeenschapsrecht erkend.³⁰⁵ In dit arrest, Defrenne II, heeft het Hof geoordeeld dat er sprake is van directe horizontale werking van het gemeenschapsrecht in de verhouding tussen werkgever en werknemer (artikel 119 VWEU), de nationale rechter moet het toepassen in zijn rechtspraak en moet de eerbiediging ervan waarborgen, anderzijds kunnen justitiabelen er subjectieve rechten aan ontleen, waarop zij zich voor een rechter kunnen beroepen. Dit is een principiële zaak voor het Hof, omdat het uitspraak heeft gedaan over de directe horizontale werking in de verhouding tussen werkgever en werknemer. Discriminatie op grond van geslacht heeft derhalve directe werking. Een particulier kan dit dus voor de nationale rechter inroepen ten aanzien van een andere burger of ten aanzien van de overheid. Bovendien heeft het Hof in dit

³⁰⁵ HvJ 8 april 1976, ECLI:EU:C:1976:56.

arrest onderscheid gemaakt tussen enerzijds ‘rechtstreekse en openlijke’ discriminatie en anderzijds ‘zijdelingse en verkapte’ discriminatie. Rechtstreekse discriminatie heeft ook rechtstreekse werking, terwijl niet-rechtstreekse discriminatie niet binnen de toepassingsfeer kunnen worden vastgesteld. Het Hof heeft in deze zaak de fundamentele waarde van gelijke beloningen tussen mannen en vrouwen bevestigd.

Op 10 april 1984 deed het Hof in twee zaken, namelijk Von Colson/Kamann³⁰⁶ en Harz³⁰⁷. In Von Colson omzeilde het Hof de rechtstreekse werking van de richtlijn 76/207 maar kwam uiteindelijk wel tot een gelijkaardig resultaat als de richtlijn wel rechtstreekse werking zou hebben. Het Hof was van oordeel dat de richtlijn met betrekking tot de op discriminatie te stellen sancties geen onvoorwaardelijke en voldoende nauwkeurige verplichting bevatte waarop een particulier zich zou kunnen beroepen om krachtens deze richtlijn een schadevergoeding te krijgen.³⁰⁸ Het Hof heeft in deze zaak de indirecte werking van de richtlijn bevestigd, maar kende aan deze richtlijn geen verticale werking toe. Verder heeft het Hof in deze zaak bevestigd dat het aan de nationale rechter is om ten volle gebruik te maken van zijn beoordelingsvrijheid en daardoor de ter uitvoering van de richtlijn vastgestelde wet uit te leggen en toe te passen in overeenstemming met de eisen van het gemeenschapsrecht.³⁰⁹ Diezelfde dag deed het Hof dus ook uitspraak in de zaak Harz, een praktisch identieke zaak als Von Colson/Kamann. De redenering en antwoorden in deze zaak zijn dan ook identiek, namelijk wordt er in beide arresten door het Hof geoordeeld dat de richtlijn niet verlangt dat als sanctie op de gemaakte discriminatie bij de toegang tot het arbeidsproces, de werkgever verplicht wordt om de gediscrimineerde sollicitant een arbeidsovereenkomst aan te bieden.³¹⁰ Bovendien oordeelt het Hof dat de richtlijn met betrekking tot de sancties geen onvoorwaardelijke en voldoende nauwkeurige verplichting heeft waarop een particulier zou kunnen beroepen.³¹¹ Als laatste wordt in beide zaken door het Hof geoordeeld dat de richtlijn lidstaten vrijlaat om voor een schending van het discriminatieverbod een sanctie te stellen, waarbij zij een keuze hebben uit verschillende oplossingen die geschikt zijn om het doel van de richtlijn te bereiken. Hierbij moet rekening worden gehouden met dat de sanctie in passende verhouding moet staan tot de geleden schade.³¹² Het enige verschil tussen deze twee zaken is dat in geval van Von Colson/Kamann, de richtlijn werd ingeroepen tegen een publieke werkgever, wat een verticale relatie inhoudt. Terwijl in Harz de richtlijn werd ingeroepen tegen een particuliere werkgever, wat een horizontale relatie inhoudt. Het Hof besloot in beide gevallen geen uitspraak te doen over de verticale dan wel horizontale werking van de richtlijn, maar oordeelde dat de richtlijn indirecte werking heeft en dat het aan de nationale rechter is, door zijn toegekende beoordelingsvrijheid, om de richtlijn uit te leggen en toe te passen conform het gemeenschapsrecht.³¹³

Indirect onderscheid inzake geslacht

Na Defrenne II sprak het Hof zich later uit in het arrest Jenkins.³¹⁴ Jenkins was een parttime werknemer bij een Engels bedrijf. De parttime werknemers kregen een loon dat 10% lager was

³⁰⁶ HvJ 10 april 1984, ECLI:EU:C:1984:153.

³⁰⁷ HvJ 10 april 1984, ECLI:EU:C:1984:155.

³⁰⁸ HvJ 10 april 1984, ECLI:EU:C:1984:153, r.o. 27.

³⁰⁹ HvJ 10 april 1984, ECLI:EU:C:1984:153, r.o. 28.

³¹⁰ HvJ 10 april 1984, ECLI:EU:C:1984:153, r.o. 18 en HvJ 10 april 1984, ECLI:EU:C:1984:155, r.o. 18.

³¹¹ HvJ 10 april 1984, ECLI:EU:C:1984:153, r.o. 27 en HvJ 10 april 1984, ECLI:EU:C:1984:155, r.o. 27.

³¹² HvJ 10 april 1984, ECLI:EU:C:1984:153, r.o. 28 en HvJ 10 april 1984, ECLI:EU:C:1984:155, r.o. 28.

³¹³ HvJ 10 april 1984, ECLI:EU:C:1984:153, r.o. 28.

³¹⁴ HvJ 31 maart 1981, ECLI:EU:C:1981:80.

dan fulltime werknemers. Opvallend was dat bijna alle parttime werknemers vrouwen waren. Uiteindelijk oordeelde het Hof dat *“een verschil in beloning tussen voltijd- en deeltijdwerknemers slechts een door artikel 119 EEG-Verdrag verboden discriminatie is, wanneer het in werkelijkheid enkel een indirect middel is om het loonpeil van deeltijdwerknemers te verlagen wegens de omstandigheid dat deze groep werknemers uitsluitend dan wel voornamelijk uit vrouwen bestaat.”*³¹⁵ Hiermee bedoelt het Hof dat een onderscheid gemaakt mag worden *“zolang het maar zijn verklaring vindt in objectieve gerechtvaardigde factoren die geen verband houden met discriminatie op grond van geslacht”*.³¹⁶

In bovenstaande zaak komt als eerste naar voren dat een indirect onderscheid verboden is, tenzij dit objectief gerechtvaardigd wordt door factoren die geen verband houden met discriminatie. Het is dus, volgens het Hof, mogelijk onderscheid tussen bepaalde mensen te maken, mits dit gerechtvaardigd is en dit voor een rechtmatig doel wordt gebruikt. In sommige gevallen kan de discriminatie voorrang krijgen op het fundamentele beginsel van gelijke behandeling.

In latere jurisprudentie wordt vaker gesproken over een objectieve rechtvaardiging door middel van een legitiem doel inzake indirecte discriminatie, is deze term voor langere tijd wel redelijk vaag gebleven. Het Hof verwijst naar deze term maar geeft verder geen duidelijke uitleg over wat het precies inhoudt, totdat het Hof in de Bilka³¹⁷ uit 1986 uitspraak had gedaan. Ook in dit geval betrof het een zaak waarbij er onderscheid werd gemaakt tussen fulltime en parttime werknemers, alleen fulltime werknemers konden namelijk aanspraak doen op een bedrijfspensioen. Mevrouw Von Hartz in deze zaak argumenteerde dat deze regel over het algemeen voornamelijk vrouwen benadeelde, omdat vrouwen vaker parttime werkten dan mannen. Bilka, de voormalige werkgever van mevrouw Von Hartz, betoogde juist dat haar geen schending van het beginsel van gelijke beloning kon worden verweten.³¹⁸ Het Hof heeft in deze zaak uitsluitend gegeven over wat precies onder die ‘objectieve rechtvaardiging door een legitiem doel’ wordt verstaan. Zo oordeelde het Hof dat Bilka artikel 119 VWEU heeft geschonden wanneer de maatregel een veel groter aantal vrouwen dan mannen treft, *“tenzij de onderneming aantoont dat de bedoelde maatregel haar verklaring vindt in factoren die objectief gerechtvaardigd zijn en niets van doen hebben met discriminatie op grond van geslacht”*. Vervolgens heeft het Hof meer duidelijkheid gegeven over de ‘objectieve rechtvaardiging’. Het Hof zegt hierover dat de beoordeling van deze objectieve rechtvaardiging toebedeeld is aan de nationale feitenrechter.³¹⁹

Hierbij is belangrijk dat het enkel aanvoeren van niet-discriminatiegronden onvoldoende is. De ondernemer, in dit geval, moest bewijzen dat het loonbeleid verklaard kon worden door objectief gerechtvaardigde motieven en dat dit middel daadwerkelijk aan een bepaalde behoefte van de onderneming beantwoordde en dat het beleid daarnaast geschikt was om het door haar beoogde doel te bereiken en dat het daarvoor ook noodzakelijk was,³²⁰ hierbij spreken we van doelmatigheid, terwijl dat doel niet te bereiken is op een andere wijze waarbij geen onderscheid gemaakt wordt (proportionaliteit). Samenvattend houdt dit het volgende in:

³¹⁵ HvJ 31 maart 1981, ECLI:EU:C:1981:80, r.o. 15.

³¹⁶ HvJ 31 maart 1981, ECLI:EU:C:1981:80, r.o. 11.

³¹⁷ HvJ 13 mei 1986, ECLI:EU:C:1986:204.

³¹⁸ HvJ 13 mei 1986, ECLI:EU:C:1986:204, r.o. 7.

³¹⁹ HvJ 13 mei 1986, ECLI:EU:C:1986:204, r.o. 36.

³²⁰ HvJ 13 mei 1986, ECLI:EU:C:1986:204, r.o. 36.

- er is voldaan aan indirect onderscheid als de discriminator een middel heeft aangewend dat op het eerste zicht neutraal is en niets te maken heeft met bijvoorbeeld een onderscheid op basis van geslacht, zoals dus in voorgaande zaken;
- maar dit middel wel leidt tot een onderscheid tussen twee groepen, waarbij de ene groep meer benadeeld wordt dan de andere.

Dit is verboden, tenzij:

- het door een objectieve rechtvaardiging door een legitiem doel wel mag. Onder de objectieve rechtvaardiging valt dat de discriminator een objectieve rechtvaardigingsgrond kan aanvoeren waarmee hij verantwoording geeft over het verschil in behandeling. Hierbij is het belangrijk dat de discriminator aantoont dat hij met dit middel een rechtmatig doel nastreeft dat cruciaal is om voorrang te krijgen op het beginsel van gelijke behandeling (artikel 119 VWEU);
- daarnaast moet de middelen noodzakelijk en evenredig zijn om dit doel te bereiken (proportionaliteit).³²¹

Het Hof heeft in dit arrest alle verwarring rondom de ‘objectieve rechtvaardiging’ weggenomen door duidelijkheid te geven over wat dit precies inhoudt. Dit arrest wordt in de literatuur ook wel gekenmerkt als hét sleutelarrest inzake indirecte discriminatie en de daaropvolgende objectieve rechtvaardiging.

Direct onderscheid inzake leeftijd

Ook het beginsel van non-discriminatie op grond van leeftijd wordt gekwalificeerd als een algemeen beginsel van gemeenschapsrecht, net zoals dat het geval is bij non-discriminatie op grond van geslacht. Dit is in het arrest Mangold³²² bepaald. Het Hof heeft ook geoordeeld dat het gemeenschapsrecht zich verzet tegen een nationale regeling op grond van waarvan het sluiten van arbeidsovereenkomsten voor bepaalde tijd onbeperkt is toegestaan wanneer de werknemer 52 jaar of ouder is, tenzij er een nauw zakelijk verband bestaat met een eerdere arbeidsovereenkomst voor onbepaalde tijd met dezelfde werkgever.³²³ In Mangold is tevens bepaald dat het aan de nationale rechter is om de volle werking van het algemene beginsel van non-discriminatie op grond van leeftijd te verzekeren door elke strijdige bepaling van nationaal recht buiten toepassing te laten, ook wanneer de termijn voor omzetting van deze richtlijn nog niet is verstreken.³²⁴

Een ander belangrijk arrest inzake non-discriminatie op grond van leeftijd is het Swedex-arrest.³²⁵ Dit arrest is een voortzetting van het Mangold-arrest en wordt bovendien gezien als een mijlpaal-arrest. In dit arrest ging het om een vrouw van 25 jaar die na ruim 10 jaar dienstverband ontslagen werd door het bedrijf Swedex. De vrouw kreeg een opzegtermijn van slechts één maand omdat bij het bepalen van de opzegtermijn, het dienstverband voor de 25^e levensjaar van de werknemer niet werd meegenomen. In dit arrest paste het Hof een analyse toe:

- Als eerste wordt vastgesteld of er een onderscheid wordt gemaakt;

³²¹ HvJ 13 mei 1986, ECLI:EU:C:1986:204, r.o. 35.

³²² HvJ 22 november 2005, ECLI:EU:C:2005:709, r.o. 75.

³²³ HvJ 22 november 2005, ECLI:EU:C:2005:709, r.o. 78.

³²⁴ HvJ 22 november 2005, ECLI:EU:C:2005:709, r.o. 78.

³²⁵ HvJ 19 januari 2010, ECLI:EU:C:2010:21.

- Daarna kijkt het Hof of er sprake is van een gerechtvaardigd verschil in behandeling op grond van, in dit geval artikel 6 lid 1 van de Richtlijn 2000/78/EG;
 - o Deze richtlijn geeft de mogelijkheid een direct onderscheid op grond van leeftijd objectief te rechtvaardigen, waarbij ook weer wordt getoetst op er sprake is van (1) een legitiem doel voor de verschillende behandeling, waarbij rekening wordt gehouden met legitieme doelstellingen van beleid op het gebied van werkgelegenheid, arbeidsmarkt of beroepsopleiding, daarnaast bekijkt het Hof ook of de gebruikte middelen om dat doel te bereiken (2) passend en (3) noodzakelijkheid zijn.³²⁶

Het Hof heeft in het Swedex-arrest vijf jaar na het Mangold-arrest uitspraak gedaan en biedt meer duidelijkheid en verbetering, door onder andere de hierboven besproken toets. Ook in deze zaak verzet het gemeenschapsrecht, met in het bijzonder het beginsel van non-discriminatie op grond van leeftijd, zich tegen een nationale regeling die het, zoals in deze zaak, mogelijk maakt om de tijdvakken van arbeid die een werknemer voor het bereiken van de 25 jaar heeft vervuld, niet in aanmerking worden genomen bij de berekening van de duur van de opzegtermijn.³²⁷ Ook hier wordt geoordeeld dat de nationale rechter de naleving moet verzekeren van het beginsel van non-discriminatie op grond van leeftijd, daarnaast komt het ook de nationale rechter toe om indien nodig elke strijdige bepaling van nationale regeling buiten beschouwing te laten.³²⁸

Opvallend is dat zowel het Mangold-arrest als het Swedex-arrest, net als Defrenne II inzake non-discriminatie op grond van geslacht, een uitspraak is inzake de horizontale doorwerking van richtlijnen in de nationale orde. In Mangold heeft het Hof tevens geoordeeld dat voor zover een richtlijn een concretisering vormt voor een algemeen beginsel, zoals in deze zaak het verbod op leeftijdsdiscriminatie, dit algemene beginsel in horizontale verhoudingen kan worden toegepast alsmede nationale bepalingen die hiermee in strijd zijn buiten toepassing moeten worden gelaten. In zo'n geval werkt het algemene beginsel namelijk door.³²⁹ Dit wordt ook bevestigd in het Swedex-arrest.³³⁰

Bovendien is in Mangold is besloten dat nationale rechters, in geschillen tussen particulieren en waarbij een beroep wordt gedaan op richtlijnen die algemene beginselen van de Europese Unie preciseren, die beginselen en de daarbij betrokken richtlijnbevestigingen moeten realiseren door de nationale regels die in strijd zijn met deze richtlijnbevestigingen buiten beschouwing te laten. Wat wel belangrijk is, is dat in het Swedex-arrest is herhaald dat richtlijnen geen horizontale directe werking hebben, wat inhoudt dat ze geen verplichtingen kunnen opleggen aan particulieren en dus niet tegen een particulier kan worden ingeroepen. Omdat dit enigszins controversieel klinkt ten aanzien wat daarvoor is beschreven, is een nadere uitleg op zijn plaats. Het Hof heeft besloten dat niet de richtlijn zelf, maar een algemeen beginsel van Europees recht horizontale doorwerking heeft. Daarnaast gaat het in deze zaak om het buiten toepassing laten van nationaal recht, als consequentie van de voorrang van het Europees recht. Zoals eerder aangegeven, is het Swedex-arrest een voortzetting op het Mangold-arrest, het arrest bevestigt dat nationale rechters voortaan, wanneer het gaat om algemene beginselen van Europees recht, die beginselen toepassen.

³²⁶ HvJ 19 januari 2010, ECLI:EU:C:2010:21, r.o. 33 en Berkhout, BJU 2010, p. 77.

³²⁷ HvJ 19 januari 2010, ECLI:EU:C:2010:21, r.o. 43.

³²⁸ HvJ 19 januari 2010, ECLI:EU:C:2010:21, r.o. 56.

³²⁹ HvJ 22 november 2005, ECLI:EU:C:2005:709, r.o. 77.

³³⁰ HvJ 19 januari 2010, ECLI:EU:C:2010:21, r.o. 51.

Een andere zaak dat vergelijkbaar is met het Swedex-arrest, is de zaak Hütter.³³¹ In deze zaak ging het om twee jongens die beide stageliepen bij de Technische Universiteit Graz (TUG). Na afloop van de stage kregen zij beide een contract voor drie maanden aangeboden. In Oostenrijk gold een wet, waarin stond dat werkervaring die werd opgedaan voor het bereiken van de leeftijd van 18 jaar niet in aanmerking kwam voor het berekenen van het loon. Hütter was nog maar net 18 jaar geworden, waardoor hij op grond van deze wet werkervaring had opgedaan van 6,5 maanden. Dit in tegenstelling tot de andere jongen, die ten tijde van de stage bijna 20 jaar was, waardoor hij 28,5 maanden werkervaring had opgedaan op grond van deze berekening. Dit terwijl beide jongens ongeveer dezelfde werkervaring hadden opgedaan. Hütter was van mening dat hier sprake was van discriminatie,³³² de TUG betwistte dat er sprake was van discriminatie.³³³ De TUG voerde aan dat de regel mogelijk leidde tot indirecte discriminatie, maar dat dit gerechtvaardigd kon worden met een legitiem doel, dat passend en noodzakelijk is.³³⁴ Dit legitieme doel was tweeledig en hield het volgende in: “(1) om personen die algemeen secundair onderwijs hebben gevolgd niet te benadelen in vergelijking met personen met een beroepsopleiding; en (2) om te vermijden dat stageplaatsen duurder worden en om de integratie van jonge personen op de arbeidsmarkt te bevorderen die dat type van opleiding hebben gevolgd.”³³⁵ Het Hof oordeelde in deze zaak dat sprake was van directe discriminatie op grond van leeftijd, omdat de maatregel voorziet in een ongunstigere behandeling van personen die jonger starten met werken.³³⁶ Het Hof vond de maatregel niet passend, omdat het niet alleen ervaring beloont, maar daarnaast ook behandeling op leeftijd wordt gehanteerd. Leeftijdsgenoten kunnen door het leeftijdscriterium ongelijk worden behandeld.³³⁷ Hierdoor is het Hof van oordeel dat de maatregel niet passend en noodzakelijk is om het legitieme doel te bereiken, er was geen objectieve rechtvaardiging geformuleerd.

Onderscheid binnen de politie

Op 27 november 1980 vorderde mevrouw Johnston bij het Industrial Tribunal in het Verenigd Koninkrijk vaststelling van haar rechten tegen de Chief Constable van de Royal Ulster Constabulary (hierna: RUC). De Chief Constable is bevoegd hulppolitieagenten aan te stellen bij de RUC reserve in Noord-Ierland en op voltijdse posten in de fulltime reserve op basis van verlengbare contracten. De Chief Constable weigerde het contract van mevrouw Johnston als lid van de RUC fulltime Reserve te verlengen en haar de opleiding voor gebruik van vuurwapens te laten volgen. De politie indertijd droeg bij alledaagse praktijken in hun uitoefening geen wapen, dit was voorbehouden aan bijzondere operaties. De Chief Constable was echter van mening dat de mannen, en niet vrouwen, in de RUC en in de RUC Reserve tijdens de normale dienst wel wapens zouden dragen, nadat een groot aantal agenten bij aanslagen om het leven waren gekomen.³³⁸ Hiertegen stelde Johnston beroep in omdat het hare inziens neerkwam op discriminatie op grond van sekse.³³⁹ Het Industrial Tribunal heeft uiteindelijk prejudiciële vragen gesteld aan het Hof:

³³¹ HvJ 18 juni 2009, ECLI:EU:C:2009:381.

³³² HvJ 18 juni 2009, ECLI:EU:C:2009:381, r.o. 18.

³³³ HvJ 18 juni 2009, ECLI:EU:C:2009:381, r.o. 19.

³³⁴ HvJ 18 juni 2009, ECLI:EU:C:2009:381, r.o. 20.

³³⁵ Handboek over het Europese non-discriminatierecht 2010, p. 57.

<https://fra.europa.eu/sites/default/files/fra_uploads/1510-FRA_CASE_LAW_HANDBOOK_NL.pdf>.

³³⁶ HvJ 18 juni 2009, ECLI:EU:C:2009:381, r.o. 38.

³³⁷ HvJ 18 juni 2009, ECLI:EU:C:2009:381, r.o. 48.

³³⁸ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 4.

³³⁹ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 7.

- 1) *Moet richtlijn 76/207 van de Raad van 9 februari 1976 betreffende de tenuitvoerlegging van het beginsel van gelijke behandeling van mannen en vrouwen (PB 1976, L 39, blz. 40; hierna: de richtlijn), in de omstandigheden als omschreven in het door partijen aanvaarde relaas van de feiten, aldus worden uitgelegd, dat een Lid-Staat handelingen van het toepassingsgebied van de richtlijn mag uitsluiten welke ten aanzien van de toegang tot het arbeidsproces discrimineren op grond van geslacht, doch worden verricht om de nationale veiligheid te waarborgen dan wel de openbare veiligheid of de openbare orde te beschermen ?*
- 2) *Moet de richtlijn in de omstandigheden als omschreven in het door partijen aanvaarde relaas van de feiten, aldus worden uitgelegd, dat de voltijdse betrekking als gewapend lid van een korps reservepolitie, of de opleiding in het hanteren en het gebruik van vuurwapens voor een dergelijke betrekking, is te beschouwen als een beroepsactiviteit en, in voorkomend geval, waarvoor vanwege hun aard of de voorwaarden voor de uitoefening ervan, het geslacht een bepalende factor is in de zin van artikel 2, lid 2 ?*
- 3) *Volgens welke beginselen en criteria dienen de Lid-Staten te bepalen of voor a) de „beroepsactiviteiten“ van een gewapend lid van een dergelijk korps en b) de „hiervoor noodzakelijke opleidingen“, vanwege hun aard of de voorwaarden voor de uitoefening ervan, „het geslacht een bepalende factor is“, als bedoeld in artikel 2, lid 2 ?*
- 4) *Kan het beleid van een Chief Constable of Police, die krachtens de wet met de verantwoordelijkheid voor de leiding en het beheer van een politiekorps is belast, dat vrouwelijke leden van dat korps geen vuurwapens mogen dragen, in de omstandigheden als omschreven in het relaas van de feiten in deze zaak, worden beschouwd als een bepaling „betreffende de bescherming van de vrouw“ als bedoeld in artikel 2, lid 3, of als een „bestuursrechtelijke bepaling... gemotiveerd door beschermende bedoelingen“ in de zin van artikel 3, lid 2, sub c, van de richtlijn ?*
- 5) *Indien het antwoord op de vierde vraag bevestigend luidt: volgens welke beginselen en criteria moeten de Lid-Staten bepalen of de „beschermende bedoelingen... gefundeerd zijn“ in de zin van artikel 3, lid 2, sub c ?*
- 6) *Kan verzoekster zich in de omstandigheden van de onderhavige zaak voor de nationale rechterlijke instanties van de Lid-Staten beroepen op het beginsel van gelijke behandeling, als neergelegd in de betrokken bepalingen van de richtlijn ?*
- 7) *Indien het antwoord op de zesde vraag bevestigend luidt:*
 - a) *Moet artikel 224 EEG-Verdrag aldus worden uitgelegd, dat Lid-Staten in geval van ernstige binnenlandse onlusten waardoor de openbare orde wordt verstoord, zich ontheven mogen achten van de ingevolge de richtlijn op hen of op werkgevers binnen hun rechtsgebied rustende verplichtingen ?*
 - b) *Zo ja, kan een particulier zich beroepen op de omstandigheid dat een Lid-Staat geen overleg heeft gepleegd met andere Lid-Staten, teneinde aldus een beroep van eerstbedoelde Lid-Staat op artikel 224 EEG-Verdrag uit te sluiten ?*

De vragen strekken ertoe te vernemen of het verenigbaar is met het gemeenschapsrecht en met richtlijn 76/207, wanneer de nationale rechter door een voorschrift als dat van artikel 53 lid 2 van de Sex Discrimination Order wordt belet zijn rechterlijke controle uit te oefenen. Verder zijn de vragen bedoeld om te kunnen beslissen of en onder welke voorwaarden op grond van de bepalingen van de richtlijn in een situatie zoals in deze zaak waarbij mannen en vrouwen verschillend mogen worden behandeld om redenen in verband met de bescherming van de openbare veiligheid. Bovendien strekt het ertoe te achterhalen of de richtlijn ingeroepen kan

worden tegen een daarmee strijdige nationale bepaling. Als laatste kan de vraag rijzen of een lidstaat zich kan onttrekken aan de verplichtingen krachtens de richtlijn.³⁴⁰ De Britse regering is van mening dat de bedoelde maatregelen niet binnen de werkingssfeer van de richtlijn vallen.³⁴¹ Bovendien is de Britse regering van mening dat lidstaten over een discretionaire bevoegdheid beschikken om te beslissen of wegen de eisen van nationale veiligheid, openbare veiligheid of openbare orde de voorwaarden waaronder een bepaalde beroepsactiviteit bij de politie wordt uitgeoefend, van dien aard zijn dat zij zich ertegen verzetten dat deze activiteiten worden uitgeoefend door een met een vuurwapen uitgeruste vrouwelijke politieagente.³⁴² Zo meende de Chief Constable dat dit beleid zo is vastgesteld omdat hij meende dat vrouwelijke agenten die gewapend waren, vaker het doelwit zouden kunnen zijn van moordaanslagen en dat vuurwapens in handen van de aanvallers zouden kunnen vallen. Ook zou het dragen van wapens door vrouwen niet van harte worden toegeroepen door het publiek, omdat dit in strijd zou zijn met het ideaal van een ongewapend politiekorps en de vrouwelijke agenten minder doeltreffend zouden zijn in politiewerk op sociaal vlak met gezinnen en kinderen. Deze redenen hielden verband met de bijzondere voorwaarden waaronder de politie moet opereren in Noord-Ierland, gelet op de eisen in verband met de bescherming van de openbare veiligheid in geval van ernstige binnenlandse onlusten.³⁴³

Het Hof oordeelt dat in bepaalde omstandigheden geslacht een bepalend factor kan zijn voor de uitoefening van activiteiten. In zo'n geval mag een lidstaat die activiteiten voorbehouden aan enkel mannen.³⁴⁴ Hierbij moet rekening worden gehouden met het evenredigheidsbeginsel en mag dit niet verder gaan dan passend en noodzakelijk is ter verwezenlijking van het nagestreefde doel, er moet een evenwicht tot stand worden gebracht tussen het beginsel van gelijke behandeling en de eisen van openbare veiligheid.³⁴⁵ Het komt de nationale rechter toe te beoordelen of er sprake is van een rechtvaardiging voor zo'n soort besluit, hierbij dient het evenredigheidsbeginsel te worden eerbiedigt. Ten aanzien van de tweede en derde vraag oordeelt het Hof dat een lidstaat de eisen van de bescherming van de openbare veiligheid, in geval van veelvuldige aanslagen, in aanmerking mag nemen door algemene politietaken voor te behouden aan met vuurwapens uitgeruste mannen.³⁴⁶ Met de vierde en de vijfde vraag wordt verzocht de begrippen "bescherming van de vrouw" en "beschermende bedoelingen" uit artikel 2 lid 3 respectievelijk artikel 3 lid 2 sub c van de richtlijn te worden uitgelegd. De Britse regering betoogt dat het beleid ten aanzien van vrouwen in de RUC full-time Reserve bedoeld is hen te beschermen door te voorkomen dat ze geen doelwit worden van moordaanslagen. Hier is de Commissie het mee eens, in de uitzonderlijke situatie die toentertijd gaande was, is het gerechtvaardigd dat vrouwen geen vuurwapen mogen dragen in het kader van de bescherming van deze vrouwen. Het Hof stelt dat deze artikelen zich richten tot zwangerschap en moederschap. Deze bepaling staat niet toe dat vrouwen tot een bepaalde betrekking worden ontzegd op grond van de publieke opinie dat vrouwen meer beschermd dienen te worden dan mannen.³⁴⁷

Het Hof geeft aan dat de gevaren waaraan vrouwen worden blootgesteld tijdens de uitoefening van hun politiewerkzaamheden niet verschillen van de gevaren waar mannen in een gelijke

³⁴⁰ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 12.

³⁴¹ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 24.

³⁴² HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 31.

³⁴³ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 35.

³⁴⁴ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 37.

³⁴⁵ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 38.

³⁴⁶ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 40.

³⁴⁷ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 44.

functie worden blootgesteld. Het argument dat zij worden beschermd om redenen van openbare veiligheid, kan niet worden tot de verschillende behandeling op grond van artikel 2 lid 3 met het oog op bescherming van de vrouw.³⁴⁸ De zesde vraag richt zich erop of particulieren zich kunnen beroepen op de richtlijn. De Britse regering is van mening dat artikel 2 lid 1 van de richtlijn geen rechtstreekse werking heeft omdat de Chief Constable in onderhavige zaak betrokken is als werkgever.³⁴⁹ Het Hof merkt op dat artikel 2 lid 1 in nationaal recht is omgezet, waardoor het zonder bezwaar wordt verklaard.³⁵⁰ De afwijking van het beginsel van gelijke behandeling wordt toegestaan, maar het verleent slechts een optie. De nationale rechter moet beoordelen of van deze optie gebruik is gemaakt en dient de inhoud van die bepalingen te beoordelen.³⁵¹ Het is in onderhavige zaak aan het Industrial Tribunal om de bepalingen van de Sex Discrimination Order uit te leggen in het licht van de richtlijn. Ook haalt het Hof aan dat de verplichtingen uit de richtlijn om maatregelen te treffen die geschikt zijn om de nakoming van die verplichting te verkeren, bij alle overheidsgezag beklede instanties alsmede rechterlijke instanties, zoals zij ook hebben gesteld in Von Colson/Kamann en Harz.³⁵² Vraag zeven behoeft geen antwoord, gelet op voorgaande antwoorden. Particulieren kunnen zich op de richtlijn beroepen, daarnaast oordeelt het Hof dat de Chief Constable niet optreedt als particulier en hij niet mag profiteren van het feit dat de Staat zich niet aan gemeenschapsrecht houdt.³⁵³

Twee vergelijkbare zaken met een andere uitkomst

Na de Johnston-zaak, kwamen in de jaren '70 van de vorige eeuw twee zaken bij het Hof die op sommige vlakken erg op elkaar lijken. Het eerste arrest is Sirdar / The Army Board³⁵⁴ en het tweede arrest is Kreil / Duitsland.³⁵⁵ Net zoals in de Johnston-zaak ging het in deze zaken om ongelijke behandeling tussen mannen en vrouwen op basis van de openbare veiligheid.

Als eerste Sirdar / The Army Board. Mevrouw Sirdar had gesolliciteerd op een functie als kok bij de Royal Marines maar werd uitgesloten van toegang tot dit korps omdat zij een vrouw was. De reden van de uitsluiting was dat de toelating van vrouwen niet verenigbaar zou zijn met het verzekeren van de interoperabiliteit van de eenheden van de Royal Marines.³⁵⁶ Zo zegt de leiding van de Royal Marines dat elke marinier in tijden van crisis of personeelstekort in staat moet zijn te allen tijde in een commando-eenheid te functioneren. Deze interoperabiliteit is bij het aanstellen van vrouwen niet gewaarborgd.³⁵⁷ Sirdar besloot na haar ontslag een vordering in te stellen bij het Industrial Tribunal, welke aan het Hof de volgende prejudiciële vragen stelde:

„1) Vallen beleidsbeslissingen die een lidstaat in vredetijd en/of ter voorbereiding op oorlog neemt ten aanzien van de toegang tot de arbeid, de beroepsopleiding, de arbeidsvoorwaarden of het inzetten van zijn strijdkrachten, indien deze beslissingen worden genomen in het belang van de gevechtskracht, buiten de werkingssfeer van het

³⁴⁸ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 45.

³⁴⁹ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 49.

³⁵⁰ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 51.

³⁵¹ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 52.

³⁵² HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 53.

³⁵³ HvJ 25 mei 1986, ECLI:EU:C:1986:206, r.o. 54.

³⁵⁴ HvJ 26 oktober 1999, ECLI:EU:C:1999:523.

³⁵⁵ HvJ 10 januari 2000, ECLI:EU:C:2000:2.

³⁵⁶ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 7.

³⁵⁷ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 7.

EG-Verdrag en/of de afgeleide wetgeving, in het bijzonder richtlijn 76/207/EEG van de Raad?

2) Vallen beslissingen die een lidstaat ter voorbereiding op oorlog en in vredetijd neemt met betrekking tot de aanwerving, de opleiding en het inzetten van soldaten in marinecommando-eenheden van zijn strijdkrachten, die zijn bedoeld voor directe confrontatie met vijandelijke strijdkrachten in geval van oorlog, buiten de werkingssfeer van het EG-Verdrag of de afgeleide wetgeving, indien die beslissingen worden genomen ter verzekering van de gevechtskracht van zulke eenheden?

3) Moet artikel 224 EG-Verdrag aldus worden uitgelegd, dat het de lidstaten toestaat, discriminatie op grond van geslacht in verband met de toegang tot de arbeid, de beroepsopleiding, de arbeidsvoorwaarden, de ontslagvoorwaarden daaronder begrepen, binnen de strijdkrachten in vredetijd en/of ter voorbereiding op oorlog ter verzekering van de gevechtskracht uit te sluiten van de werkingssfeer van richtlijn 76/207/EEG?

4) Kan ingevolge artikel 224 het beleid van een lidstaat om in vredetijd en/of ter voorbereiding op oorlog alle vrouwen uit te sluiten van de dienst als interoperabele marinier, van de werkingssfeer van richtlijn 76/207/EEG worden uitgesloten? Zo ja, volgens welke maatstaven of criteria moet worden beslist, of een dergelijk beleid aldus terecht op grond van artikel 224 is uitgesloten van de werkingssfeer van richtlijn 76/207/EEG?

5) Kan het beleid van een lidstaat om in vredetijd en/of ter voorbereiding op oorlog alle vrouwen uit te sluiten van de dienst als interoperabele marinier, worden gerechtvaardigd op grond van artikel 2, lid 2, van richtlijn 76/207/EEG?

6) Zo ja, volgens welke criteria moet de nationale rechter dan beslissen, of de toepassing van dit beleid gerechtvaardigd is ? "

De eerste twee vragen zijn bedoeld om te achterhalen of de beroepsopleiding en arbeidsvoorwaarden bij strijdkrachten, met name ten aanzien van commando-eenheden van de marine, buiten de werkingssfeer van het gemeenschapsrecht vallen.³⁵⁸ Volgens Sirdar sluit geen bepaling de strijdkrachten uitdrukkelijk uit van de werkingssfeer van het Verdrag. In tegenstelling tot de overtuiging van Sirdar is de regering van het Verenigd Koninkrijk van mening dat besluiten inzake en het beheer van de strijdkrachten, met name besluiten ter verzekering van de gevechtskracht voor bijvoorbeeld de voorbereiding op een oorlog, buiten de werkingssfeer van het Verdrag vallen.³⁵⁹ Het Hof oordeelt uiteindelijk dat beslissingen op het gebied van toegang tot de arbeid, beroepsopleiding en arbeidsvoorwaarden bij strijdkrachten niet in algemene zin aan de werkingssfeer van het gemeenschapsrecht zijn onttrokken.³⁶⁰ Vervolgens neemt het Hof een besluit over de vijfde en zesde vraag. Hierbij wordt gekeken onder welke voorwaarden de uitsluiting van vrouwen in commando-eenheden binnen de Royal Marines gerechtvaardigd is.³⁶¹ Volgens Sirdar, de commissie en andere lidstaten die opmerkingen hebben ingediend, moet hierbij getoetst worden aan de criteria aangehaald in het arrest Johnston en het evenredigheidsbeginsel. De regering van het Verenigd Koninkrijk is van mening dat het rechterlijk toezicht op dit gebied beperkt is en alleen betrekking kan hebben op de vraag of de nationale autoriteiten het betrokken beleid in redelijkheid passend en noodzakelijk kon achten.³⁶²

³⁵⁸ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 11.

³⁵⁹ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 12.

³⁶⁰ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 20.

³⁶¹ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 21.

³⁶² HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 22.

Het Hof heeft erkend dat het geslacht een bepalende factor kan zijn voor bepaalde beroepen, zoals een bewaker in een gevangenis,³⁶³ of voor bepaalde politietaken, zoals is bepaald in Johnston. Een lidstaat mag zulke activiteiten voorbehouden aan mannen of aan vrouwen, naar mate van het geval. Voorwaarde hierbij is wel dat de lidstaten voorbehouden zijn zulke activiteiten op gezette tijden te onderzoeken om na te gaan of het gerechtvaardigd is de afwijking van de algemene regeling te handhaven.³⁶⁴ Bij zo'n afwijking moet het evenredigheidsbeginsel in acht worden genomen. Op grond van dit beginsel mogen afwijkingen niet verder gaan dan het nagestreefde doel en is het belangrijk dat een evenwicht tot stand wordt gebracht tussen het beginsel van gelijke behandeling en de eisen van openbare veiligheid die betrokken zijn bij de voorwaarden voor het uitoefenen van bepaalde beroepsactiviteiten.³⁶⁵ De lidstaten beschikken over een zekere *margin of appreciation*. Hierom moet ook worden nagegaan of de opgestelde maatregelen, zoals de weigering van vrouwen, in het kader van de *margin of appreciation*, werkelijk gericht zijn om de openbare veiligheid te verzekeren en of deze passend en noodzakelijk zijn om dat doel te verwezenlijken. De derde en de vierde vraag behoeven geen beantwoording, doordat het antwoord op de vijfde en zesde vraag al volledig is geweest.³⁶⁶ De organisatie van de Royal Marines is wezenlijk anders dan die van andere onderdelen van het Britse leger, het is namelijk een kleine strijdmacht die in de voorste linies moet opereren. Dit houdt in dat ook koks in de voorste linie moeten dienen. Hier wordt geen uitzondering op gemaakt. Mede gelet hierop en de regel van interoperabiliteit valt het te rechtvaardigen dat de Royal Marines uitsluitend mannen aannemen, zo oordeelt het Hof.³⁶⁷

In Kreil / Duitsland ging het om mevrouw Kreil, zij solliciteerde voor een baan als elektrisch ingenieur bij de gewapende strijdkrachten. Zij kreeg de baan niet, omdat vrouwen alleen toegang hadden tot de geneeskundige dienst en de militaire muziekkorpsen. Hierop stelde zij beroep in bij het Verwaltungsgericht Hannover en voerde aan dat de afwijzing in strijd was met het gemeenschapsrecht. Ook in deze zaak werden er aan het Hof een prejudiciële vraag gesteld, in dit geval door het Verwaltungsgericht Hannover:

Wordt richtlijn 76/207/EEG van de Raad van 9 februari 1976 — in het bijzonder artikel 2, lid 2, daarvan — geschonden door § 1, lid 2, derde volzin, Soldatengesetz in de versie van 15 december 1995 (BGBl. I, blz. 1737), laatstelijk gewijzigd bij wet van 4 december 1997 (BGBl. I, blz. 2846), en § 3a Soldatenlaufbahnverordnung in de op 28 januari 1998 gepubliceerde versie (BGBl. I, blz. 326), volgens welke vrouwen die vrijwillig dienst hebben genomen, enkel bij de geneeskundige dienst en de militaire muziekkorpsen kunnen worden aangesteld en steeds van de gewapende dienst zijn uitgesloten?

Deze vraag is bedoeld om duidelijk te krijgen of de richtlijn in de weg staat aan de toepassing van nationale bepalingen die vrouwen uitsluiten van gewapende dienst in de krijgsmacht en hen enkel toegang verlenen tot de geneeskundige dienst en de militaire muziekkorpsen.³⁶⁸ Kreil beoogt dat dit een directe discriminatie jegens vrouwen is en acht dit ontoelaatbaar.³⁶⁹ De Duitse

³⁶³ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 23, zie voor verdere uitleg HvJ 30 juni 1988, ECLI:EU:C:1988:352.

³⁶⁴ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 25.

³⁶⁵ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 26.

³⁶⁶ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 33.

³⁶⁷ HvJ 26 oktober 1999, ECLI:EU:C:1999:523, r.o. 31.

³⁶⁸ HvJ 10 januari 2000, ECLI:EU:C:2000:2, r.o. 10.

³⁶⁹ HvJ 10 januari 2000, ECLI:EU:C:2000:2, r.o. 11.

regering is daarentegen van mening dat het gemeenschapsrecht niet in de weg staat van dit beleid, omdat het gemeenschapsrecht niet geldt voor kwesties van landverdediging. Bovendien, mocht de richtlijn wel van toepassing kunnen zijn, zijn de nationale bepalingen waarin vrouwen worden beperkt, gerechtvaardigd op grond van artikel 2 lid 2 en 3 van de richtlijn.³⁷⁰ Zoals het Hof al eerder heeft besloten, in onder andere het Sirdar-arrest, dat afwijkingen van de richtlijn toelaatbaar wanneer de openbare veiligheid op het spel kan staan, deze artikelen hebben nauwkeurig omschreven uitzonderingsgevallen. Het begrip openbare veiligheid dekt zowel de interne veiligheid, zoals in de zaak Johnson, als de externe veiligheid, zoals in de zaak Sirdar.³⁷¹ Sommige uitzonderingen hebben echter betrekking op de vier vrijheden en niet op sociale bepalingen zoals het gelijkheidsbeginsel. Dit beginsel heeft een algemene strekking en daardoor geldt de richtlijn ook voor publiekrechtelijke dienstverhoudingen.³⁷² Derhalve is de richtlijn van toepassing op deze zaak.³⁷³ Verder merkt het Hof in zijn uitspraak de algemene overwegingen van het Sirdar-arrest op met betrekking tot de toepasselijkheid van de richtlijn inzake gelijke behandeling en de uitzondering van artikel 2 lid 2,³⁷⁴ alsmede de margin of appreciation die de lidstaten toekomt.³⁷⁵ Net zoals in Sirdar toetst het Hof of in de omstandigheden van het geval de maatregelen die door de nationale autoriteiten in de uitoefening van hun toegekende beoordelingsmarge zijn vastgesteld, werkelijk gericht zijn op het doel de openbare veiligheid te verzekeren, en of zij passend en noodzakelijk zijn ter verwezenlijking van dat doel.³⁷⁶ De weigering van de vrouw berust op een bepaling die vrouwen volledig uitsluiten van de gewapende dienst, zo'n bepaling die geldt voor nagenoeg alle functies is wegens de omvang niet te beschouwen als een afwijking die wordt gerechtvaardigd.³⁷⁷ Ondanks dat het Hof rekening houdt met de beoordelingsmarge die lidstaten toekomt, komen zij toch tot een schending van het evenredigheidsbeginsel door het uitgangspunt dat alle gewapende onderdelen uitsluiting uit mannen moet bestaan. Hoewel de Duitse regering zich beroept op artikel 2 lid 3 van de richtlijn, stelt het Hof dat deze bepaling niet toe laat dat vrouwen de toegang wordt ontzegd tot bepaalde functies puur op grond van de bepaling dat vrouwen meer beschermd dienen te worden tegen gevaren dan mannen.³⁷⁸ Het Hof beantwoordt de prejudiciële vraag dus bevestigend; de richtlijn staat aan de toepassing van nationale bepalingen, waarbij vrouwen worden uitgesloten van functies in de gewapende dienst, in de weg.

Hoewel deze twee zaken in hetzelfde jaar door het Hof werden behandeld, beide zaken over een vergelijkbare situatie gaan, heeft het Hof toch anders geoordeeld. Het Sirdar-arrest geeft een algemene regel, welke vervolgens niet van toepassing is op het arrest en een uitzondering wel van toepassing is. In het Kreil-arrest is de algemene regel wel van toepassing. De twee genoemde zaken hebben betrekking op het gelijkheidsbeginsel binnen militaire eenheden van een staat, waaruit blijkt dat het Europeesrechtelijke gelijkheidsbeginsel niet in weg staat om vrouwen de toegang tot bepaalde onderdelen van de militaire eenheid te ontfeggen. Hoewel er binnen dit veld voorbijgegaan mag worden aan het gelijkheidsbeginsel, worden er wel eisen gesteld aan de afwijking van het gelijkheidsbeginsel. Er kan bijvoorbeeld niet ten aanzien van elke functie worden beroepen door de staat op een uitzondering om vrouwen af te wijzen, dit is

³⁷⁰ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 12.

³⁷¹ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 17.

³⁷² HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 18.

³⁷³ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 19.

³⁷⁴ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 20.

³⁷⁵ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 24.

³⁷⁶ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 25.

³⁷⁷ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 26 en 27.

³⁷⁸ HvJ 10 januari 2000, ECLI:EU:C:2000:2. r.o. 30.

alleen te rechtvaardigen ten aanzien van specifieke functies, zoals in het geval van Sirdar. Wanneer het betrekking heeft op een meer algemene bepaling, zoals bij Kreil, dan vindt het Hof geen rechtvaardiging op grond van het evenredigheidsbeginsel.

Gelet op zaken Johnston, Sirdar en Kreil kan er worden geconcludeerd dat een afwijking van hetgeen in de richtlijn bepaald is, toelaatbaar is in het kader van openbare veiligheid, wel moet hierbij worden gezegd dat het hierbij niet kan gaan om de gehele strijdkrachten. In geval van dreigingen die de openbare veiligheid zouden kunnen aantasten, wordt het als gerechtvaardigd gezien dat een lidstaat de keuze maakt om vrouwen uit te sluiten van bepaalde functies, om zo deze veiligheid te kunnen waarborgen. In alle andere gevallen, is dit niet gerechtvaardigd.

Gerechtvaardigd onderscheid

Bovenstaande arresten gingen in op situaties waarin een vrouw ongelijk werd behandeld ten opzichte van een man. Het komt de afgelopen jaren dan ook steeds vaker voor dat bedrijven speciale maatregelen hanteren, zoals een voorkeursbeleid ten aanzien van vrouwen. Het Hof heeft meerdere zaken behandeld die in gaan op zo'n speciale maatregel; namelijk de zaak Kalanke, de zaak Marschall en de zaak Abrahamsson. Samen hebben deze zaken bepaald hoe ver speciale maatregelen mogen gaan om nadelen uit het verleden te corrigeren.³⁷⁹ In Kalanke ging het erom dat regelgeving op regionaal niveau voorkeur toekende aan vrouwelijke kandidaten binnen een sollicitatieproces.³⁸⁰ De personeelsraad van de park- en plantsoendienst van de stad Bremen weigerde de heer Kalanke promotie te geven. Hoewel een bemiddelingspoging in een aanbeveling resulteerde van de heer Kalanke. Uiteindelijk besloot de verzoeningscommissie, die door de personeelsraad werd betrokken in de zaak, dat zowel de heer Kalanke als de vrouwelijke kandidaat gelijke kwalificaties hadden maar dat op grond van de Landesgleichstellungsgesetz (= Wet gelijke kansen) de vrouw voorrang moest krijgen.³⁸¹ In geval van een mannelijk en een vrouwelijke sollicitant die even gekwalificeerd zijn, zou dus een vrouw de voorkeur krijgen indien vrouwelijke werknemers binnen die sector ondervertegenwoordigd waren. Deze zogenaamde quota-regeling zou ertoe kunnen bijdragen dat in de toekomst het een gewoon beeld wordt dat vrouwen ook bepaalde functies bekleden.³⁸² Het Hof oordeelde dat een nationale regeling welke vrouwen met dezelfde kwalificaties als hun mannelijke medekandidaten in sectoren waarin zij ondervertegenwoordigd zijn bij promotie automatisch voorrang genieten, discriminatie op grond van geslacht inhoudt.³⁸³ Het Hof besloot dat voorkeursbeleid mag, maar dat dit strikt dient te worden geïnterpreteerd.³⁸⁴ Artikel 2 lid 1 en 2 van de richtlijn staan in casu in de weg van deze nationale regeling, in de gevallen dat bij gelijke kwalificaties van voor in promotie in aanmerking komende kandidaten van verschillend geslacht, automatisch voorrang wordt verleend aan vrouwen in geval van ondervertegenwoordiging.³⁸⁵ Hiermee wordt bedoeld dat positieve acties ten aanzien van vrouwen toelaatbaar zijn, maar dat deze grenzen niet mogen worden overschreden omdat er anders wel sprake zal zijn van discriminatie. In het geval van Kalanke werd deze speciale maatregel toegestaan omdat het de gelijkheid op de werkvloer nastreefde.

³⁷⁹ Handboek over het Europese non-discriminatierecht 2010, p. 41.

³⁸⁰ HvJ 17 oktober 1995, ECLI:EU:C:1995:322, r.o. 2.

³⁸¹ HvJ 17 oktober 1995, ECLI:EU:C:1995:322, r.o. 5.

³⁸² HvJ 17 oktober 1995, ECLI:EU:C:1995:322, r.o. 14.

³⁸³ HvJ 17 oktober 1995, ECLI:EU:C:1995:322, r.o. 16.

³⁸⁴ HvJ 17 oktober 1995, ECLI:EU:C:1995:322, r.o. 21.

³⁸⁵ HvJ 17 oktober 1995, ECLI:EU:C:1995:322, r.o. 24.

Twee jaar later oordeelde het Hof in de zaak Marschall, hierin werd een gelijkaardige kwestie als in Kalanke voorgelegd aan het Hof. Het verschil tussen deze twee zaken was wel dat in Marschall niet automatisch voorrang werd verleend aan vrouwen, terwijl dat in Kalanke wel het geval was. In Marschall ging het erom dat gelijkwaardig gekwalificeerde vrouwen prioriteit moesten krijgen, ‘voor zover met de persoon van een mannelijke kandidaat verband houdende redenen de balans niet in diens voordeel doen doorslaan.’³⁸⁶ De heer Marschall was namelijk geweigerd voor de functie van een hoger ambt aan een scholengemeenschap ten voordele van een vrouwelijke kandidaat. Hij spande een zaak aan tegen de deelstaat Land Nordrhein-Westfalen. De deelstaat gaf aan dat de voorrangsregel en aanvullende bevorderingscriterium, namelijk de hoedanigheid van vrouw, dat de ongelijke situatie beoogt te neutraliseren waarin vrouwelijke kandidaten zich ten opzichte van hun mannelijke concurrenten bevinden, bij gelijke kwalificaties zou de werkgever namelijk geneigd zijn eerder een man dan een vrouw te bevoordelen.³⁸⁷ Bovendien merkt de deelstaat op dat de voorrang aan vrouwen een tegenwicht moet vormen tegen de traditionele bevorderingscriteria, zonder deze echter opzij te zetten.³⁸⁸ Als laatste kaart de deelstaat aan dat de bestreden bepaling vrouwen geen absolute en onvoorwaardelijke voorrang garandeert.³⁸⁹ Zoals in Kalanke is bepaald dat een dergelijke bepaling waarbij vrouwen automatisch voorrang krijgen discriminatie inhoudt, wordt dit in Marschall genuanceerd. Het ging er namelijk bij Marschall om dat de bepaling een clauseule bevat volgens welke vrouwen geen voorrang dienen te krijgen, indien de balans in het voordeel van een man dient door te slaan.³⁹⁰ Het Hof oordeelt uiteindelijk dat anders dan in het Kalanke-arrest aan de orde was, een nationale regeling de grenzen van discriminatie niet wordt overschreden indien er een objectieve beoordeling plaatsvindt, welke rekening houdt met de kwalificaties van sollicitanten en de toegekende voorrang aan vrouwen buiten toepassing laat, wanneer één of meer criteria in het voordeel van de mannelijke kandidaat uitslaat.³⁹¹

Zoals hierboven eerder is genoemd, is ook hier weer van belang dat er objectief wordt gekeken naar de situatie. Als laatste wordt er gekeken naar de zaak Abrahamsson en Anderson, in deze zaak oordeelde het Hof anders dan in voorgaande zaken. In totaal hadden acht kandidaten gesolliciteerd, waaronder mevrouw Abrahamsson en de heer Anderson op een functie aan de universiteit van Göteborg. Bij het sollicitatieproces werd rekening gehouden met de wetenschappelijke kwalificaties van de kandidaten en een bepaling uit een verordening die de gelijke behandeling in het beroepsleven wil bevorderen. De heer Anderson kwam puur op basis van wetenschappelijke kwalificaties op de eerste plaats, maar uiteindelijk besloot de benoemingscommissie dat een andere vrouw de functie diende te krijgen.³⁹² Nadat deze vrouw haar sollicitatie had ingetrokken, werd er opnieuw gekeken naar de kandidaten. Het ging toen tussen de heer Anderson en mevrouw Fogelqvist. Hoewel het verschil tussen deze twee aanzienlijk achtte, werd uiteindelijk mevrouw Fogelqvist gekozen op basis van de verordening.³⁹³ Daarop stelde de heer Anderson en mevrouw Abrahamsson beroep in. De heer Anderson beoogde dat het besluit in strijd was met het gelijkheidsbeginsel, terwijl mevrouw Abrahamsson betoogde dat de kandidaten niet gelijkelijk worden beoordeeld.³⁹⁴ In deze zaak ging het dus om beleid dat bepaalde dat kandidaten van het ondervertegenwoordigde geslacht

³⁸⁶ HvJ 11 november 1997, ECLI:EU:C:1997:533, r.o. 3 en 5.

³⁸⁷ HvJ 11 november 1997, ECLI:EU:C:1997:533, r.o. 4.

³⁸⁸ HvJ 11 november 1997, ECLI:EU:C:1997:533, r.o. 15.

³⁸⁹ HvJ 11 november 1997, ECLI:EU:C:1997:533, r.o. 17.

³⁹⁰ HvJ 11 november 1997, ECLI:EU:C:1997:533, r.o. 24.

³⁹¹ HvJ 11 november 1997, ECLI:EU:C:1997:533, r.o. 33.

³⁹² HvJ 6 juli 2000, ECLI:EU:C:2000:367, r.o. 18.

³⁹³ HvJ 6 juli 2000, ECLI:EU:C:2000:367, r.o. 19.

³⁹⁴ HvJ 6 juli 2000, ECLI:EU:C:2000:367, r.o. 21.

die over voldoende kwalificaties beschikte, dus niet per se gelijke kwalificaties als een andere kandidaat, voor de functie prioriteit moest krijgen boven een kandidaat van het andere geslacht, tenzij het verschil tussen de kwalificaties van kandidaten zo groot is, dat dit in strijd zou zijn met het vereiste van Mangold-objectiviteit.³⁹⁵ Dit achtte het Hof in strijd met het EG-verdrag, omdat dit eigenlijk automatisch voorkeur gaf aan kandidaten van het ondervertegenwoordigde geslacht en hier niet voldoende gewicht wordt gegeven aan kwalificaties.

Het Hof heeft zich in veel arresten uitgesproken over (in)directe discriminatie. Opvallend is dat het Hof het begrip ‘objectieve rechtvaardiging middels een legitiem doel’ gebruikt, maar dit begrip is voor lange tijd erg vaag gebleven. Het Hof stelt zich vaak terughoudend op en kent lidstaten een beoordelingsruimte toe. In het Bilka-arrest heeft het Hof meer duidelijkheid gegeven over deze objectieve rechtvaardigingsgrond. In de jaren daarna hield het Hof zich vast aan deze term en wordt deze nog altijd aangehouden als uitgangspunt.

3.3.3 Buitenlandse jurisprudentie

De tien landen die leidend zijn op het gebied van AI-systemen, zijn China, Amerika, Engeland, Canada, Rusland, Duitsland, Noorwegen, Zweden, Frankrijk en India. Waarbij China op de eerste plaats staat en India op de tiende plaats. Op basis van deze gegevens, is voor dit hoofdstuk voornamelijk gekeken naar jurisprudentie uit deze landen. Er is met name gekeken naar Engelstalige jurisprudentie, dit is gedaan uit praktische overwegingen. De gevonden jurisprudentie wordt in de navolgende paragrafen besproken, hiervoor is een onderscheid gemaakt in de verschillende onderwerpen binnen Artificial Intelligence. Tevens zal per zaak duidelijk worden aangegeven in welk land dit heeft afgespeeld. Dit betreft slechts een indicatieve quickscan.

Zelfsturende voertuigen

Nilsson v. General Motors (3:18-cv-00471)

In december 2017 spande een motorrijder, de heer Nilsson, een rechtszaak aan tegen General Motors in San Francisco. De heer Nilsson reed op de weg toen de heer Salazar in zijn Cruise Automation 2016 Chevrolet Bolt van General Motors reed. De heer Salazar was de auto aan het testen, waarbij hij zijn handen niet aan het stuur had terwijl de automatische modus was ingeschakeld. De auto gaf aan in te voegen op de linkerbaan, maar staakte deze poging en reed vervolgens terug naar de oorspronkelijke baan, de auto veranderde dus nooit van rijstrook, waardoor hij in botsing kwam met de heer Nilsson. De heer Nilsson viel toen van zijn motor af en liep verwondingen op aan zijn nek en schouder. Deze verwondingen waren dermate ernstig, dat de heer Nilsson arbeidsongeschiktheidsverlof moest opnemen. De heer Nilsson besloot vervolgens een rechtszaak aan te spannen, hierin beweerde hij dat General Motors een zorgplicht verschuldigd was om zijn zelfrijdende voertuig te laten werken op een manier waarop het voldoet aan de verkeersregels en voorschriften, maar dat zij deze zorgplicht hebben geschonden doordat het zelfrijdende voertuig zo nalatig reed, dat het in een aangrenzende rijstrook terecht kwam zonder rekening te houden met een passerend voertuig. Uit de beweringen is gebleken dat de heer Nilsson met zijn motor op de middelste rijbaan reed toen de heer Salazar op dat exacte moment ook op die rijbaan reed. Op het moment dat geautomatiseerde functies het voertuig over namen, zou de heer Salazar naar buiten hebben gekeken, waardoor het zelfrijdende voertuig autonoom kon manoeuvreren. De heer Nilsson

³⁹⁵ HvJ 6 juli 2000, ECLI:EU:C:2000:367, par. 3.

reed alvorens het ongeluk gebeurde, achter de heer Salazar. Toen hij zag dat het linker knipperlicht aan stond, reed hij door op dezelfde rijstrook. De auto keerde vervolgens terug naar de oorspronkelijke rijstrook waarin de heer Nilsson zich op dat moment bevond. Volgens General Motors ging het verkeer op dat moment 20 km/u, wat de zelfrijdende auto ook volgde. De motor reed daarentegen 27 km/u volgens General Motors. Uiteindelijk stemde General Motors ermee in de rechtszaak te schikken voor een niet bekendgemaakt bedrag.

Hudson v. Tesla and Oscar Enrique Gonzalez-Bustamante (2018-CA-011812-O)

De heer Hudson spande in 2018 bij de rechtbank in Florida een rechtszaak aan tegen Tesla, omdat zijn Tesla model S in de automatische piloot-modus bevond toen hij met 130 km/uur op een andere gehandicaptenauto inreed en crashte, de auto van de heer Gonzalez-Bustamante. Hij liep hierbij rug- en hersenletsel op. De auto was tevens volledig vernield. De heer Hudson stelde dat Tesla via een misleidende campagne consumenten, waaronder zichzelf, heeft doen geloven dat de automatische piloot passagiers veilig met hoge snelheden kon vervoeren met minimale toezicht. In werkelijkheid werkt het systeem niet zoals Tesla beweert en kan er niet volledig op het stuurautomaat systeem worden vertrouwd. Hoewel Tesla beweert dat het systeem betrouwbaar is, is het systeem niet in staat stilstaande objecten te detecteren, zoals de gehandicaptenauto bij deze botsing. Dit resulteert, volgens de heer Hudson, in een buitengewoon hoog risico op botsingen op hoge snelheid, ernstig letsel en dood zowel aan Tesla's passagiers als aan de andere weg deelnemers. De heer Hudson stelt vordering in ten aanzien van aansprakelijkheid, nalatigheid, schending van impliciete garantie, verkeerde voorstelling van zaken, misleidende reclame en schending van Florida's Deceptive and Unfair Trade Practices Act. Bovendien vordert hij een nalatigheidsclaim in ten aanzien van de heer Gonzalez-Bustamante omdat hij zijn plicht niet zou zijn nagekomen om zijn gehandicaptenauto van de linkerrijbaan te verwijderen voordat Hudson erop botste. De advocaat van Hudson betoogt dat Tesla het stuurautomaat systeem beschrijft als rijkhulpsysteem, wat impliceert dat de bestuurder een actief toezicht moet houden, maar dat Tesla het aan zijn consumenten verkoopt als een autonoom systeem. Tevens wordt Tesla ervan beschuldigd zijn consumenten als proefkonijnen te gebruiken, maar er geen enkele verantwoordelijkheid over wil dragen. Tesla verzocht vervolgens om een sepot. Hierbij voeren zij aan dat de handleiding voor de modellen duidelijk maakt dat de stuurautomaat functie niet in staat is om stilstaande objecten

Sz Hua Huang et al v. Tesla Inc., The State of California, (19CV346663)

Op 23 maart 2018 overlijdt Walter Huang nadat met zijn Tesla tegen een betonnen middenberm aanreed, op het moment dat dit gebeurde stond het Autopilot-systeem van zijn Tesla aan. In april van dat jaar heeft zijn vrouw een rechtszaak aangespannen tegen Tesla. Hierin wordt beweerd dat de Autopilot-functie defect was en dat dit de dood van haar man heeft veroorzaakt. Het navigatiesysteem van Huang's Tesla interpreteerde de rijstrooklijnen op de rijbaan verkeerd, kon de betonnen middenberm niet detecteren en remde de auto niet af, maar versnelde in plaats daarvan de auto tot de middenberm. De aantijgingen omvatten onder andere productaansprakelijkheid, gebrekkig productontwerp, het gebrek aan waarschuwing, schending van de garantie, opzettelijke en nalatige verkeerde voorstelling van zaken geven en het vals adverteren.

Harcourt v. Tesla (HG19019709)

Mevrouw Mallory Harcourt, op dat moment acht maanden zwanger, had haar auto geparkeerd in haar garage. Nadat ze de auto in de parkeerstand had gezet, stapte zij uit en haalde haar tweejarige zoon uit het kinderzitje. Ze liet vervolgens de deur aan de bestuurderskant open, omdat ze de boodschappen wilde pakken nadat ze haar zoon een schone luier had aangetrokken.

Terwijl ze haar kind naar binnen begeleidde, die overigens achter haar liep, keerde haar zoon terug naar de auto en ging op de bestuurdersstoel zitten. Mevrouw Harcourt liep terug om haar zoontje terug te roepen. In de tussentijd had haar zoon zich tussen het gaspedaal en de stoel genesteld, trok aan de versnellingspook en versnelde direct richting mevrouw Harcourt. Hierdoor raakte mevrouw bekneeld tussen de auto en de garagemuur. Uiteindelijk hebben burens haar gered en is zij naar het ziekenhuis gebracht. Door het ongeluk brak mevrouw haar bekken, haar been en door haar gebroken bekken moest zij vervroegd bevallen van haar dochter. De vrouw spant een rechtszaak aan waarin zij klaagt over productaansprakelijkheid, nalatigheid, schending van garantie, schending van de Song-Beverly Consumer Warranty Act, opzettelijk verkeerde voorstelling van zaken en consumentenfraude.

Robotchirurgie

Da Vinci Surgery

Het Da Vinci Surgical System is een robot chirurgisch systeem dat is gemaakt door het Amerikaanse bedrijf Intuitive Surgical. Het gebruik van dit systeem heeft geleid tot duizenden gevallen van letsel, ernstige complicaties en soms zelfs tot de dood. In 2013 zijn meer dan 3000 casussen afgehandeld, door een zogenoemde 'settlement'. Onderstaande zaken zijn zaken die wel door de rechtbank zijn behandeld.

Mracek v. Bryn Mawr Hospital, 363 F. (09-2042 363 App'x 925)

In 2005 onderging de heer Mracek een prostatectomie nadat bij hem prostaatkanker was vastgesteld. Zijn chirurg was voornemens om de da Vinci-robot te gebruiken voor de operatie. Mracek beweerde dat de robot tijdens de operatie niet goed functioneerde en foutmeldingen toonde. Het chirurgisch team probeerde de robot operationeel te maken, maar slaagde daar niet in. Vervolgens is een vertegenwoordiger van Da Vinci komen helpen, maar ook deze persoon kon de robot niet functioneel maken. Voor de rest van de operatie werd andere apparatuur gebruikt dan de da Vinci-robot. Een week na de operatie kreeg Mracek erectiestoornissen en pijn in zijn lies, waaraan hij voor de operatie geen last van had.³⁹⁶ De heer Mracek diende een klacht in tegen het ziekenhuis en Intuitive Surgical, waarin hij klaagt over strikte productaansprakelijkheid, risicoaansprakelijkheid, nalatigheid en schending van de garantie. Uiteindelijk is het ziekenhuis vrijwillig ontslagen, wat inhoudt dat de eiser in de rechtszaak de beëindiging van de rechtszaak heeft bevolen. Intuitive Surgical heeft een motie tot kort geding ingediend bij het United States Court of Appeals for the Third Circuit (Hof van Beroep), welke door de rechtbank is toegewezen. Hiertegen ging Mracek tijdig in beroep.³⁹⁷ Hij betoogt dat het Hof ten onrechte een summiere uitspraak heeft gedaan over zijn claim wegens strikte aansprakelijkheid. De uitspraken van het Hof over de drie andere vorderingen werden niet door Mracek betwist. In Amerika gelden de regels dat een eiser voor een risicoaansprakelijkheidsclaim moet aantonen dat het product defect was en dat het defect het letsel veroorzaakte en het defect bestond op het moment dat het product de controle van de fabrikant verliet. De eiser mag dit zowel met direct als indirect bewijs ondersteunen. De District Court (arrondissementsrechtbank) oordeelde dat Mracek geen bewijs had aangedragen om redelijke, secundaire oorzaken voor het defect van de robot weg te nemen of om aan te tonen dat het defect zijn verwonding veroorzaakte. Hier was Mracek het niet mee eens en gaf aan dat hij zijn eigen getuigenis en de getuigenissen van twee behandelende artsen, waaronder de chirurg die de behandeling uitvoerde. Mracek had geen deskundigenrapport ingediend, wat in zijn optiek ook niet nodig was omdat het vermeende effect duidelijk is en gemakkelijk is te

³⁹⁶ <https://cite.case.law/f-appx/363/925/>, onder I.

³⁹⁷ <https://cite.case.law/f-appx/363/925/>, onder I.

begrijpen voor een jury. Uiteindelijk werd geoordeeld dat het terecht was dat een kort geding is uitgesproken, dit omdat de heer Mracek geen bewijs heeft aangeleverd waaruit blijkt dat de erectiestoornissen en liespijn was veroorzaakt door de vermeende storing van de da Vinci-robot.³⁹⁸

O'Brien v. Intuitive Surgical, Inc. 2011 WL 3040479 (N.D. Ill July 25, 2010)

Eiser Daniel O'Brien spant een rechtszaak aan tegen Intuitive Surgical, hij klaagt over strikte productaansprakelijkheid, onjuiste voorstelling van zaken en risicoaansprakelijkheid. Hij beweert dat tijdens een pancreatectomie en de transplantatie van eilandjescellen de da Vinci-robot defect was, die tijdens de operatie werd gebruikt. Hierdoor liep O'Brien verwondingen op. Intuitive Surgical had aanvankelijk de klacht afgewezen, waarna de rechtbank het toestond dat hij een gewijzigde klacht zou indienen. De rechtbank verwierp uiteindelijk de gewijzigde klacht op grond van de overtuiging dat O'Brien leek te beweren dat hij slachtoffer was van nalatigheid van artsen en dat de rechtbank niet in staat was om een basis te kunnen vinden in hetgeen de eiser had aangedragen voor deze claim. Hierna diende O'Brien een nieuwe klacht in, deze klacht werd weer door Intuitive Surgical afgewezen. Ook de rechtbank verwierp de klacht opnieuw. De rechtbank gaf aan dat hoewel zij de klacht van aanklager begreep, de robot tijdens of vlak voor de operatie faalde, waardoor de chirurg overging op de klassieke manier van opereren. Eiser ging ervanuit dat, omdat er een terugroepactie voor deze robot is geweest, de producent aansprakelijk zou zijn voor de verwondingen. De rechtbank ziet dit echter anders, zij beschouwen de robot als een hulpproduct bij operaties. Toen het product faalde, schakelde de chirurg over op een open operatie. Uit de klacht blijkt verder niet of het falen zich voor of tijdens de operatie heeft voorgedaan. De rechtbank zegt vervolgens dat de enige manier waarop de eiser een claim kan indienen, is als het falen van het apparaat bijvoorbeeld een breuk of enig letsel veroorzaakte, of wanneer het vertraging veroorzaakte, waardoor er verwondingen zijn ontstaan. De claim voor risicoaansprakelijkheid vereist een bewijs van een naaste oorzaak. De claim voor onjuiste voorstelling vereist bewijs dat O'Brien werd beschadigd door zijn afhankelijkheid van Intuitive's vermeende verkeerde voorstelling van zaken met betrekking tot het apparaat, wat alleen kan zijn opgetreden als het apparaat zijn verwondingen in de buurt heeft veroorzaakt; als het apparaat O'Brien's verwondingen niet had veroorzaakt, had hij immers niet kunnen worden beschadigd door te vertrouwen op iets dat Intuïtief erover zei. De rechtbank gaf twee niet-exclusieve voorbeelden van hoe het apparaat de verwondingen van O'Brien zou kunnen hebben veroorzaakt:

- (1) de storing van het apparaat veroorzaakte direct een lichamelijke breuk of een of andere verwonding; en
- (2) de storing en de resulterende omschakeling naar "open chirurgie", wat betekent dat chirurgie zonder hulp van het apparaat, vertraagde voltooiing van de operatie en dat de vertraging op zijn beurt de verwondingen van O'Brien veroorzaakte.

Zonder enige uitleg over hoe de storing zijn verwondingen veroorzaakte, kon O'Brien niet 'aannemelijk' suggereren 'een nabije oorzaak of, daaruit volgt,' een recht op uitstel'.

Gezichtsherkennings tools

State v. Loomis (881 N.W.2d 749 (Wis. 2016))

In 2013 werd de heer Loomis vijf tenlasteleggingen opgelegd in verband met een drive-by-schietpartij. Hij ontkende deelname aan de schietpartij, maar gaf toe dat hij diezelfde avond met dezelfde auto had gereden. Loomis pleitte schuldig aan twee van de minder zware aanklachten,

³⁹⁸ <https://cite.case.law/f-appx/363/925/>, onder III.

namelijk het proberen om een verkeersagent te ontluchten en een motorvoertuig te besturen zonder de toestemming van diens eigenaar. Ter voorbereiding op de veroordeling heeft een officier van het Wisconsin Department of Corrections een Pre Sentence Investigation (PSI) opgesteld met een COMPAS-beoordeling. Deze COMPAS-beoordeling schat de kans op recidive in op basis van zowel een interview met de dader als informatie uit het strafrechtelijk verleden van deze dader. De methodiek achter deze beoordeling is een bedrijfsgeheim, waardoor alleen de inschattingen van het recidiverisico aan de rechtbank worden gerapporteerd. Tijdens de hoorzitting verwees de rechtbank naar deze beoordeling en veroordeelde Loomis tot zes jaar gevangenisstraf en vijf jaar uitgebreid toezicht. Loomis diende een motie in voor vrijstelling na de veroordeling bij de rechtbank, met het argument dat het vertrouwen van de rechtbank op COMPAS zijn rechten op een eerlijk proces schond. Zo zei hij dat COMPAS-rapporten gegevens verstrekken die alleen relevant zijn voor bepaalde groepen en omdat de methodologie die wordt gebruikt om de rapporten te maken een bedrijfsgeheim is. Loomis beweerde dat het gebruik van de COMPAS-beoordeling door de rechtbank inbreuk maakte op zowel zijn recht op een geïndividualiseerde straf als op zijn recht om precieze informatie. Loomis voerde bovendien op grond van een eerlijk proces aan dat de rechtbank bij de veroordeling ongrondwettelijk rekening hield met het geslacht door te vertrouwen op een risicobeoordeling die rekening hield met het geslacht. Het gerechtshof ontkende de motie na de veroordeling en het Wisconsin Court of Appeals bevestigde het beroep bij het Wisconsin Supreme Court. Ook het Hooggerechtshof van Wisconsin bevestigde dit. De argumenten van de heer Loomis werden verworpen.

Rechter Bradley ontdekte dat het gebruik van geslacht als een factor bij de risicobeoordeling het niet-discriminerende doel diende om nauwkeurigheid te bevorderen en dat Loomis niet voldoende bewijs had geleverd dat de strafrechter daadwerkelijk rekening had gehouden met geslacht. Bovendien, aangezien COMPAS alleen openbaar beschikbare gegevens en gegevens gebruikt die door de beklaagde zijn verstrekt, concludeerde de rechtbank dat Loomis alle informatie die bij het opstellen van het rapport was gebruikt, had kunnen ontkennen of toelichten en daarom de juistheid van de informatie die bij de veroordeling werd gebruikt, had kunnen verifiëren. Met betrekking tot individualisering benadrukte rechter Bradley het belang van geïndividualiseerde straffen en gaf toe dat COMPAS alleen geaggregeerde gegevens verstrekt over het recidiverisico voor groepen die vergelijkbaar zijn met de dader. Maar ze legde uit dat, aangezien het rapport niet de enige basis voor een beslissing zou zijn, veroordeling die een COMPAS-beoordeling beschouwt, nog steeds voldoende geïndividualiseerd zou zijn, omdat rechtbanken de discretie en informatie hebben die nodig is om het waar nodig niet eens te zijn met de beoordeling. Echter, Justice Bradley voegde eraan toe dat rechters voorzichtig moeten zijn bij het gebruik van dergelijke risicobeoordelingen. Om ervoor te zorgen dat rechters risicobeoordelingen op de juiste manier afwegen, heeft de rechtbank zowel voorgeschreven hoe deze beoordelingen aan rechtbanken moeten worden voorgelegd als in hoeverre rechters ze mogen gebruiken. De rechtbank legde uit dat risicoscores niet mogen worden gebruikt "om te bepalen of een dader is opgesloten" of "om de ernst van de straf vast te stellen".

Rechters die gebruik willen maken van die risicobeoordelingen, zullen andere factoren die de gekozen straf ondersteunen ook moeten uitleggen. De PSI's moeten schriftelijke waarschuwingen bevatten;

- ten eerste verhindert het "eigendomsrecht van COMPAS" de openbaarmaking van hoe risicoscores worden berekend;
- ten tweede zijn COMPAS-scores niet in staat om specifieke individuen met een hoog risico te identificeren, omdat deze scores afhankelijk zijn van groepsgegevens;

- ten derde, hoewel COMPAS vertrouwt op een nationale gegevenssteekproef, is er "geen kruisvalidatiestudie voor een populatie in Wisconsin";
- ten vierde hebben onderzoeken "de vraag opgeworpen over de vraag of [COMPAS-scores] onevenredig veel delinquenten met een hoger risico op recidive classificeren";
- en ten vijfde, COMPAS is specifiek ontwikkeld om het Department of Corrections te helpen bij het nemen van beslissingen na de veroordeling.

Bij het geven van deze waarschuwingen heeft de rechtbank duidelijk gemaakt dat zij zowel algemene scepsis over de nauwkeurigheid van de tool als een meer gerichte scepsis wil opwekken met betrekking tot de beoordeling van de risico's van minderheidsovertreders door de tool.

Een rechter was het hiermee eens, maar vroeg zich wel af of de rechters algoritmische risicobeoordelingen zouden kunnen begrijpen. Ze bekritiseerde met name de beslissing van de rechtbank om Northpointe, het bedrijf dat COMPAS heeft ontwikkeld, de mogelijkheid te ontzeggen om een amicus brief in te dienen. Ze zou een uitgebreider verslag van strafrechtbanken nodig hebben gehad over "de sterke en zwakke punten en de relevantie van de geïndividualiseerde straf die wordt uitgesproken door het op bewijzen gebaseerde instrument". Deze verklaring, zo stelde ze, was nodig in het licht van de kritiek die dergelijke beoordelingen hebben gekregen van zowel overheidsfunctionarissen als wetenschappers. De mening van de rechtbank van Loomis suggereert een poging om het huidige enthousiasme voor algoritmische risicobeoordelingen bij veroordelingen te temperen. Maar het is onwaarschijnlijk dat de schriftelijke disclaimer die de rechtbank vereist voor PSI's, inclusief dergelijke beoordelingen, "rechtbanken in staat zal stellen om de nauwkeurigheid van de beoordeling en het juiste gewicht dat aan de risicoscore wordt toegekend, beter te beoordelen." Door de kracht van de kritiek op COMPAS niet te specificeren, het gebrek aan informatie waarover rechters beschikken, en de externe en interne druk om dergelijke beoordelingen te gebruiken, over het hoofd te zien, is het onwaarschijnlijk dat de oplossing van de rechtbank het gewenste gerechtelijke scepticisme creëert.

Janecyk v. International Business Machines (1: 20-cv-00783 - 2020CH00833)

Op 22 januari 2020 dient Tim Janecyk een klacht in tegen International Business Machines uit eigen naam en namens andere in een vergelijkende situaties. Janecyk is een fotograaf die foto's van zichzelf en anderen tijdens gevoelige situaties, zoals politieke bijeenkomsten, uploadde op de fotodeelsite Flickr. Vervolgens verzamelde het moederbedrijf van Flickr deze foto's en gebruikte ze om een database van 99 miljoen afbeeldingen te maken die als referentiebibliotheek gebruikt kon worden om AI-modellen te trainen. Met behulp van deze database heeft IBM een subset van de foto's gencodeerd om het uiterlijk van de mensen op de foto's te beschrijven. IBM bood zijn collectie vervolgens aan aan onderzoekers als een hulpmiddel om vooringenomenheid in gezichtsherkenningsmodellen te verminderen. Janecyk beweert schendingen van meerdere onderafdelingen van het BIPA en eiste \$1.000 tot \$5.000 per overtreding namens alle inwoners van Illinois 'die hun biometrische identificatiegegevens hadden, inclusief scans van gezichtsgeometrie, verzameld, vastgelegd, ontvangen of anderszins verkregen door IBM van foto's in zijn Diversiteit in Faces Dataset.

Vance v. International Business Machines (20-c-577)

Een paar dagen nadat de klacht van Janecyk tegen IBM werd ingediend, diende een ander advocatenkantoor een tweede klacht in tegen IBM. De klacht betreft schendingen van meerdere onderafdelingen van het Biometric Information Privacy Act (BIPA) en bovendien wordt er claims voor onrechtvaardigde verrijking en een voorlopige voorziening neergelegd. Volgens

het BIPA, dat bestaat sinds 2008, mag een bedrijf de biometrische identificatie van een consument (inclusief gezichtsgeometrie) niet verzamelen, tenzij het de consument eerst heeft geïnformeerd over het verzamelen, het doel en het gebruik van de identifier en schriftelijke toestemming heeft gekregen van de klant. Samen met de heer Janecyk daagt Vance IBM voor de rechter. Net zoals Janecyk heeft Vance een Flickr-account waarop hij foto's heeft geüpload. De eisers beweren dat IBM de Flickr-dataset heeft gebruikt om zijn eigen dataset te kunnen creëren, welke zij vervolgens hebben verspreid onder de naam 'DIF Dataset'. Elke afbeelding in de DIF Dataset zou naar verluidt kunnen worden herleid tot het individuele Flickr-account waarnaar het oorspronkelijk was geüpload. Op basis van deze feiten dienden de twee eisers een collectieve klacht in en stellen dat IBM geen openbaar beschikbaar bewaarschema en richtlijnen heeft opgesteld voor het vernietigen van biometrische informatie in strijd met artikel 15 (a) van de Illinois Biometric Information Privacy Act, 740 ILCS § 14/1 en volgende. De klagers beweren namelijk dat IBM er niet in is geslaagd een openbaar beschikbaar bewaar- en vernietigingsschema bij te houden. De rechtbank oordeelt hierover dat dit niet leidt tot concreet letsel, zoals vereist is. Om deze reden vindt de rechtbank dat de eisers niet bevoegd waren om deze claim in te dienen bij de federale rechtbank. IBM heeft zich tegen de klacht verweerd en eiste dat de claim van de eisers werd afgewezen, door onder andere te stellen dat de toepassing van de BIPA in deze zaak in strijd zou zijn met de extraterritorialiteitsdoctrine van de staat Illinois. De betrokken partijen zijn het er uiteindelijk over eens dat de BIPA niet extraterritoriaal kan worden toegepast, waardoor de zaak zich verder richt op de vraag of de gebeurtenissen voornamelijk en grotendeels in de staat Illinois hebben plaatsgevonden. Verder stelt IBM dat het BIPA uitdrukkelijk foto's en biometrische informatie afgeleid van foto's uitsluit, hiertegen in zeggen de eisers dat gedrag vergelijkbaar met het vermeende gedrag van IBM niet wordt uitgesloten door de foto-uitsluitingen van het BIPA. Hierin gaat de rechtbank mee. De vordering van IBM om de vordering van de eisers af te wijzen, wordt daarom afgewezen. Uiteindelijk stelt IBM dat de eisers geen schade hebben geleden, hierop reageerden de klagers dat zij niet hoeven te beweren dat ze financiële schade hebben geleden, dit wordt vervolgens door de rechtbank bevestigd. Uiteindelijk wijst de rechtbank de vordering van IBM toe waarin wordt gesteld dat de eisers geen voorlopige voorziening kunnen instellen. De rest wordt afgewezen.

Mutnick v. Clearview AI, et al., (1:20-cv-00512)

Op dezelfde dag als de hierboven besproken zaak, klaagde Mutnick Clearview AI aan, eveneens de oprichter Hoan Ton-Thot en de directeur Richard Schwartz, op basis van de oprichting van een gezichtsherkenningsdatabase van miljoenen Amerikanen op basis van foto's die van het internet waren gehaald. Clearview zou die database naar verluidt hebben verkocht aan meer dan 600 wetshandhavingsinstanties en de eiser beweert dat de database van Clearview wordt gebruikt door wetshandhavingsinstanties en andere particuliere entiteiten om personen biometrisch te identificeren. Net als Janecyk beweert Mutnick dat de mensen van wie de biometrische informatie in de database is opgenomen, geen kennis hadden van, en niet instemden met, het vermeende gedrag van Clearview. De eiser stelt dat er sprake is van schending van Biometric Information Privacy Act (BIPA), maar daarnaast ook een schending van het eerste, vierde en veertiende amendement. Het belang dat Mutnick behartigt is enerzijds de belangen van de inwoners van Illinois, waarop de BIPA van toepassing is, daarnaast wil het alle burgers die in de Clearview-database staan beschermen. Mutnick eist een schadevergoeding, daarnaast heeft hij ook gevraagd om een voorlopige voorziening, waarmee hij elke verdere verspreiding van de biometrische gegevens wil stoppen.

Burke v. Clearview AI, Inc., (20-cv-00370-BAS-MSB)

In februari 2020 is er in Californië tegen Clearview een klacht ingediend. Burke beweert dezelfde feiten als Mutnick, maar daarnaast beweert Burke ook dat Clearview handelt in strijd met de California Consumer Privacy Act (CCPA). Volgens de klacht gebruikt de gezichtsherkenningsoftware van Clearview de miljarden geschraapte afbeeldingen in zijn database om een soort biometrische informatie te genereren, bekend als een "gezichtsafdruk", om een gezicht te matchen met andere persoonlijk identificeerbare informatie; Vervolgens verkoopt het toegang tot de faceprint-database aan wetshandhavingsinstanties en particuliere bedrijven. In de klacht wordt beschuldigd dat Clearview op onjuiste wijze persoonlijke informatie heeft verzameld zonder de consument hiervan op de hoogte te stellen. Krachtens de CCPA, die op 1 januari 2020 in werking is getreden, moet een bedrijf consumenten informeren "op of vóór het punt van verzameling" over de categorieën persoonlijke informatie - inclusief biometrische informatie - die wordt verzameld en de doeleinden waarvoor deze zal worden gebruikt, en mag geen aanvullende categorieën verzamelen of de reeds verzamelde persoonlijke informatie voor aanvullende doeleinden gebruiken zonder de consument hiervan op de hoogte te stellen.

Dinerstein v. Google, (19-cv-04311)

In 2017 kondigde Google aan dat zij een partnerschap met University of Chicago aangingen. Dit had als doel om tools te ontwikkelen voor machine learning die toekomstige gezondheidsproblemen en nadelige medische gebeurtenissen van patiënten kunnen voorspellen. Om dit doel te bereiken, heeft de universiteit geanonimiseerde dossiers van volwassen patiënten in de periode tussen 1 januari 2010 en 30 juni 2016 met Google gedeeld. Een van de patiënten van dit medisch centrum was de heer Dinerstein. Dinerstein beweerde dat onder de gegevens die zijn gedeeld met Google onder andere ook vrije tekstnotities vielen, welke niet voldoende waren geanonimiseerd. Volgens Dinerstein is dit in strijd met de Healthcare Insurance Portability and Accountability Act van 1996 (HIPAA), die het delen van 'individueel identificeerbare gezondheidsinformatie' beperkt. Volgens Dinerstein ontstond hierdoor het risico om opnieuw te worden geïdentificeerd door Google, doordat Google via een van zijn andere services andere informatie over personen heeft kunnen verkrijgen. Zowel Google als de universiteit hebben zich hiertegen verweerd door te stellen dat Dinerstein geen claim had ingediend waarop herstel kan worden verleend. Uiteindelijk wijst de rechtbank de claim van Dinerstein af, omdat hij niet heeft kunnen aantonen dat hij schade heeft geleden.

Patel v. Facebook, Inc., (3:15-cv-03747-JD)

In deze class action wordt door Facebookgebruikers beweerd dat de gezichtsherkenningstechnologie van Facebook in strijd is met de BIPA, Facebook zij namelijk biometrische gegevens van gebruikers hebben verzameld via de tool Tag Suggestion zonder voorafgaande kennisgeving of toestemming. Deze tool zou mensen scannen en identificeren op foto's die de gebruiker heeft geüpload naar Facebook. Zoals hierboven ook al is aangegeven, vereist de BIPA dat een bedrijf die biometrische gegevens krijgt, de gebruiker op de hoogte moet stellen van het verzamelen van deze gegevens en hoe lang deze gegevens worden verzameld, opgeslagen en gebruikt. Facebook heeft meerdere pogingen gedaan om de zaak te laten seponeren, met het argument dat de eisers geen daadwerkelijke schade hebben geclaimd. Uiteindelijk is het gekomen tot een schikking, waarbij Facebook 650 miljoen dollar heeft betaald.

Conclusie

Hoewel elk van bovenstaande zaken anders is, valt een trend op. Het moge duidelijk zijn dat het in het geval van Artificial Intelligence, het niet echt uitmaakt in waar een zaak voor de rechter komt. Opvallend is dat geen enkele staat of land speciale Artificial Intelligence-wetgeving heeft, die als basis wordt gehouden voor de uitkomst van een zaak. De besproken zaken hierboven worden allemaal het bekende recht beoordeeld, is het niet via productaansprakelijkheid, dan is het wel via privacywetgeving. Opvallend is dat, ook al speelt Artificial Intelligence zo'n grote rol, uiteindelijk voor de 'old school'-wetgeving wordt gekozen, belangrijk is dus niet dat het om Artificial Intelligence gaat, maar belangrijk is wat de precieze schade is of wat het achterliggende probleem is. Duidelijk is dat Artificial Intelligence alleen maar een bijzaak is, maar dat het uiteindelijk toch blijft gaan om het probleem erachter.

4. Ontwikkeling handreiking

Op basis van het overzicht van standaarden, de input van de drie testgroepen en het aanvullende onderzoek is een eerste concepthandreiking gemaakt. De 'kale' versie is hieronder opgenomen. Deze is vervolgens opgemaakt, vormgegeven en gedesigned door Studio Julia Janssen. Een eerste versie van dat design is in de bijlages te vinden.

Non-discriminatie

by

design



Inleiding

Artikel 1 van de Nederlandse Grondwet verbiedt discriminatie: ‘Allen die zich in Nederland bevinden, worden in gelijke gevallen gelijk behandeld. Discriminatie wegens godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht of op welke grond dan ook, is niet toegestaan.’ Non-discriminatie is dus de basis van ons rechtssysteem en onze samenleving.

Dit document is bedoeld voor systeembouwers, data analisten en AI-experts en legt uit welke vragen en principes leidend zijn bij het ontwikkelen en het implementeren van een AI-systeem met het oog op het discriminatieverbod. Stel je wilt een AI-systeem zo non-discriminair mogelijk maken, waar moet je dan aan denken?

Deze handreiking geeft uitgangspunten en vragen en geen ge- of verboden. Vaak is de hoop dat juristen ‘gewoon zeggen’ wat wel en wat niet mag, maar zo werkt het recht niet. Het recht geeft een standaard, een uitgangspunt of een principe, maar daarop bestaan altijd uitzonderingen. Bovendien zal een rechter altijd rekening houden met de context, of wat juristen noemen ‘de omstandigheden van het geval’.

Dat geldt ook voor het non-discriminatierecht. Onderscheid maken op basis van geslacht is bijvoorbeeld verboden, tenzij dit een relevante factor is. Zoekt een castingbureau een actrice om de vrouwelijke hoofdrol op zich te nemen, dan mag het natuurlijk mannen uitsluiten. Sterker nog, positieve discriminatie kan in sommige gevallen geoorloofd zijn; als een organisatie bijvoorbeeld met name mannelijke werknemers heeft, dan mag die besluiten dat bij gelijke geschiktheid aan vrouwen de voorkeur wordt gegeven.

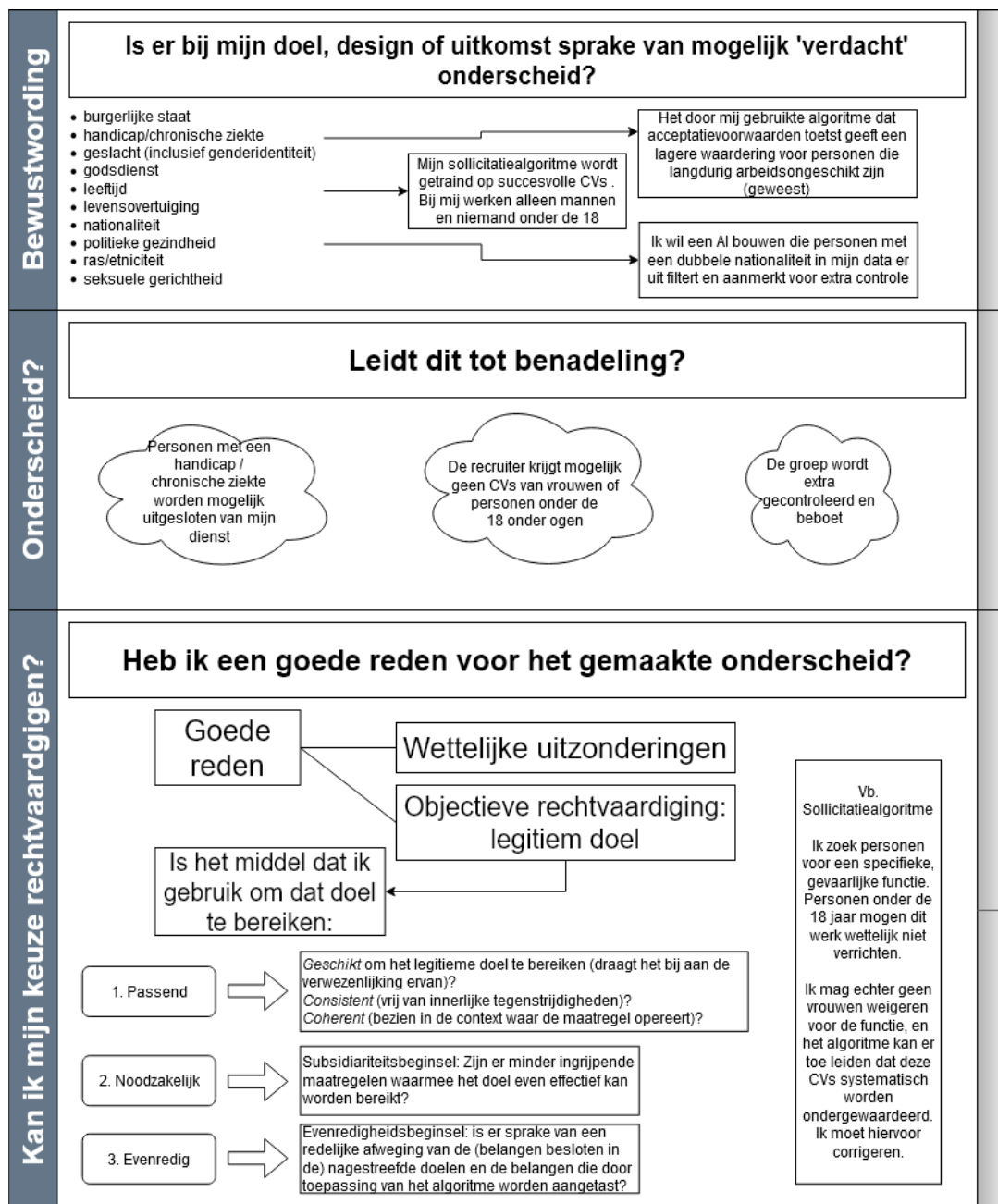
Ook de vraag wat nu precies ‘onderscheid maken’ is, is niet eenduidig te beantwoorden. Het uitgangspunt in het recht is dat gelijke gevallen gelijk behandeld moeten worden, en ongelijke gevallen ongelijk. Maar wat is precies gelijk en wat niet? Welke factoren zijn relevant voor het besluit en welke niet? Al dat soort vragen kunnen niet in het algemeen worden beantwoord in deze handreiking, omdat ze afhangen van de context en de vraag hoe een AI-systeem functioneert, waartoe het dient en welke waarborgen er zijn getroffen.

Belangrijk is dat het recht niet alleen uitgaat van een beperkt aantal gronden waarop in principe geen besluiten mogen worden genomen, zoals ras, geslacht of geaardheid, maar dat het ook discriminatie ‘op welke grond ook’ verbiedt. Dit vergt van een systeembouwer dat die zich goed bewust is op basis van welke groepen het AI-systeem onderscheid maakt en of dat onderscheid te rechtvaardigen is. Is het onderscheid bedoeld of onbedoeld; is het relevant of niet?

Daarnaast is het van belang dat niet alleen ‘directe discriminatie’ verboden is, maar ook ‘indirecte discriminatie’. Als de politie bijvoorbeeld in het kader van een project vooral personenauto’s met een kenteken uit midden of Oost-Europa controleert, dan is dat verboden omdat de eigenaars van die auto’s veelal personen uit midden of Oost-Europa zijn. Zij worden dus meer getroffen door deze maatregel dan personen met een andere nationaliteit. Zo ook kan bijvoorbeeld een postcodegebied een proxy zijn voor ras en etniciteit.

Voorbeelden van toegestane besluiten	Voorbeelden van verboden besluiten
Een medische organisatie selecteert alleen mensen met een bepaalde etnische achtergrond voor het testen van een medicijn, omdat er	Een medische organisatie weigert orthodox gereformeerden te behandelen, omdat die zich niet vaccineren tegen

aanwijzingen zijn dat er binnen die groep grotere en ernstigere bijwerkingen zijn.	ziekten en artsen geloven dat ze het dan aan zichzelf te wijten hebben als ze ziek worden.
De belastingdienst besluit Thaise-massagesalons aan een nadere check te onderwerpen, omdat er aanwijzingen zijn dat die worden gebruikt voor het witwassen van drugsgelden.	De belastingdienst besluit een groep nader te controleren en doet dat niet op basis van etniciteit, maar op basis van de vraag of iemand een tweede huis heeft, waarbij 90% van die groep een immigratieachtergrond heeft.



Deze handreiking heeft ten doel systeembouwers te helpen bij het inrichten van het data-management, het bouwen van een algoritme en het ordenen van processen om tot besluiten te

komen. Het gaat dus met name om de stappen die voorafgaan aan een uitkomst, terwijl het discriminatierecht over deze voorfasen weinig zegt. Het gaat in dit rechtsgebied om een reeds genomen of op zijn hoogst een voorgenomen besluit. Dat wordt vervolgens getoetst op twee vragen: (1) is het besluit tot stand gekomen op basis van een van de verboden gronden (directe discriminatie) of (2) heeft het besluit onevenredig nadelig effect op categorieën personen met bepaalde kenmerken en indien (1) of (2), is dat legitiem?

Deze handreiking vervaalt deze *ex post* toets (oftewel een toets nadat een beslissing is genomen) naar *ex ante* zorgvuldigheidsnormen (oftewel normen die voorafgaan aan een beslissing). Hieruit volgen twee beperkingen. Allereerst kan niet worden gezegd dat als aan de principes uit deze handreiking wordt voldaan er nooit een discriminatoire uitkomst kan zijn; daarop moet altijd *ex post* worden getoetst. Daarnaast is het omgekeerde ook niet waar; als de in deze handreiking vervatte principes niet worden gevolgd betekent dat niet dat de discriminatieregels zullen worden overtreden. Als een AI-systeem wordt gebruikt voor het voorspellen van astrologische processen, dan zal er immers niet snel een discriminatie-aspect mee gemoeid zijn.

Omdat het discriminatierecht zelf weinig *ex ante* regels bevat, wordt in deze handreiking naast dit rechtsgebied ook gekeken naar het privacy- en gegevensbeschermingsrecht, statistische principes, organisatorische handvaten en technologische best practices. Samen bij elkaar vormt dat een pakket aan principes voor het bouwen van een non-discriminatoire AI-systeem. De relevante vragen en uitgangspunten worden in deze handreiking weergegeven aan de hand van fasen die zijn gebaseerd op het Cross-industrie standaardproces voor datamining ([CRISP-DM](#)). Per stap zullen steeds eerst de vragen worden geclusterd, daarna zullen ze ter illustratie kort worden uitgewerkt aan de hand van drie hypothetische casus, waarbij niet alle, maar een aantal vragen worden uitgewerkt, en tot slot zullen concrete juridische, organisatorische en technische handvatten worden gegeven. Drie uitgangspunten zijn leidend voor het hele proces en zullen in alle fasen in acht moeten worden genomen.

1. Diversiteit

De meeste vragen die in dit document staan kunnen niet in isolement worden beantwoord. Het document probeert een brug te slaan tussen de juridische wereld en de technische wereld. Terwijl juristen gewend zijn aan hele algemene uitgangspunten zoals ‘wees transparant’ en ‘vermijd discriminatie’ moeten bij het bouwen van een AI-systeem concrete keuzes worden gemaakt. Welk deel van het proces moet bijvoorbeeld transparant zijn en voor wie moet het proces begrijpelijk zijn: de systeembouwer zelf, een rechter of voor een leek? Hetzelfde geldt voor discriminatie: de keuze om de ene bias te vermijden betekent vaak dat een andere bias de kop op steekt.

Het is daarom belangrijk om het team dat aan een AI-project werkt zo divers mogelijk samen te stellen. Dat geldt voor de verschillende expertises en professionele achtergronden van mensen, maar ook voor de persoonlijke achtergrond binnen het team. Let daarbij op etniciteit, geslacht, gaardheid, culturele en religieuze achtergrond, leeftijd en andere aspecten die bij de concrete toepassingsfunctionaliteit van het AI-systeem relevant kunnen zijn.

2. Contextgevoeligheid

AI-systemen hebben uiteindelijk invloed op de praktijk, vaak binnen een specifieke context. Daarom is het belangrijk om al bij het ontwerp van het systeem specifieke domeinkennis binnen het team te hebben. Vaak worden AI-systemen gebouwd in isolatie, met een mismatch met de

realiteit ten gevolg. Dat heeft niet alleen nadelige effecten op de effectiviteit, maar kan ook ongerechtvaardigde discriminatie in de hand werken, bijvoorbeeld omdat de bouwers van een AI-systeem niet hebben gezien dat een bepaald datapunt een proxy is voor een van de gronden die het anti-discriminatierecht een rol spelen.

Daarnaast is het belangrijk om in een vroegtijdig stadium belanghebbenden bij het ontwerp te betrekken. Stel dat een AI-systeem met 80% accuraatheid een bepaalde diagnose kan stellen, terwijl artsen dat slechts met 60% kunnen, maar het nadeel van het systeem is wel dat het is getraind op data met betrekking tot mannen en dus een veel hogere foutmarge heeft ten aanzien van vrouwen. Dat betekent niet dat het per definitie verboden is om zo'n systeem in te voeren, maar wel dat het belangrijk is om al in een vroegtijdig stadium in gesprek te gaan met patiëntenorganisaties over de inrichting van het ontwikkel- en besluitvormingsproces en het mitigeren van potentiële problemen. Kunnen er alsnog data worden verzameld die het beeld compleet maken? Moet het AI-systeem alleen worden toegepast ten aanzien van diagnoses bij mannen? Kan een deel van de besparing in kosten en middelen worden ingezet voor extra capaciteit bij het diagnosticeren van vrouwen zonder een AI-systeem? Etc.

3. Controleerbaarheid

Het proces en de stappen die worden gezet moeten inzichtelijk, systematisch en controleerbaar zijn. Daarom is het belangrijk dat alle keuzes goed worden gedocumenteerd en verantwoord. Dit zorgt er voor dat achteraf kan worden gecontroleerd of er fouten zijn gemaakt en het systeem later beter kan worden geüpdatet. Het uitgangspunt is daarnaast dat onderdelen in het proces zo concreet mogelijk worden gedocumenteerd, zodat alle stappen herhaalbaar en verifieerbaar zijn.

Belangrijk is om te zorgen voor interne en externe controle op de procesonderdelen. Dat kan bijvoorbeeld door een second opinion te vragen van externe experts, door testen te doen op dezelfde data met een ander algoritme of door het proces twee keer te doorlopen met een andere fairness definitie. Belangrijk is daarbij dat het geheel een iteratief proces is: het AI-systeem is een lerend systeem en is voortdurend in ontwikkeling en dus moet er constant worden gemonitord of aan alle voorwaarden en non-discriminatie principes is voldaan. Het doorlopen van deze handreiking is dus geen garantie voor het volledig uitbannen van discriminatie bij AI-systemen. Ook is het nadrukkelijk niet bedoeld als een lijst met vragen die simpelweg afgevinkt kunnen worden. Het is opgesteld om op een gesystematiseerde en bewuste wijze keuzes te maken over hoe non-discriminatie principes in AI processen kunnen worden ingebed.

Leeswijzer

Tot slot. Deze handreiking is geschreven voor de projectleider van een AI-systeem en heeft ten doel de projectleider de juiste fasen te laten onderscheiden, de juiste mensen op juiste momenten bij elkaar te zetten en hen de juiste vragen te laten stellen. Uiteraard moet worden bedacht dat besluitvormingsprocessen die geen gebruik maken van AI nu ook vaak gebiased zijn, maar dat er weinig kennis en data over dit soort beoordeelbaarheid is. Het werken met AI is dan ook niet per sé een risico in termen van discriminatie, het is juist een kans om processen neutraler en rechtvaardiger te maken en beter in beeld te krijgen welke groepen worden getroffen door bepaald handelen, beleid of besluiten.

(1) Probleemdefinitie

Doel en noodzaak

Wat is het probleem en hoe gaat AI helpen het probleem op te lossen?

Wat is het doel van het project?

Is het noodzakelijk om met AI te werken of kan het probleem ook zonder een AI-systeem worden geadresseerd?

Welke veronderstellingen over het onderscheid tussen verschillende groepen liggen ten grondslag aan de formulering van het doel van het systeem?

Bestaan er verschillende opvattingen over het doel van het systeem en zijn de verschillende belanghebbenden daarin gehoord?

Wat is het probleem en volgens wie en waarom moet daar iets aan gebeuren?

Welke groepen worden onderscheiden in de probleemdefinitie(s) en waarom?

Welke verandering en voor wie moet het systeem teweegbrengen en waarom?

Impact

Moeten er voor dit project meer data worden verzameld en verwerkt dan reeds beschikbaar zijn binnen de organisatie en welke gevolgen heeft dat voor burgers?

Welke impact heeft het systeem op burgers en de maatschappij ten positieve en ten negatieve?

Wordt het systeem gebruikt om informatie te verkrijgen, om besluiten voor te bereiden of om zelfstandige besluiten te nemen en welke gevolgen heeft dat voor de mate waarin AI bepalend zal zijn in de praktijk?

Welke impact hebben fout-positieven en fout-negatieven op de burger en de maatschappij?

Welke procedures zijn ingebouwd voor belanghebbenden om een beslissing (foutnegatieven/foutpositieven) aan te vechten?

Benchmarks

Wat zijn de financiële, computationele en organisatorische kosten voor dit systeem en welke kosten zouden er zijn als er voor een niet-AI gedreven oplossing zou worden gekozen?

Wanneer is het AI-systeem een succes, bijvoorbeeld bij welk percentage van effectiviteit, en wanneer moet deze benchmark zijn gehaald, bijvoorbeeld na 1 maand of 2 jaar?

Welk percentage in foutnegatieven en foutpositieven is acceptabel?

Wat betekenen de verschillende succescriteria voor verschillende groepen?

Voorbeeld arbeidsmarkt	AI-inzet	Voorbeeld strafrechtkenen	AI-inzet	Voorbeeld medische domein	AI-inzet
Het probleem is dat het manueel beoordelen van brieven veel tijd kost en manuele selectie ge-biased is (human bias).		Het probleem is dat de politie nu vaak pas na de feiten ter plaatse is.		Het probleem is dat het manueel beoordelen van scans veel tijd kost, duur is en niet altijd accuraat is.	
Dit systeem dient om een voorselectie te maken van sollicitatiebrieven.		Dit systeem dient om te voorspellen waar en wanneer een overtreding zal plaatsvinden en aldus preventief en gericht patrouilles in te zetten.		Dit systeem dient om betere diagnoses te stellen voor een specifieke vorm van kanker.	
Dit systeem heeft ten doel het proces efficiënter en objectiever te maken.		Dit systeem heeft ten doel het proces kosten-efficiënter, nauwkeuriger en objectiever te maken.		Dit systeem heeft ten doel het proces kosten-efficiënter en nauwkeuriger te maken.	
Voor de training van dit systeem wordt gebruik gemaakt van reeds binnengekomen sollicitatiebrieven.		Gebruik van reeds bestaande databases, aangevuld met gescrapte social media data.		Data worden verzameld gedurende de trialfase van dit systeem in verschillende instituten over de wereld.	
Het systeem maakt zelf beslissingen.		Het systeem maakt voorspellingen en informeert de corpsleiding.		Het systeem maakt zelf beslissingen bij de duidelijke gevallen en geeft de keuze aan een arts bij een hoge onzekerheidsmarge.	
Foutpositieven betekent een minder efficiënt systeem, foutnegatieven betekent geen baankans voor een sollicitant.		Foutpositieven betekent efficiëntieverlies, foutnegatieven betekent dat overtreding zal plaatsvinden.		Foutpositieven leiden tot angst, foutnegatieven leiden in extremo tot de dood.	
Succes bij 20% kostenreductie en 5% minder onterechte afwijzing t.o.v. het bestaande proces. Moet zich uiterlijk na 1 jaar hebben bewezen.		Succes bij 10% hogere effectiviteit van politiesurveillances, gemeten aan het aantal arrestaties en staande houdingen. Moet zich na 3 jaar hebben bewezen.		Succes bij 10% hogere precisie van de diagnoses, gemeten ten opzichte van de foutnegatieven die nu bestaan. Kosten mogen niet hoger zijn dan het nu bestaande proces. Er zit geen einddatum aan dit proces, maar bij de constatering van een hogere foutmarge t.a.v. foutnegatieven wordt het project onmiddellijk stopgezet.	
Maximaal fout positief: 60% (t.o.v. de mensen die worden uitgenodigd voor een gesprek) en fout negatief: 2% (t.o.v. de mensen die zouden worden uitgenodigd voor een gesprek door manuele selectie)		Gedurende de 3 jaar worden surveillances uitgevoerd op basis van het bestaande proces en anderen met behulp van het AI-systeem. Die laatsten mogen geen hogere foutpositieven of -negatieven opleveren.		Artsen blijven diagnoses uitvoeren zoals nu, daarnaast doet een AI-systeem diagnoses. Een onafhankelijke arts vergelijkt en controleert de resultaten.	

Juridisch	Technisch	Organisatorisch
Er dient vooraf een helder en concreet doel voor het systeem te worden gedefinieerd. Ga na of er persoonsgegevens worden verwerkt, dat wil zeggen gegevens die kunnen worden herleid tot concrete individuen. Klik hier voor uitleg.	Beargumenteer en documenteer de keuze voor AI in relatie tot context: b.v. keuze voor doelvariabele, classificatietaak, prestatiedoelen, etc. Formuleer en documenteer de logica en het waarom van het AI systeem. Ga na en documenteer wat de beperkingen zijn van de verschillende aanpakken met betrekking tot de representatie van het probleem en de belanghebbenden.	Loop de volgende vragen langs: 1. Heeft het team de toegang tot de benodigde middelen? 2. Heeft het team de benodigde bevoegdheden? 3. Is alle relevante expertise aanwezig in het team? 4. Is het team zo divers mogelijk? 5. Is een persoon met domeinkennis betrokken bij het project? 6. Wie is er voor welk onderdeel van het proces verantwoordelijk en hoe vindt de interne communicatie en besluitvorming plaats?
Er dient een Impact Assessment te worden uitgevoerd, waarin wordt gekeken naar de impact van het systeem op alle mensenrechten en hoe die gemitigeerd worden. Als er hoge risico's blijven bestaan na aanvullende maatregelen dan moet voorafgaande toestemming van de Autoriteit Persoonsgegevens worden gevraagd. Klik hier voor een model.	Beargumenteer de keuze voor een type AI-systeem: is het een eenvoudige beslisboom, een zelflerend systeem of deep learning? Waartoe wordt het systeem ingezet: het verkrijgen van inzichten, het voorbereiden van besluiten of het nemen van zelfstandige besluiten? Klik hier voor uitleg.	Breng in kaart binnen welk domein het systeem een rol zal gaan spelen en wat de maatschappelijke context is. Breng in kaart wie de belanghebbenden zijn en bij wie de voordelen van het systeem en bij wie de nadelen komen te liggen. Betrek de relevante belanghebbenden of vertegenwoordigende (burgerrechten) organisaties in een vroegtijdig stadium.
Juridisch gezien moeten systemen noodzakelijk, proportioneel en subsidiair zijn. Dat betekent dat de nadelen van zo'n systeem niet groter mogen zijn dan de voordelen en dat er geen andere, minder ingrijpende manieren zijn om hetzelfde doel te verwezenlijken. Klik hier voor uitleg.	Bepaal de benchmarks ten aanzien van foutpositieven en -negatieven, de effectiviteit en de looptijd van het project. Voer een technische evaluatie uit. Een voorbeeld is de Ethics Canvas als ontwikkeld door het Open Data Institute. Klik hier voor de Canvas.	Documenteer alle stappen en keuzes in het proces en overleg die keuzes zowel binnen het team als met belanghebbenden. Bij twijfel, contacteer een externe expert voor een second opinion. Leg vast wanneer het proces zal worden stopgezet en wat de exit strategie is.

(2) Dataverzameling

Doel en noodzaak

Welke data zijn nodig voor dit project en waarom?

In hoeverre zijn deze gegevens al binnen de organisatie beschikbaar en in hoeverre moeten ze van buiten worden gehaald?

Is het toegestaan om deze data voor dit project te verzamelen en te verwerken?

Datakwaliteit

Welke bias zit er in de data (van binnen, buiten of gecombineerd) en welke consequenties heeft dat?

Zijn de data representatief en zijn alle relevante groepen in gelijke mate vertegenwoordigd?

Als verschillende databronnen worden gebruikt, hoe wordt er gezorgd dat deze data compatibel en vergelijkbaar zijn?

Dataopslag

Hoe lang worden de gegevens bewaard en hoe?

Worden de gegevens veilig en vertrouwelijk behandeld; welke gevolgen heeft een datalek voor groepen of categorieën personen?

Worden data gedeeld met andere partijen en wat is het gevaar dat die misbruik maken van de data met negatieve gevolgen voor groepen of categorieën personen?

Voorbeeld arbeidsmarkt	AI-inzet	Voorbeeld strafrechtketen	AI-inzet	Voorbeeld medische domein	AI-inzet
Data van 30 random vroegere sollicitatieprocedures zullen worden beoordeeld op basis van daadwerkelijke uitkomst en de uitkomst als gesuggereerd door het AI-systeem. Data is afkomstig van de organisatie zelf. Het gaat hier om secundair gebruik van data dat is geoorloofd vanwege het publieke belang van het voorkomen van gebiasde sollicitatieprocedures.		Data over effectiviteit van vroegere surveillance en criminaliteitscijfers (beschikbaar binnen de organisatie). Deze data worden gekoppeld aan open data afkomstig van sociale media. De data worden geanonimiseerd en geaggregeerd op groepsniveau. De verwerking vindt plaats op basis van statistische data en niet op persoonsgegevens.		Er zijn zo veel mogelijk data nodig over diagnoses gesteld door artsen en second opinions door het AI-systeem. Hoe presteert het AI-systeem ten opzichte van een arts? Er wordt slechts gewerkt met data die patiënten zelf afgeven. Hiervoor geven zij geïnformeerde toestemming.	
Er zit een grote bias in de bestaande dataset, onder meer op basis van etniciteit en afkomst. De bias in bestaande data is de reden om dit systeem te introduceren. Succes van het AI-systeem zal dus niet worden beoordeeld op basis van de mate waarin de uitkomsten gelijk zijn aan vroegere processen. Een onafhankelijk team zal beoordelen welke van de selectiemethoden de meest rechtvaardige uitkomst oplevert.		Er zit een historische bias in de dataset van de politie, onder meer ten aanzien van bepaalde wijken, personen met een immigratieachtergrond, lage sociaaleconomische status en mannen. Er zit een bias in de social media die vooral door jongeren worden gebruikt. Deze twee databronnen worden onafhankelijk van elkaar beoordeeld en niet direct gekoppeld. Er zal worden gecorrigeerd op deze bias.		Er zal een bias zitten in deze patiëntpopulatie, namelijk vrouwen van 50+ met overgewicht. Het AI zal dus leren om met name ten aanzien van deze groepen accurate voorspellingen te doen. Dat kan betekenen dat op termijn, het AI-systeem wordt ingezet voor diagnoses bij deze groep en artsen voor andere groepen, al naar gelang de accuratesse daartoe aanleiding geeft.	
Data worden slechts voor de duur van dit project bewaard en daarna verwijderd. Gedurende het project staat de data op een interne cloud Deze is slechts toegankelijk is voor medewerkers van de afdeling werving en selectie en voor ontwikkelaars van het AI-systeem		Data worden in geanonimiseerde vorm oneindig bewaard omdat ook historische patronen van gro waarde zijn voor toekomstige processen. De data worden binnen een beveiligde ruimte opgeslagen. Slechts leden van een speciale unit hebben toegang tot deze data.		Data worden voor 20 jaar bewaard, om kwaliteitscontroles en assessments mogelijk te maken. Alleen onderzoekers en data-analisten hebben toegang tot de data. De data van de verschillende instituten worden aan elkaar gekoppeld en worden gedeeld in een secured private cloud.	

Juridisch	Technisch	Organisatorisch
Beoordeel of er een legitieme grondslag is voor het verwerken van persoons-gegevens, zoals een wettelijke plicht, informed consent of publiek belang. Klik hier voor meer uitleg. Als er gegevens over iemands ras, gezondheid, strafrechtelijk verleden of andere gevoelige gegevens worden verwerkt, dan is expliciete toestemming, een wettelijke grondslag of een zwaarwegend publiek belang nodig. Klik hier voor uitleg. Als er gegevens uit andere bron(nen) worden verkregen, dan moet worden geverifieerd of die data op een legitieme wijze zijn verkregen.	Zorg dat data zo worden verzameld dat ze kunnen worden benut door het AI-systeem. Leg daarbij van te voren technische vereisten vast. Beoordeel data op: 1. Verdeling van attributen (bijvoorbeeld het doelattribuut van een voorspellingstaak) 2. Relaties tussen paren of een klein aantal attributen 3. Resultaten van eenvoudige aggregaties 4. Eigenschappen van significante subpopulaties Beoordeel data op mogelijk aanwezige bias aan de hand van o.a.: 1. Verdeling van de doelvariabele in verschillende groepen. <i>Voorbeeld:</i> Wanneer de proportie 'positieven' voor vrouwen anders is dan voor mannen, kan dit een teken zijn van historische bias. 2. Relaties tussen paren of een klein aantal attributen (<i>features</i>): leidt dit tot proxies? <i>Voorbeeld:</i> postcode kan een proxy zijn voor etniciteit. 3. Verdeling van attributen en representatie van relevante subpopulaties. <i>Voorbeeld:</i> wanneer een training dataset van gezichtsherkenningsoftware weinig (i.e., lage representatie) foto's bevat van donkergetinte gezichten (i.e., andere verdeling van attributen) bevat, bestaat het risico dat het systeem slechter zal werken voor deze groepen.	Zorg dat data zo neutraal en objectief mogelijk worden verzameld en dat dit zo transparant en controleerbaar mogelijk geschiedt. De wijze waarop data worden verzameld, hoe, door wie, waar en met gebruikmaking van welke technieken kunnen bepalend zijn voor de neutraliteit en betrouwbaarheid van de verkregen data. Leg dit proces nauwgezet vast en bewaar te allen tijde gegevens over hoe de dataset tot stand is gekomen, om zo de processen controleerbaar en herhaalbaar te maken. Doe dat ook voor de data die via derden zijn verkregen.
Data moeten correct zijn, up to date worden gehouden en betrokkenen moet de mogelijkheid worden geboden om aanvullende gegevens aan te dragen. Klik hier voor een voorbeeldbrochure.	Controleer data nogmaals op bias: 1. Historische bias: ongelijkheid in de samenleving is vaak ook terug te zien in op zich neutraal verzamelde data 2. Representatie bias: ontstaat wanneer delen van de inputruimte zijn onder- of overgerepresenteerd 3. Measurement bias: meten we wel wat we willen meten? Zijn verkoopcijfers bijvoorbeeld wel een goede proxy voor de verkoopkwaliteiten van werknemers?	Controleer op ontbrekende waarden, nauwkeurigheid en representativiteit en of de verdeling voor verschillende groepen gelijk is. Leg verschillen uit en overweeg mitigerende maatregelen. Overweeg om nieuwe gegevens te verzamelen of om het doel van het project te herzien: ga dan terug naar fase 1 (probleemdefinitie).
Als gegevens de EU-grens overgaan, en bijvoorbeeld naar de VS of China worden doorgevoerd, moet de organisatie waarmee de gegevens worden gedeeld zich aan de in de Europese Unie geldende gegevens-beschermingsregels houden. Klik hier voor nadere uitleg.	Zorg voor de technische veiligheid en vertrouwelijkheid van de data. Zorg dat de data geëncrypteerd en gecompartmentaliseerd zijn.	Zorg voor organisatorische veiligheid en vertrouwelijkheid van de data. Zorg dat data alleen met dubbele authenticatie in te zien zijn door een geselecteerd aantal werknemers of betrokkenen.

(3) Datavoorbereiding

Inclusie en exclusie

Welke van de verzamelde data zijn relevant voor het model en waarom?

Aan de hand van welke criteria wordt deze keuze gemaakt?

Wat gebeurt er met de data die niet worden gebruikt?

Aan de hand van welke criteria wordt de keuze voor dataselectie gemaakt en hoe reflecteren die onderscheid tussen groepen?

Beïnvloedt de keuze voor bepaalde data of databewerkingen de probleemdefinitie?

Welke aspecten van het probleem worden buiten beschouwing gelaten?

Integratie en aggregatie

Hoe wordt gezorgd dat historische data en nieuw verzamelde data op elkaar aansluiten; zijn de data vergelijkbaar en welke aannames ten aanzien van groepen en categorieën zit er in de reeds verzamelde data en de nieuw te verzamelen data?

Hoe wordt gezorgd dat data uit verschillende bronnen op elkaar aansluiten; zijn de data vergelijkbaar en welke aannames ten aanzien van groepen en categorieën zit er in de verschillende databronnen?

Op welke wijze worden data geaggregeerd en welke gevolgen heeft dat voor de representativiteit van de data?

Wat betekent dit voor de representatie van het probleem en de belanghebbenden? B.v. betekent dit een herformulering van een groep of aggregatie van groepen?

Leidt de combinatie van verschillende data tot proxies en zo ja welke?

Labelen

Hoe worden data gelabeld en waarom?

Sluit dit aan bij hoe andere organisaties data labelen en datasets gebruiken waarop het algoritme is getraind?

Sluit dit aan bij hoe andere belanghebbenden en domein experts data zouden labelen?

Zitten er gevoelige labels over bijvoorbeeld etniciteit, geaardheid of geslacht bij of labels die daar indirect naar verwijzen, zoals postcodegebieden, en zo ja, waarom?

Voorbeeld arbeidsmarkt	AI-inzet	Voorbeeld strafrechtketen	AI-inzet	Voorbeeld medische domein	AI-inzet
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	
Data wordt gelabeld op basis van kwaliteiten van de sollicitanten, om zo een objectieve keuze te kunnen maken voor een sollicitant.		Data wordt gelabeld op basis van wijken/postcodegebieden, om zo in kaart te brengen welke wijken meer aandacht behoeven.		Data wordt gelabeld op basis van onder andere leeftijd, etniciteit en geslacht, om zo te zien of deze bepaalde vorm van kanker meer voor komt in bepaalde groepen.	
Er zitten gevoelige gegevens tussen, maar deze spelen geen rol bij de labeling.		Er zitten gevoelige labels tussen, namelijk postcodegebieden, Deze kunnen verwijzen naar bepaalde etniciteit en naar sociaaleconomische status.		Er zitten gevoelige labels tussen, deze zijn nodig om in kaart te kunnen brengen in welke groepen een groter percentage deze vorm van kanker heeft.	

Juridisch	Technisch	Organisatorisch
Controleer welke gevolgen de dataselectie heeft voor de representatie, met name van categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid.	Pas pre-processing technieken voor bias-mitigation toe, zoals: 1. Instance class modification 2. Instance selection 3. Instance weighting Welke technische beperkingen brengen de data met zich mee? Doe aanpassingen aan het technisch systeem al naar gelang de datakwaliteit dat vereist. Heeft de datakwaliteit gevolgen voor fase 1 (probleemdefinitie)?	Leg nauwgezet de keuzes vast ten aanzien van inclusie en exclusie. Welke rationale is er gekozen voor de inclusie of exclusie van de data en waarom? Welke gevolgen heeft de kwaliteit en representativiteit voor de werkzaamheid van het systeem en de gestelde benchmarks? Verzamel eventueel meer gegevens als blijkt dat bepaalde groepen over- of onder gerepresenteerd zijn: ga dan terug naar fase 2.
Controleer wederom welke gevolgen de integratie en aggregatie van data hebben voor categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status. Zelfs het samenvoegen van twee neutrale databases kan samen een database met sterke bias opleveren.	1. Zijn de data op dezelfde wijze verzameld? 2. Welk invloed heeft het verschil in de wijze (methodologie, tijd, plaats, etc.) van de dataverzameling op de mogelijkheid van integratie en vergelijkbaarheid van de data? 3. Welke categorieën worden gekozen ten aanzien van het aggregatieproces en waarom? Controleer op aggregatie bias: die kan ontstaan bij het bouwen van een model. Houd er rekening mee dat de vertaling van input naar labels niet altijd voor alle groepen hetzelfde is. Vermijd een “all-size-fits-all” model.	Datasets, analysemethode en beslissingen worden gekozen met het oog op objectiviteit; fouten worden vastgelegd en direct gecorrigeerd. Als ze van belang zijn voor andere organisaties worden die onmiddellijk gedeeld. Methodes, procedures, definities en classificaties worden consistent toegepast; deze worden zo veel mogelijk gestandaardiseerd tussen organisaties, om vergelijkingen en controle mogelijk te maken.
Controleer welke gevolgen het labelen van data heeft voor de categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status.	Controleer op meting bias: Deze kan ontstaan als de keuze van labels bevooroordeeld zijn. Houd er ook rekening mee dat de granulariteit en kwaliteit van data tussen verschillende groepen kan verschillen en dat elke classificatie een versimpeling is van de werkelijkheid.	Controleer de gevolgen van het labelen op de kwaliteit en representativiteit en de werkzaamheid van het systeem. Verzamel eventueel meer gegevens als blijkt dat bepaalde groepen over- of onder gerepresenteerd zijn: ga dan terug naar fase 2.

(4) Modelering

Pre-modelering

Welk algoritme wordt er gekozen en waarom?

Welk modeltype wordt er nagestreefd en waarom?

Hoe worden criteria op het gebied van uitlegbaarheid en *fairness* vertaald naar de modelselectiestrategie?

Model(selectie)

Welke parameters worden er voor het model gekozen en waarom?

Is het voldoende om één model te bouwen, of is het beter om meerdere modellen te bouwen en naast elkaar te leggen?

Is het model gebaseerd op bestaande modellen en waarom wel of niet?

Test

Hoe presteert het model op effectiviteit?

Hoe presteert het model op de gekozen fairness definitie?

Hoe presteert het model op de gekozen benchmark van foutpositieven en foutnegatieven?

Voorbeeld arbeidsmarkt	AI-inzet	Voorbeeld strafrechtkenen	AI-inzet	Voorbeeld medische domein	AI-inzet
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	

Juridisch	Technisch	Organisatorisch
Wordt er gekozen voor een model gebaseerd op causaliteit of correlatie? Wees bewust van het feit dat het juridische domein is gebaseerd op causaliteit; juridische uitlegbaarheid en juridische rechtvaardiging kunnen vaak niet worden gevonden in statistische correlaties. Ga dus na of bij een systeem gebaseerd op correlatie, de uitkomsten kunnen worden gekoppeld aan causaliteit op individueel niveau. Deep-learning systemen met black box elementen die besluiten nemen zijn bijna altijd verboden.	Algoritme selectie: Verklaar en documenteer keuzes voor een ML-algoritme in het licht van interpreteerbaarheid en verklaarbaarheid. Modelkeuze: 1. Overweeg de vooraf gedefinieerde fairness-maatstaven bij het selecteren van een model. 2. Controleer of aangeleerde relaties een afspiegeling zijn van bestaande domeinexpertise en niet willekeurig zijn 3. Gebruik het eenvoudigste model dat de prestatiespecificaties mogelijk maken Selecteer de kandidatenpoel van AI-algoritmes in het licht van benodigde <i>fairness</i> interpreteerbaarheid, en uitlegbaarheid. <i>Voorbeeld:</i> wanneer <i>fairness</i> criteria gesteld zijn, overweeg het gebruik van in-processing of post-processing <i>bias mitigation</i> methodes om hiervoor te optimaliseren. Kijk bijvoorbeeld hier voor een selectie van <i>bias mitigation</i> algoritmes. <i>Voorbeeld:</i> wanneer causaliteit een voorwaarde is, wordt gezorgd dat de kandidatenpoel bestaat uit direct interpreteerbare algoritmes, zoals <i>linear regression</i>. <i>Voorbeeld:</i> wanneer voorspellingen als doel hebben om menselijke beslissingen te ondersteunen, wordt gezorgd dat de kandidatenpoel bestaat uit methodes die uitlegbaar zijn voor mensen met een minder technische achtergrond, zoals een simpele <i>decision tree</i>. Vertaal succescriteria naar technische maatstaven, bijvoorbeeld op het gebied van accuraatheid, <i>false positives</i>, <i>false negatives</i>, en <i>fairness</i>. <i>Voorbeeld:</i> om te voorkomen dat het model minder goed werkt voor minderheden, zal de fairness metric <i>equalized odds</i> worden meegenomen in de modelselectie. Indien gebruik wordt gemaakt van bestaande modellen, neem deze dan mee in de modelselectiestrategie. <i>Voorbeeld:</i> in een NLP-applicatie wordt het gebruik van <i>pre-trained word embeddings</i> overwogen, de evaluatie van de <i>embeddings</i> op de succes criteria is daarom onderdeel van de modelselectiestrategie.	De modelselectiestrategie worden vastgelegd, gedocumenteerd en openbaar gemaakt. Het ontwerp is het liefst universeel, zodat de modellen goed met elkaar kunnen worden vergeleken op uitkomst en fairness. Betrek belanghebbenden, zoals eindgebruikers en besluitvormers, bij de selectie van de kandidatenpoel om te zorgen dat de interpreteerbaarheid en uitlegbaarheid van de modellen past bij het doel van het model en de achtergrond van de gebruikers.
AI-systemen die zijn gebaseerd op statistische modellen kunnen rekening houden met de principes die hiervoor leidend zijn, zoals: 1. Betrouwbaarheid 2. Onpartijdigheid 3. Objectiviteit 4. Vergelijkbaarheid 5. Consistentie Kijk hier voor meer informatie.	Bias-beperking: 1. Gebruik indien nodig voorbehandelingsmethoden om het algoritme aan te passen om vertekende resultaten te minimaliseren, bijvoorbeeld door aanpassing van het classificatiemodel; 2. Pas indien nodig nabewerkingsmethoden toe om bias te beperken na training van het classificatiemodel. White-box-methoden passen het model aan; black-box-methoden passen de voorspellingen aan. Kijk bijvoorbeeld hier voor een selectie van <i>bias mitigation</i> algoritmes. <i>Voorbeeld:</i> wanneer causaliteit een voorwaarde is, wordt gezorgd dat de kandidatenpoel bestaat uit direct interpreteerbare algoritmes, zoals <i>linear regression</i>. <i>Voorbeeld:</i> wanneer voorspellingen als doel hebben om menselijke beslissingen te ondersteunen, wordt gezorgd dat de kandidatenpoel bestaat uit methodes die uitlegbaar zijn voor mensen met een minder technische achtergrond, zoals een simpele <i>decision tree</i>. Vertaal succescriteria naar technische maatstaven, bijvoorbeeld op het gebied van accuraatheid, <i>false positives</i>, <i>false negatives</i>, en <i>fairness</i>. <i>Voorbeeld:</i> om te voorkomen dat het model minder goed werkt voor minderheden, zal de fairness metric <i>equalized odds</i> worden meegenomen in de modelselectie. Indien gebruik wordt gemaakt van bestaande modellen, neem deze dan mee in de modelselectiestrategie. <i>Voorbeeld:</i> in een NLP-applicatie wordt het gebruik van <i>pre-trained word embeddings</i> overwogen, de evaluatie van de <i>embeddings</i> op de succes criteria is daarom onderdeel van de modelselectiestrategie. Controleer of aangeleerde relaties een afspiegeling zijn van bestaande domeinexpertise en niet willekeurig zijn. Controleer hoe de modellen presteren t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid. Selecteer het eenvoudigste model dat de prestatiespecificaties mogelijk maakt. <i>Voorbeeld:</i> zowel een <i>logistic regression</i> model als een <i>random forest</i> model zijn voldoende accuraat, er wordt daarom gekozen voor het <i>logistic regression</i> model.	Verantwoording: 1. Metadata worden bewaard en vastgelegd; 2. Data zijn voor zover mogelijk toegankelijk voor derden; 3. Het model moet uitlegbaar zijn aan en inzichtelijk voor belanghebbenden 4. Welk vorm van uitlegbaarheid wordt door het systeem geboden? 5. Voor wie is deze uitleg begrijpelijk?
Controleer hoe het model presteert t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid.	Hoe presteert het model op de benchmarks van effectiviteit, de gekozen fairness definitie en de benchmark van foutpositieven en foutnegatieven. Stel het model bij aan de hand van de uitkomsten. Controleer op Evaluatie bias: dit kan voorkomen bij het testen van het model. Let erop dat de benchmarks die gebruikt worden om het systeem te testen overeenstemmen met de beoogde doelgroep.	Vraag een onafhankelijk team van experts met een diverse persoonlijke en professionele achtergrond voor een second opinion.

(5) Implementatie

Praktijktest

Wat is de toepassingsstrategie?

Welke beperkte en afgebakende testcase is representatief en kan goed worden gemonitord?

Hoe werkt het model en is dat volgens verwachting?

Aanpassing model

Welke aanpassingen zijn er nodig om de werkzaamheid te verhogen?

Welke aanpassingen zijn er nodig om de fairness van het model te verhogen?

Welke aanpassingen zijn er nodig om de foutmarges te verkleinen?

Toepassing

Welke beperkingen volgen uit de vorige stappen voor de toepassingsmogelijkheden en het implementatietraject?

Welke aandachtspunten zijn er voor de toepassing en hoe kan er bij de implementatie voor worden gezorgd dat deze goed kunnen worden gemonitord?

Hoe worden belanghebbenden en anderen op de hoogte gesteld en betrokken?

Voorbeeld arbeidsmarkt	AI-inzet	Voorbeeld strafrechtkenen	AI-inzet	Voorbeeld medische domein	AI-inzet

Wordt nog ingevuld

Wordt nog ingevuld

Wordt nog ingevuld

Wordt nog ingevuld

Wordt nog ingevuld

Wordt nog ingevuld

Wordt nog ingevuld

Wordt nog ingevuld

Wordt nog ingevuld

Juridisch	Technisch	Organisatorisch
Stel de belanghebbenden op de hoogte van hun rechten: 1. Recht op informatie, inclusief informatie over algoritme 2. Recht om het besluit aan te vechten 3. Recht om aanvullende informatie aan te dragen 4. Recht om niet onderworpen te worden aan automatische besluitvorming Klik hier voor meer informatie.	Beantwoord de volgende vragen: 1. Wat is de toepassingsstrategie? 2. Welke beperkte en afgebakende toepassing is representatief en kan goed gemonitord worden? 3. Hoe werkt het model en is dat volgens verwachting? 4. Wat is de exitstrategie? Overweeg gebruikerstesten om gebruiksgemak en toegankelijkheid voor mensen met een beperking te toetsen. Wanneer het model gebruik maakt van persoonsgegevens, neem dan een vergelijking van de prestaties van het model en de prestaties binnen het huidige besluitvormingsproces op in de evaluatiestrategie. <i>Voorbeeld:</i> de prestaties van het model worden getoetst in een <i>pilot study</i> bij een kleine groep gebruikers. <i>Monitoring en onderhoud</i> Wanneer het model regelmatig wordt geüpdatet op basis van nieuwe data, zorg dat ook deze nieuwe modellen grondig worden geëvalueerd. Plan voor het monitoren van veranderingen in de datadistributie, zoals <i>concept drift</i> en een verschuiving in de demografie van <i>data subjects</i>.	Stel betrokkenen en belanghebbende op de hoogte van het feit dat het AI-systeem in een testsetting wordt geïmplementeerd. Doe dat waar mogelijk voordat zij in aanraking komen met of de gevolgen ondervinden van het systeem.
Controleer hoe het model presteert t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status.. Als het model significant onderscheid maakt, direct of indirect, op basis van een van deze groepen, is daar een rechtvaardiging voor en zo ja welke? Leg dit voor aan een jurist.	Evalueer het model op: 1. Effectiviteit 2. Fairness 3. Foutpositieven en foutnegatieven Voer verbeteringen door. Beoordeel wat de context van de gekozen testcase zegt over de algemene werking van het systeem en denk na in hoeverre andere toepassingsgebieden andere contextgevoeligheden met zich brengen. Neem daarop passende maatregelen.	Overleg met zowel belanghebbenden als externe experts over de functionaliteit, fairness en accuratesse van het model. Implementeer voor zo ver mogelijk hun adviezen en suggesties. Leg veranderingen in het systeem vast. Documenteer het gekozen model en de resultaten van de evaluatie, bijvoorbeeld in een model card.
Stel wederom belanghebbenden op de hoogte van hun rechten: 1. Recht op informatie, inclusief informatie over algoritme 2. Recht om het besluit aan te vechten 3. Recht om aanvullende informatie aan te dragen 4. Recht om niet onderworpen te worden aan automatische besluitvorming	Monitoring en controle: 1. Plan evaluatiemomenten 2. Documenteer de gebruiksresultaten 3. Controleer op exitcriteria 4. Maak API's beschikbaar voor externe auditors 5. Maak databladen en modelkaarten voor zover mogelijk openbaar.	Stel wederom iedereen op de hoogte van het feit dat het AI systeem wordt geïmplementeerd en voer een klachtenprocedure in. Vraag een onafhankelijk team van experts met een diverse persoonlijke en professionele achtergrond voor een second opinion.

(6) Evaluatie

Evaluatievoorbereiding

Wordt er gekozen voor een permanente evaluatie, specifieke evaluatiemomenten of beide?

Wordt er gekozen voor een interne evaluatie, een evaluatie door externen of beide?

Hoe wordt de evaluatie getest en met welke meetpunten?

Evaluatie

Hoe functioneert het systeem ten aanzien van de benchmarks?

Welke aanpassingen zijn er ten aanzien van de beschermde categorieën nodig?

Hoe zou het systeem functioneren met een ander model, fairness definitie en/of algoritme?

Actiepunten

Moet het systeem al dan niet tijdelijk worden stopgezet?

Kunnen gevonden problemen en obstakels worden verholpen?

Wat vinden belanghebbenden en externe experts van de evaluatieresultaten?

Voorbeeld arbeidsmarkt	AI-inzet	Voorbeeld strafrechtkenen	AI-inzet	Voorbeeld medische domein	AI-inzet
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	
Wordt nog ingevuld		Wordt nog ingevuld		Wordt nog ingevuld	

Juridisch	Technisch	Organisatorisch
Als er persoonsgegevens worden verwerkt moet het systeem duidelijk effectiever zijn dan de status quo zonder het AI-systeem om te voldoen aan de noodzakelijkheids-, proportionaliteits- en subsidiariteitstest. Verantwoord dit.	Leg de volgende punten nauwgezet vast: 1. Hoe wordt de evaluatie ingericht en waarom zo? 2. Waar wordt de verantwoordelijkheid voor de evaluatie belegd en waarom? 3. Welke meetpunten worden gekozen en waarom? Plan een uitgebreide evaluatie waarin wordt Overweeg gebruikerstesten om gebruiksgemak en toegankelijkheid voor mensen met een beperking te toetsen. Wanneer het model gebruik maakt van persoonsgegevens, neem dan een vergelijking van de prestaties van het model en de prestaties binnen het huidige besluitvormingsproces op in de evaluatiestrategie. <i>Voorbeeld:</i> de prestaties van het model worden getoetst in een <i>pilot study</i> bij een kleine groep gebruikers.	Betrek belanghebbenden bij de evaluatie. <i>Voorbeeld:</i> Doe een survey of focusgroep met belanghebbenden om hun ervaringen in kaart te brengen. Stel betrokkenen en belanghebbende op de hoogte van het feit dat het AI-systeem in een testsetting wordt geïmplementeerd. Doe dat waar mogelijk voordat zij in aanraking komen met of de gevolgen ondervinden van het systeem. Ga na of het systeem, het model, de data en/of de evaluatie openbaar worden gemaakt, al dan niet in geanonimiseerde vorm.
Controleer hoe het systeem presteert t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status. Als het model significant onderscheid maakt op basis van een van deze groepen, is daar een rechtvaardiging voor en zo ja welke? Ga na of het systeem al dan niet tijdelijk moet worden stopgezet of dat naar een van de vorige fasen in het proces moet worden teruggegaan om aanpassingen te doen.	Evalueer het model op: 1. Effectiviteit 2. Fairness 3. Foutpositieven en foutnegatieven Voldoen de presentaties aan de voorgestelde benchmarks. Zo nee, stop het project tijdelijk of permanent. Zo ja, zijn er mogelijkheden om het systeem nog effectiever, rechtvaardiger of accurater te maken? Hoe functioneert het systeem: 1. Op andere data? 2. Met een ander algoritme? 3. Met een andere fairness definitie? 4. Met een ander model?	Documenteer de evaluatiemethodes, motivaties voor keuzes, en verantwoordelijkheid voor de evaluatie. Documenteer het gekozen model en de resultaten van de evaluatie, bijvoorbeeld in een model card.
Vraag een externe jurist om een second opinion te doen en advies te geven ten aanzien van de juridische randvoorwaarden voor het systeem.	Vraag een externe data-analist om een second opinion te doen en advies te geven ten aanzien van het technisch systeem.	Vraag een bedrijfskundige om een second opinion te doen en advies te geven ten aanzien van de procesinrichting van het systeem.

5. Validatie standaarden

De eerste concepthandreiking is voorgelegd in drie workshops. De volledige verslagen van die discussies zijn in de bijlages te vinden. Hieronder zijn de belangrijkste inzichten weergegeven. Daarnaast is de concepthandreiking ook besproken met de klankbord van het Ministerie van Binnenlandse Zaken, waarin vertegenwoordigers van tal van Ministeries, overheidsorganisaties en semipublieke organisaties zitting namen.

5.1 Workshop 1

Aan de eerste workshop namen deel:

- Josine van der Brug (Tactlegal)
- Wouter van der Bij (VWS)
- Sander Klous (KPMG & UvA)
- Egge van der Poel (TIAS)
- Erwin Redeman (Neurocast)

Centraal stond het gebruik van AI door [Neurocast](#), dat wordt ingezet om vroegtijdig ziektes als MS en alzheimer op te sporen. Hoe kan de handreiking worden toegepast op deze casus, welke positieve punten volgen uit die toepassing en welke verbeterpunten zijn er voor de handreiking? Daaruit kwamen de volgende aandachtsgebieden naar voren:

- Het belangrijkste voor het waarborgen van maatschappelijke waarden is het begin, de start van een project. In de praktijk wordt een project vaak zonder groot plan aangevangen; hoe een project zich precies ontwikkelt is dan nog onbekend. Wel is het mogelijk om op dat moment een aantal mijlpalen te slaan, benchmarks en doelstellingen te formuleren. Het gaat daarbij om een brede visie op wat het project beoogt, voor wie en ten behoeve van wat.
- CRISP-DM model is een goede basis, maar kent ook beperkingen. Een van de kritiekpunten is dat de evaluatie niet dynamisch en permanent genoeg is. In wezen zouden er een veelheid van kleine feedbackloopjes moeten zijn in het proces: van stap 5 naar 3 en van stap 2 naar 4, etc.
- Het uitgangspunt moet zijn dat de burger eigenaar van zijn data is. De kernvraag is dus ook of de AI toepassing de burger of de gemeenschap ten goede komt. In de medische sector is daarom ook dynamische toestemming van belangrijk.
- Het is belangrijk om in de testfase niet alleen te focussen op zo min mogelijk data; juist gevoelige data kunnen van belang zijn om in latere stadia te controleren op bias.
- Er is een verschil in foutmarge/bias die wordt geaccepteerd ten aanzien van menselijke en computergestuurde beslissingen. Daarom kan het aanbevelingswaardig zijn om voor de start van het systeem te analyseren hoe de huidige (menselijke) praktijk al dan niet discriminatoir is en hoe het AI systeem daar een verbetering op zou kunnen zijn.
- Ten aanzien van transparantie is een onderscheid te maken tussen: (1) technische uitlegbaarheid (bv voor technische audit), (2) procesuitlegbaarheid (die juridisch het meest relevant is) en (3) uitlegbaarheid ten behoeve van de belanghebbende (die wil vaak met name kunnen begrijpen of hij het met de uitleg eens is).

5.2 Workshop 2

Aan de tweede workshop namen deel:

- Thomas Attema (Cryptoloog)
- Dennis Craandijk (AI lab nationale politie)
- Jeannine Florie (Eenheid Midden-Nederland)
- Anna Garghetti (Programma manager bij de politie)
- Nils Hulzebosch (Politie Amsterdam)
- Jelte Mense (Onderzoeker Data Science)
- Amir Niknam (Adviseur Innovatie bij Politie Nederland)
- Jeroen van Rest (Consultant TNO)

Centraal stond de toepassing van AI bij de digitale Perimeter Gezichtsherkenning, een vorm van niet-coöperatieve gezichtsherkenning waarbij de politie samenwerkt met de Johan Cruijff-Arena. Hoe kan de handreiking worden toegepast op deze casus, welke positieve punten volgen uit die toepassing en welke verbeterpunten zijn er voor de handreiking? Daaruit kwamen de volgende aandachtsgebieden naar voren:

- Het zou goed zijn om voor ‘techneuten’ een korter document te maken waarin zij in een oogopslag kunnen zien wat de belangrijkste uitgangspunten van het non-discriminatierecht is. Data-analisten zullen vaak geen groter document doorspitten voordat zij beginnen met een project.
- Het zou ook goed zijn om de principes voor de verschillende casus verder uit te werken. Bij de generieke uitgangspunten in de handreiking komt het uiteindelijk neer op de vraag: hoe dan, welke benchmarks gelden er, wie is verantwoordelijk, etc?
- Soms gaat het ene principe ten koste van de andere. Een model kan niet én optimaal transparant én optimaal effectief én optimaal non-discriminerend én.... zijn.

5.3 Workshop 3

Aan de derde workshop namen deel:

- Mark Bovens (UU; WRR)
- Rob van den Hoven van Genderen (VU)
- Roos van Leeuwen (BZK)
- Egge van der Poel (TIAS)
- Frank van Vonderen (VKA)
- Johan Wolswinkel (TiU)

Deze workshop had een algemeen karakter en was met name gericht op een discussie tussen experts met verschillende achtergronden. Daaruit kwamen de volgende aandachtsgebieden naar voren:

- Het is wellicht een idee om factsheets te maken: 1 a4 met kernpunten voor juristen; 1 a4 met kernpunten voor systeembouwers; etc.
- Het is wellicht een idee in het begin een verklarende woordenlijst te plaatsen.
- Een korte samenvatting zou prettig zijn en een duidelijke hiërarchie in de mate waarin principes en uitgangspunten belangrijk/essentieel zijn.
- Evaluatie is een continu proces. Het is dus niet een kwestie van afvinken en weer door, maar steeds teruggaan als er aanpassingen zijn gedaan: en hoe zit het dan nu met deze vragen?
- Soms zijn gevoelige gegevens juist nodig om te controleren op bias/discriminatie, natuurlijk wel beperkt tot wat noodzakelijk is.

6. Finalisering standaarden

Op basis van de workshops en eerste input van de klankbordleden van het Ministerie van Binnenlandse Zaken is een aangepaste versie van de handreiking gemaakt. Deze is vervolgens wederom besproken met de klankbord. Ook is het concept voorgelegd aan een interne klankbordgroep van het schrijfteam. Daarin namen zitting: Mark Bovens (WRR; UU), Francien Dechesne (LU), Ronald Leenes (TiU), Egge van der Poel (TIAS) en Johan Wolswinkel (TiU). Tot slot is het concept voorgelegd aan het College voor de Rechten van de Mens. Op basis daarvan heeft het team een definitieve versie van de handreiking gemaakt.

7. Annex

7.1 Enquête 1

7.1.1 Enquête

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, ...

Nee want, ...

Acceleratoren zijn: ...

Obstakels zijn:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft gesignaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

7.1.2 Survey

Survey

Name:

Organization:

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Which three examples of (current or future) AI use come to mind first?

Which three dangers / negative sides of AI are you most concerned about?

Which three opportunities / positive sides of AI come to mind first?

Do you think that AI applications and their use (for example by governmental organisations, companies and non-profit organizations) will increase in the near future? Why / why not - what are accelerators and what are obstacles?

1. Problem analysis

Yes because, ...

Accelerators are: ...

No because, ...

Obstacles are:

Different phases / steps play a role in the development and use of AI systems, such as: (1) the creation of a data set, (2) the development and training of an algorithm, (3) the production of patterns, models and profiles through data analysis, (4) making decisions / policies based on these models and (5) a feedback loop in which the results are ‘fed’ to the self-learning algorithm. [If you envisage a different categorisation/other phases, please indicate so]

When using AI, (unintended) discrimination can occur. What do you think are the underlying causes and can you give concrete examples of these? Where (which sectors, applications, etc.) do you foresee that the problem of discrimination in AI systems will occur in particular? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

Transparency of AI-based policy and decision making is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there with regard to transparency of AI processes and AI based policy- and decision-making? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

What do you think are other bottlenecks with regard to AI, for example technical, social or societal obstacles/problems? Can you give concrete examples of this? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

Could AI also help reduce / remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how? Please indicate as specifically as possible where in the process these benefits can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>
<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>

2. Existing Standards

2.1 Do you know of existing policy documents, standards, codes of conduct or guidelines with regard to AI? If so, which are most promising and why (please add source)?

Which principles or standards from those documents are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>

2.2 Are you aware of existing legislation or case law regarding AI? If so, which (please add source)?

Which principles or standards from that legislation or case law are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>

2.3 Are you aware of literature laying down additional standards / principles for AI? If so, which studies do you find most important and why (please add source)?

Which principles or standards from that literature are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>

2.4 Do you believe that these standards (identified in questions 2.1, 2.2 and 2.3) adequately address the problems you have identified (question group 1)? If not, what are the main shortcomings / which principles are missing? Please indicate as specifically as possible where in the process these additional principles could be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>

3. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☐☐

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain:

☐☐

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain:

☐☐

The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain:

- ☐ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain:

- ☐ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain:

- ☐ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain:

- ☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles

Explain:

- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

7.2 Enquête 2

7.2.1 Enquete

Enquête 2

Samen met collega's van diverse universiteiten doen we in opdracht van het Ministerie van Binnenlandse Zaken onderzoek naar de regulering van Kunstmatige Intelligentie/Artificial Intelligence. Voordat wij het zoveelste rapport over AI produceren willen wij eerst de klok gelijk zetten. Is er eigenlijk wel behoefte aan een nieuw document over dit onderwerp en zo ja, wat ontbreekt er dan aan de bestaande richtsnoeren? Zijn er bijvoorbeeld nieuwe of andere standaarden nodig of is er met name winst te behalen met meer en beter toezicht op de naleving van de bestaande regels?

Om inzicht op deze en andere punten te krijgen sturen we een korte enquête naar een aantal experts. We hopen dat jij bereid zou willen zijn de vragenlijst in te vullen; dat kost ongeveer 25 tot 35 minuten en je zou ons er enorm mee helpen!

Gaandeweg de enquête wordt meer verteld over ons onderzoek en hoe wij dat insteken; dat gebeurt stap voor stap, omdat wij willen voorkomen dat onze inkleuring meespeelt bij de beantwoording (gelieve dus niet vooruit lezen).

Als je liever anoniem wilt blijven (de enquêtes worden mogelijk als bijlage bij een openbaar rapport gevoegd) laat dan de eerste pagina leeg.

Als dat binnen de mogelijkheden ligt zouden we willen vragen om het document voor 22 augustus toe te sturen aan Bart van der Sloot (b.vdrsloot@uvt.nl) terug te sturen, maar eerder mag natuurlijk ook.

Naam:
Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificiële Intelligence (AI) definiëren?

Wat is uw functie/rol (b.v. data scientist, product owner, project manager, data-analist, data engineer, machine learning engineer, manager, onderzoeker, leidinggevende)?

In welke industrie bent u werkzaam (b.v. gezondheidszorg, financiën, energie, online retail, wetenschap)?

Binnen welke AI-domein(en) vinden uw werkzaamheden met name plaats (b.v. computer vision, natural language processing, recommender systems, (geautomatiseerde) besluitvorming)

Bij de ontwikkeling of het gebruik van welk soort AI-applicaties bent u betrokken (b.v. fraudedetectie, vraagvoorspelling, customer service)?

1. Probleemanalyse

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden.

Is het optreden van discriminatie door het gebruik van AI **in het algemeen** volgens u:

Ja Nee

☐ ☐ Een **belangrijk** probleem
Leg uit:

☐ ☐ Een **urgent** probleem
Leg uit:

☐ ☐ Een **veelvoorkomend** probleem
Leg uit:

1.2. Is het optreden van discriminatie door het gebruik van AI in **uw werkdomein** volgens u:

Ja Nee

☐ ☐ Een **belangrijk** probleem
Leg uit:

☐ ☐ Een **urgent** probleem
Leg uit:

☐ ☐ Een **veelvoorkomend** probleem
Leg uit:

1.3 In welke AI toepassingen (sectoren/domeinen) verwacht u dat ongewenste discriminatie/ongelijke behandeling zich met name zal voordoen?

Ja Nee

☐ ☐ Gezondheidszorg
Leg uit:

☐ ☐ Welzijn
Leg uit:

☐ ☐ Politie en justitie
Leg uit:

☐ ☐ Onderwijs
Leg uit:

Ja Nee

☐ ☐ Huisvesting
Leg uit:

☐ ☐ Retail
Leg uit:

☐ ☐ Arbeidsmarkt
Leg uit:

☐ ☐ Anders, namelijk...
Leg uit:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) Probleemdefinitie (identificeren bedrijfsprobleem en vertaling naar ML taak), (2) Initiële dataverzameling en analyse, (3) Datavoorbereiding, (4) Modelleren (verschillende modellen uitproberen en een model selecteren), (5) Evaluatie (geselecteerde model evalueren aan de hand van succes criteria vastgesteld in stap 1 en een ‘test set’) en (6) Implementatie (inclusief monitoring/maintenance activiteiten en terugkoppeling over performance in gebruik). [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

1.4 Waar in het proces zijn volgens u de oorzaken van ongewenste discriminatie te situeren? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzameli ng	(3) datavoorberei ding	(4) modelleren	(5) evaluatie	(6) implementatie
Problemen:	Problemen:	Problemen:	Problemen:	Problemen:	Problemen:
Voorbeelden:	Voorbeelden:	Voorbeelden:	Voorbeelden:	Voorbeelden:	Voorbeelden:
Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:

1.5 Transparantie van AI-besluiten is een *best practice*; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzameli ng	(3) datavoorberei ding	(4) modelleren	(5) evaluatie	(6) implementatie
Problemen:	Problemen:	Problemen:	Problemen:	Problemen:	Problemen:
Voorbeelden:	Voorbeelden:	Voorbeelden:	Voorbeelden:	Voorbeelden:	Voorbeelden:
Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:

1.6 Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe?

2. Mogelijke aanpakken

Verschillende organisaties en bedrijven hebben principes en standaarden ontwikkeld om discriminatie door AI systemen te minimaliseren. Data scientists hebben het hierbij vaak over het reduceren van bias en het zorgdragen voor *fairness*.

2.1 Neemt u of uw organisatie zorgen omtrent discriminatie momenteel mee in keuzes en beslissingen aangaande het gebruik of de inrichting van AI? Zo ja, ten aanzien van welke fase(s)? Leg aub uit.

(1) <i>probleem definitie</i>	(2) <i>dataverzame ling</i>	(3) <i>datavoorber eiding</i>	(4) <i>modelleren</i>	(5) <i>evaluatie</i>	(6) <i>implementati e</i>

2.2 Zo ja, maakt u of uw organisatie gebruik van specifieke standaarden/principes/richtlijnen (voeg aub bron toe)?

2.3 Welke activiteiten/technieken gebruikt u of uw organisatie om mogelijke pijnpunten in het kader van ongelijke behandeling/discriminatie **te identificeren**?

Ja Nee

- ☐ ☐ Formele risicoanalyse
Leg uit:
- ☐ ☐ Checklists
Leg uit:
- ☐ ☐ Toolkits
Leg uit:
- ☐ ☐ In gesprek met gebruikers
Leg uit:
- ☐ ☐ In gesprek met andere betrokkenen
Leg uit:
- ☐ ☐ Anders, namelijk...
Leg uit:

2.4 Welke activiteiten/technieken gebruikt u of uw organisatie om discriminatie/ongelijke behandeling **te meten**?

Ja Nee

- ☐ ☐ **Data op biases** checken aan de hand van fairness criteria. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:
- ☐ ☐ Model evalueren aan de hand van formele fairness criteria bij **model selectie**. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:
- ☐ ☐ Model evalueren aan de hand van formele fairness criteria bij **model in productie**. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:
- ☐ ☐ Actieve evaluatie ten behoeve van discriminatie/ongelijke behandeling met Gebruikers; Zo ja, *leg uit*....
Domeinexpert; Zo ja, *leg uit*....
Andere betrokkenen; Zo ja, *leg uit*
- ☐ ☐ Anders, namelijk...
Leg uit:

2.5 Welke activiteiten/technieken worden in uw team gebruikt om discriminatie/ongelijke behandeling **te reduceren**?

Ja Nee

- ☐ ☐ Meer data verzamelen
Leg uit:
- ☐ ☐ Analyse van selectie variabelen
Leg uit:
- ☐ ☐ Data aanpassen/filteren (pre-processing)
Leg uit:
- ☐ ☐ Speciale technieken om unfairness te reduceren tijdens het leren (in-processing)
Leg uit:
- ☐ ☐ Speciale technieken om bias te mitigeren van predictions (post processing)
Leg uit:
- ☐ ☐ Anders, namelijk ...
Leg uit:

3. Juridisch raamwerk

3.1 Bent u op de hoogte van bestaande wetgeving, richtsnoeren of Codes of Conducts ten aanzien van AI? Zo ja, welke en welke spelen specifiek een rol bij de ontwikkeling of het gebruik van AI systemen binnen uw organisatie als het gaat om discriminatie?

3.2 Op welke moment komt deze wetgeving aan bod in de ontwikkeling of het gebruik van AI systemen binnen uw organisatie en op welke manier?

3.3 Zijn er in het verleden knelpunten geweest bij het operationaliseren van de wetgeving in de ontwikkeling, implementatie en gebruik van AI systemen binnen uw organisatie en hoe heeft uw organisatie (of is zij voornemens) deze knelpunten te adresseren?

3.4 Wat zou u idealiter van een wet of code of conduct verwachten ten aanzien van het bestrijden van discriminatie en ongelijke behandeling in AI-systemen?

4. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analisten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analisten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analisten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis,

interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:

☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:

☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:

☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.2.2 Survey

Survey 2

Together with colleagues from various universities, we are conducting a research project on the regulation of Artificial Intelligence on behalf of the Dutch government. Before we produce yet another report on AI, we want to do a reality check. Is there really a need for yet another document on this subject and if so, what is missing in the existing guidelines? For example, are new or different standards required; is more and better supervision and enforcement of the existing rules for AI needed?

To gain insight into these and other questions, we will send a short survey to a number of experts, such as you. Would you be willing to complete the questionnaire? It takes about 25 to 35 minutes and you would help us a lot.

More is told in the survey about the research project we are conducting and how we conduct it; this is done step by step, so as to prevent as much as possible our choices and perceptions from playing a role in your answers (so please don't read ahead).

If you prefer to remain anonymous (the surveys might be included in an annex to a public report), leave the first page blank.

If possible, we would like to ask you to send the document to Bart van der Sloot (b.vdrsloot@uvt.nl) before 22 August, though sooner, as always, is better.

Name:

Organization:

0. General

How would you define Artificial Intelligence (AI) in one sentence?

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI in general:

Yes No

☐ ☐ An important problem
Explain:

☐ ☐ An urgent problem
Explain:

☐ ☐ A common problem
Explain:

1.2. The occurrence of discrimination through the use of AI in your work domain is:

Yes No

☐ ☐ An important problem
Explain:

☐ ☐ An urgent problem
Explain:

☐ ☐ A common problem
Explain:

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes No

☐ ☐ Healthcare
Explain:

☐ ☐ Wellness
Explain:

☐ ☐ Policing/criminal
law
Explain:

☐ ☐ Education
Explain:

Yes No

☐ ☐ Housing
Explain:

☐ ☐ Retail
Explain:

☐ ☐ Labourmarket/employment
Explain:

☐ ☐ Otherwise, namely
Explain:

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

2.3 What activities / techniques do you or your organization deploy to identify potential issues of unequal treatment / discrimination?

Yes No

☐ ☐ Formal Risk Assessment
Explain:

☐ ☐ Checklists
Explain:

☐ ☐ Toolkits
Explain:

☐ ☐ Conversations with users
Explain:

☐ ☐ Conversations with other parties involved
Explain:

☐ ☐ Otherwise, namely...
Explain:

2.4 What activities / techniques do you or your organization deploy to measure discrimination / unequal treatment?

Yes No

- ☐ ☐ Checking for biases in the data based on fairness criteria. If so, which criteria do you use?
Explain:
- ☐ ☐ Evaluating the model based on formal fairness criteria for model selection. If so, which criteria do you use?
Explain:
- ☐ ☐ Evaluating the model based on formal fairness criteria for the model in production. If so, which criteria do you use?
Explain:
- ☐ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain....
- Other people involved; If so, explain....
- ☐ ☐ Otherwise, namely...
Explain:

2.5 What activities / techniques are used in your team to reduce discrimination / unequal treatment?

Yes No

- ☐ ☐ Collect more data
Explain:
- ☐ ☐ Analysis of selection variables
Explain:
- ☐ ☐ Adjusting / filtering data (pre-processing)
Explain:
- ☐ ☐ Special techniques to reduce unfairness during learning (in-processing)
Explain:
- ☐ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☐☐

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain:

☐☐

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain:

☐☐

The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain:

- ☐ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain:

- ☐ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain:

.....

- ☐ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain:

.....

- ☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles

Explain:

.....

- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

7.3 Ingevulde enquêtes (enquête 1)

7.3.1 Respondent 1

Naam: Paul De Hert & Bram Visser

Organisatie: Vrije Universiteit Brussel, Onderzoeksgroep LSTS

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Artificiële intelligentie is het vermogen van een computer of softwareprogramma om zijn gedrag aan te passen teneinde een vooraf bepaald doel te bereiken.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

*Het detecteren van credit card fraude.
GPS systemen die obv de huidige verkeerssituatie het verkeer omleiden.
Gepersonaliseerde advertenties/content.*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Gebrek aan transparantie vanwege de complexiteit van de berekeningen en de hoeveelheid aan data.
Het reproduceren van biases in de dataset op geaggregeerd niveau.
Aansprakelijkheid ingeval het algoritme een foute beslissing/aanbeveling maakt.*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

*Door de hoeveelheid aan data die verwerkt kan worden zullen de aanbevelingen/beslissingen die daaruit volgen doorgaans zeer accuraat zijn.
De beslissingen/aanbevelingen gebeuren zeer snel: in real-time en zelfs voorspellend.
Door de hoge efficiëntie ervan kunnen er grote kostenbesparingen gedaan worden.*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ik denk dat die hoge vlucht al een tijdje bezig is om de simpele reden dat het past in een economische logica van het optimaliseren van beslissingen.

Mogelijke obstakels zijn sectorafhankelijke toepassingen die het momenteel te duur maken om te implementeren. Hoe complexer het vereiste algoritme, hoe duurder.

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Discriminatie kan optreden wanneer AI gebruikt wordt door de politie om mensen een risicoscore toe te kennen. Als er in de oorspronkelijke criminele data waarop het algoritme getraind wordt reeds een deel van de bevolking onterecht disproportioneel vertegenwoordigd is, zal deze discriminatie alleen maar vergroot worden wanneer men dit op grote schaal zal toepassen. Het wordt in dit geval duidelijk in de besluit- en beleidsvorming, terwijl de oorzaak ervan zich in de het aanleggen van dataset bevindt, waarop het algoritme wordt getraind en de patronen worden gedestilleerd.

Het verdient hierbij ook vermelding dat net een zelflerend algoritme hiervoor ook een oplossing kan bieden. Wanneer blijkt dat de voorgestelde beslissingen/aanbevelingen niet lijken te kloppen met de realiteit (te veel valse positieven), kan het algoritme dit opnemen in de feedbackloop en op basis daarvan de data, het algoritme, de patronen en de besluitvorming aanpassen.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van

transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

De complexiteit van de berekeningen maakt het soms niet mogelijk om na te gaan welke data precies in acht werd genomen. Ook vormen intellectuele eigendomsrechten en bedrijfsgeheimen obstakels voor transparantie. Dit probleem situeert zich dan voornamelijk op het niveau van ontwikkelen en trainen van het algoritme.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

De nauwkeurigheid van de uitkomst van het algoritme dat zich vertaalt in de besluit- en beleidsvorming is van cruciaal belang als het gaat om aangelegenheden die een impact hebben op het leven van mensen. Denk bijvoorbeeld aan vrijheidsberoving; hierbij is het cruciaal dat het algoritme een juiste uitkomst genereert. Het probleem van de nauwkeurigheid leunt dicht aan tegen het probleem van transparantie hoewel het mogelijk is om de nauwkeurigheid van een systeem te testen zonder dat systeem transparant te maken.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Zoals reeds vermeld onder 1.1 kan een zelflerend algoritme per zaak zijn database updaten wat de nauwkeurigheid van het algoritme bevordert en zodoende ongelijkheden wegwerkt.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Voornamelijk op internationaal vlak ben ik bekend met [OECD Principles on AI](#), en in mindere mate met de Europese communicaties omtrent een '[human-centric](#)' AI. Op nationaal vlak zijn mij zulke standaarden niet bekend. Beide standaarden benadrukken ongeveer dezelfde zaken.

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Technical robustness and safety (EU and OECD)	Transparency (EU and OECD)	Diversity, non-discrimination and fairness (EU)	Accountability (EU and OECD)	
Privacy and data governance (EU)			Societal and environmental well-being (EU) Respect the rule of law (OECD)	
			Human agency and oversight (EU)	

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

De GDPR geeft datasubjects meer controle over hun data waardoor ze zelf kunnen beslissen welke data ze afstaan aan welke bedrijven en zodoende het trainen van die algoritmes kunnen bepalen.

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
GDPR	GDPR			

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Van Brakel (2020) stelt in [Een reflectie over het huidige toezicht van het gebruik van surveillancetechnologie door de lokale politie in België](#) voor om, naast andere aanbevelingen die voornamelijk tot doel hebben het democratisch toezicht en transparantie te bevorderen, een Surveillance Impact Assessment (naar het model van een DPIA) in te voeren in het kader van de implementatie van surveillance technologie door de politie teneinde een toetsing uit te voeren die breder gaat dan louter databescherming maar dus ook kijkt naar de bredere maatschappelijke impact van zulke implementatie. Voorts ziet zij voor specifiek het probleem van discriminatie aan te pakken een rol voor de interfederale instelling UNIA weggelegd.

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
			Toezicht en transparantie	

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
			Transparantie wordt overal erkend als een belangrijk onderdeel maar het is duidelijk dat dit niet eenvoudig is dit te bereiken. Een goed idee zou zijn om een bepaalde mate van transparantie minimaal te garanderen afhankelijk van het toepassingsdoemein (bijvoorbeeld een zeer grote mate van transparantie en nauwkeurigheid als het gaat om predictive policing algoritmes.)	Zoals reeds aangehaald kan de feedbackloop na verloop van tijd discriminatie in de data wegwerken. Het grote nadeel hiervan is echter wel dat de discriminatie in de oorspronkelijke data eerst voor een bepaalde periode zal moeten gekopieerd worden om ze te kunnen wegwerken.
			Er lijkt een consensus te zijn dat de aansprakelijkheid bij ontwikkelaars ligt	

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: Ja, ik denk dat deze mensen voornamelijk bezig zijn met optimaliseren van hun product en niet met de bredere maatschappelijke implicaties ervan.

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: Ja, niet-bindende documenten die het werk van ontwikkelaars enkel bemoeilijken motiveert hen niet om in acht te nemen.

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: Neen, ik denk dat het zorgen maken om de algemeenheid ervan pas in een later stadium optreedt vanaf er zekerheid is over het bekend zijn van de problematiek bij de ontwikkelaars.

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: Neen, er zijn genoeg context-specifieke voorbeelden.

☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Ja, er wordt veel en uiteenlopend over geschreven wat een duidelijk overzicht bemoeilijkt. Echter, ik ben niet zeker of ze in de eerste plaats weet hebben van alles wat er over geschreven wordt.

☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Ja

☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Ja, het zijn dan ook slechts standaarden en principes, geen bindende regelgeving.

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.3.2 *Respondent 2*

Name: Diana Dimitrova

Organization: FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Technologies which process, including analyse, a huge amount of data according to pre-set rules and produce output without human intervention, and which have also the ability to learn independently with time.

Which three examples of (current or future) AI use come to mind first?

Passenger Name Record (PNR), biometric identification/verification, research data analysis.

Which three dangers / negative sides of AI are you most concerned about?

Lack of transparency of decision-making processes and results; results are difficult to challenge and reverse; too high trust in AI per se (not only in particular applications).

Which three opportunities / positive sides of AI come to mind first?

Processing speed, new insights, possibility to process high volumes of data.

Do you think that AI applications and their use (for example by governmental organisations, companies and non-profit organizations) will increase in the near future? Why / why not - what are accelerators and what are obstacles?

Yes because, AI technologies are quick to impose themselves on the market, new laws have in-built AI processing elements (see e.g. PNR) and the broader public is not aware of the risks and thus not opposing its deployment.

Accelerators are: AI developers and their marketing tactics; law-makers who see technology as a universal solution.

1. Problem analysis

No because, ...

Obstacles are:

Different phases / steps play a role in the development and use of AI systems, such as: (1) the creation of a data set, (2) the development and training of an algorithm, (3) the production of patterns, models and profiles through data analysis, (4) making decisions / policies based on these models and (5) a feedback loop in which the results are ‘fed’ to the self-learning algorithm. [If you envisage a different categorisation/other phases, please indicate so]

When using AI, (unintended) discrimination can occur. What do you think are the underlying causes and can you give concrete examples of these? Where (which sectors, applications, etc.) do you foresee that the problem of discrimination in AI systems will occur in particular? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: lack of accuracy; negative impact on the rights of certain groups of people</i>	<i>Problems: AI does not achieve the purposes for which it is set up.</i>	<i>Problems: negative impact on individuals who have been wrongly placed in a certain category.</i>	<i>Problems: impact on individuals and society at large.</i>	<i>Problems: lock-in effect of the initial discriminatory algorithm construction</i>
<i>Examples: biometric algorithms not using data from people coming from different parts of the world</i>	<i>Examples: in law enforcement the behaviour of suspects/criminals might change over time and the initial algorithm might not detect them, while casting suspicion on innocent persons.</i>	<i>Examples: in risk profiling, one might have barely passed the threshold of a certain category of risk (and be innocent), but will still fall within the (in casu) wrong category.</i>	<i>Examples: individuals being denied credit, etc, or research results being wrong, which could have negative impact on science (e.g. medicine), quality of the laws, etc.</i>	<i>Examples: ...</i>
<i>Causes: data sets might not be representative</i>	<i>Causes: the logic of the algorithm may be discriminatory if it does not objectively reflect the problem it seeks to solve or if the problem</i>	<i>Causes: the patterns might be inaccurate or might be too “stiff”, i.e. produce categories in neither of which some</i>	<i>Causes: the resulting decisions will not be accurate and may prejudice both individuals and the quality of the whole</i>	<i>Causes: the initial mistakes do not get corrected and replicate themselves.</i>

	<i>changes over time.</i>	<i>people/events will actually fit, but will have to be anyway placed into one.</i>	<i>decision/policy-making process.</i>	

Transparency of AI-based policy and decision making is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there with regard to transparency of AI processes and AI based policy- and decision-making? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: the decision-makers cannot themselves know/understand how the algorithm works.</i>	<i>Problems: no possibility to challenge decisions by affected persons.</i>	<i>Problems: biases are replicated.</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: treated as a commercial secret.</i>	<i>Causes: treated as a commercial secret.</i>	<i>Causes: the distilled patterns are not challenged by the developers/decision-makers and are also unexplainable to them.</i>	<i>Causes: the results cannot be explained to the concerned individuals either because they are not understood by the decision-makers or these do not want to be transparent,</i>	<i>Causes: ...</i>

			<i>e.g. by evoking trade secrets and similar.</i>	

What do you think are other bottlenecks with regard to AI, for example technical, social or societal obstacles/problems? Can you give concrete examples of this? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: Sometimes the problem to be solved is not clearly formulated and the data set is thus not selected on the basis of objective criteria.</i>	<i>Problems: ...</i>	<i>Problems: technical complexity which hinders human understanding. Humans do not dare challenge the patterns and modify them.</i>	<i>Problems: decisions are not transparent because of lack of human understanding, also because of lack of clear explanation obligations in law all cases where AI technologies are involved, and lack of clarity how to be transparent without breaching IP rights, which gives room for abuse of these rights by the controller. Another problem is that knowledge (about different things in life) might be taken away from humans and be transferred to the machines. In addition, it is not clear how data subject rights apply in different AI contexts.</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

Could AI also help reduce / remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how? Please indicate as specifically as possible where in the process these benefits can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Existing problems: Problems not studied on a bigger scale</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>
<i>AI-driven solutions: The capacity of AI to process more data. However, this depends on whether the quality of the data is better, as more quantity does not mean better quality.</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>

2. Existing Standards

2.1 Do you know of existing policy documents, standards, codes of conduct or guidelines with regard to AI? If so, which are most promising and why (please add source)?

-
1. Article 29 Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (WP251rev.01, 6 February 2018).
 2. ICO, 'Explaining decisions made with AI' <https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/>.
 3. German Data Ethics Commission, 'Opinion of the Data Ethics Commission' https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_la ng.pdf?__blob=publicationFile&v=3.

Which principles or standards from those documents are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
Quality requirements	Quality requirements	Quality requirements	Explainability of individual decisions;	
			How other data subjects rights apply (partial guidance);	
			overall human agency, dignity, privacy, democracy and rule of law requirements; respect for fundamental rights.	

2.2 Are you aware of existing legislation or case law regarding AI? If so, which (please add source)?

Council of Europe Convention 108+, Article 9
Opinion 1/15 of the Court (Grand Chamber) [2017], ECLI:EU:C:2017:592 (EU PNR Canada);
 Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime [2016] OJ L119/132

Which principles or standards from that legislation or case law are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
	Quality standards in relation to the abstract algorithms and necessity for independent supervision		Understandability and explainability of decisions	

2.3 Are you aware of literature laying down additional standards / principles for AI? If so, which studies do you find most important and why (please add source)?

Council of Europe, 'Consultative Committee of the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (Convention 108) Report on Artificial Intelligence' <https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6>.
 EDPB, 'New Rules for Credit Reporting Systems in the Digital Economy' (referring to Italy's IL Garante guidance) https://edpb.europa.eu/news/national-news/2019/new-rules-credit-reporting-systems-digital-economy_de.

Which principles or standards from that literature are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
			explainability	
			Human dignity and agency	
			Self-determination	

2.4 Do you believe that these standards (identified in questions 2.1, 2.2 and 2.3) adequately address the problems you have identified (question group 1)? If not, what are the main shortcomings / which principles are missing? Please indicate as specifically as possible where in the process these additional principles could be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
Concrete and comprehensive quality standards	Concrete and comprehensive quality standards	Concrete and comprehensive quality standards	How safeguards, such as data subjects rights, apply or should be interpreted in light of profiling and AI decision-making.	Concrete and comprehensive quality standards
			Clarification that these standards should apply also when a decision is only partially automated.	
			Impact assessment obligations of AI technologies as a whole on fundamental	

			rights, rule of law and democracy. More public debates as to whether specific AI technologies are needed and if yes - under what conditions.	

3. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☒ ☐ The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: Law/ethics is not considered to be important in the development and technical professionals sometimes do not receive proper training in these fields.

☒ ☐ The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: In addition, law and guidelines are rarely clear enough and do not give technical professionals concrete guidance, which discourages them.

☒ ☐ The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain: see previous answer.

- ☒ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI
Explain: Technical professionals need to be accompanied at every step by a legal/ethical expert. However, the latter sometimes do not have sufficient understanding of the technology. More inter-disciplinary courses/training are needed. Sometimes the guidance by ethicists/lawyers runs against the market interests of software/technology developers and thus legal/ethical advice is sidelined.
- ☒ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)
Explain: These standards could differ also because of specific national provisions across the EU. There is also no overall agreement between legal/ethics professionals about how to apply law and ethics to new technologies. However, this might be due not only to legal issues (how to interpret laws), but also to commercial interests.
- ☐ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice
Explain: Partially agree. The adoption in practice will depend also on whether developers/engineers will have an incentive to follow them.
- ☒ ☐ There is a lack of supervision and enforcement of the existing standards and principles
Explain: and consistency within the EU about this enforcement. Probably DPAs lack resources and technical knowledge...
- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

I would suggest paying attention to the issues also in the law enforcement field, which is discussed less often than other applications. You could also emphasize better the necessity to link the technology (AI) and the underlying problem it is trying to solve. Finally, continual assessments of the impact of the technology and possibilities to stop their operation are essential.

7.3.3 Respondent 3

Enquête

Naam: Marlies van Eck

Organisatie: Hooghiemstra & Partners en Universiteit Leiden (dit wordt per 1-9 Radboud Universiteit)

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Ik begrijp meer dan gemiddeld wat AI is maar altijd vanaf de buitenkant.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Risico-classificering van burgers/klanten/patienten/gebruikers/mensen
Google Maps
Zoekmachines voor vinden van relevante informatie

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Dat mensen gaan geloven in de 'waarheid' zoals die gepresenteerd wordt door de systemen
Dat mensen denken dat ze het zelf beter weten en specifieke kennis die in ontwerp wel is meegenomen, negeren
Dat we nog niet weten wat de impact is.

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Beter geïnformeerde beslissingen
Sneller de juiste en relevante informatie vinden
Verlost zijn van repeterende en saaie taken

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, ...

Nee want, de achterliggende systemen en legacyproblemen zijn veel te groot

Acceleratoren zijn: ...

Obstakels zijn: zie hierboven. Ik vermoed dat het eerst een afzwakking zal kennen om daarna wel een vlucht te nemen.

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van

transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde

beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit:

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit:

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

Ja maar ik heb geen tijd om het op te schrijven, excuus! Mijn enige tip is: als je een handreiking schrijft doe dit dan met meer disciplines en vooral met AI deskundigen

7.3.4 Respondent 4

Enquête

Naam: Linda van de Fliert

Organisatie: Gemeente Amsterdam

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Software die geautomatiseerd voorspellingen doet, besluiten neemt of adviezen geeft door gebruik van data-analyse, statistiek of zelflerende logica, die leiden tot directe of indirecte impact op burgers, bedrijven of fysieke assets in de buitenruimte

> Dit is de definitie die we met experts hebben geformuleerd voor de 'inkoopvoorwaarden voor verantwoord gebruik van algoritmes' die we hebben opgesteld en in de samenwerking met verschillende overheden, waarmee we instrumenten voor publieke controle op algoritmes en AI willen ontwikkelen

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

-
- Beeldherkenning en automatische koppeling, bijvoorbeeld door parkeercontrole-auto's of milieuzonecamera's;*
 - Prioriteren van data, bijvoorbeeld meldingen van [vakantieverhuurfraude](#);*
 - Patronen of gelijkenissen herkennen in grote hoeveelheden data, bijvoorbeeld het meldingenalgoritme van Amsterdam ([zie uitleg](#))*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

-
- Onbedoelde vooroordelen die altijd al onderdeel van menselijke keuzes waren vastleggen in een systeem*
 - Fouten of onvolkomenheden in data die leiden tot verkeerde conclusies en vooral degene treffen waar de meeste data over beschikbaar is*
 - AI gebruiken doe je om te optimaliseren; maar waarvoor optimaliseer je? In geval van de overheid bestaat het te automatiseren werk toch vooral uit reguleren en voorkomen; boetes uitdelen, handhaving, etc.*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

-
- *Efficiënter en effectieve maken van processen*
 - *In sommige gevallen: besluiten objectiever maken als die objectief moeten en kunnen zijn (voorbeeld parkeercontrole: een parkeerbeamte zou het merk van de auto of een woordenwisseling met de bestuurder kunnen meewegen in het besluit een boete uit te schrijven)*
 - *Vooroordelen in menselijke besluitvorming komen aan het licht / worden expliciet gemaakt als ze in AI geprogrammeerd worden*
-

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, in ieder geval beeld- en stemherkenning; je ziet dat dat nu uit de fase van experiment is en bewezen technologie is. Je kan bijv. je creditcard scannen zodat je niet meer je gegevens hoeft in te vullen online.

Overigens denk ik niet dat de technologie per se veel geavanceerder gaat worden komende tijd (kan ik niet voorspellen), maar ik voorspel wel dat het, in de huidige staat, veel meer alom toegepast gaat worden

Acceleratoren zijn: grote bedrijven investeren er nu veel in (o.a. spraakpalen in huis)

Nee want, m.n. beeldherkenning zou best geremd kunnen worden, door zorgen over gezichtsherkenning. Het valt me op dat er ineens echt een maatschappelijk debat over privacy is; dat had ik 5 jaar geleden niet verwacht.

Obstakels zijn: mogelijke schandalen met grootschalige privacyschending of afnemen van vrijheden na gezichtsherkenning, ook als dat evt. in andere landen is. Vervolgens zal er wet- en regelgeving komen die (in principe terecht) remmend werkt

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: Vertekend beeld door oververtegenwoordiging van groep</i>	<i>Problemen: er worden bij het ontwikkelen aannames gedaan en keuzes gemaakt die op dat moment legitiem zijn; bijvoorbeeld trainen met een onvolledige dataset. Maar dat wordt nergens vastgelegd waardoor 5 jaar later niet meer achterhaald kan worden waarom het algoritme werkt zoals het werkt</i>	<i>Problemen: AI die patronen destilleert vindt meestal correlaties. Dat er geen causaal verband onder het patroon ligt is prima als gaat om zoiets als een logistiek proces efficiënter inrichten. Maar het is niet prima als het patroon leidt tot besluitvorming die impact heeft op mensenlevens. Dan moet er een causaal verband aan te tonen zijn.</i>	<i>Problemen: ...</i>	<i>Problemen: Er is vaak geen feedbackloop. Beleidskeuzes zijn altijd een tijdsopname van bepaalde politieke voorkeuren en keuzes op dat moment. Als je dat niet regelmatig evalueert, dan verandert het algoritme dus niet mee met de ontwikkelingen binnen een organisatie en de politieke kleur van het bestuur</i>
<i>Voorbeelden: Ik ken het voorbeeld niet precies maar in A'dam is een bepaalde vluchtelingengroep (ik meen Eritreeërs)</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: Maranke Wieringa (PhD UU) vertelde me dit voorbeeld; zij kent het beter dan ik: In R'dam</i>

een tijd lang gemonitord – destijds met legitieme redenen – en daardoor is die groep en mensen met vergelijkbare kenmerken als die groep oververtegenwoordigd in de data en alle volgende toepassingen waar de set bij wordt gebruikt				konden mensen aanspraak maken op de schuldhelpverlening. Een algoritme herkent of mensen voldoen aan de eisen. Als dat niet het geval is of bijvoorbeeld te laat betaald wordt, worden ze automatisch uit het programma gezet (overigens ook als per ongeluk een typfout gemaakt wordt). Dat was logisch toen er een tijd terug een vrij rechtlijnig bestuur was. In het huidige politieke klimaat is dat veel minder logisch en zouden andere keuzes gemaakt worden. Toch is het systeem nooit aangepast omdat dat gewoon niet onderdeel was van het proces
Oorzaken: mensen zien data als objectief en feitelijk, terwijl het juist altijd een abstractie van de werkelijkheid is, waar ook dingen weggelaten of extra uitgelicht worden (bewust of onbewust)	Oorzaken: ...	Oorzaken: ook hier geldt: mensen zien data als objectief en feitelijk, terwijl het juist altijd een abstractie van de werkelijkheid	Oorzaken: ...	Oorzaken: algoritmes zitten nog in de sfeer van nieuwe technologie (al gebruiken we al jaren rekenmodellen). Dat betekent dat er nog veel ‘gepilot’ wordt en

		<i>is, waar ook dingen weggelaten of extra</i>		<i>dat er dus nog geen eenduidig kader is binnen organisaties; zoals dat er wel voor bijvoorbeeld financiële controle is</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: datasets bevatten vaak deels niet-openbare informatie waardoor het openbaarmaken van wel openbare informatie een extra handeling vereist.</i>	<i>Problemen: veel ontwikkelaars leggen niet goed vast welke keuzes ze maken en waarom. Welke risico's heeft het algoritme, hoe wordt daarmee omgegaan? Is er een discussie over geweest met het team? Bestuur? Raad? Of alleen keuze van ontwikkelaar? Was de trainingsdata incompleet? Al die dingen worden vaak nergens opgeschreven, waardoor uiteindelijk moeilijk te achterhalen is hoe het algoritme</i>	<i>Problemen: Mensen doen hier vaak heel moeilijk over, dat je niet kán laten zien hoe het algoritme werkt, zeker bij black box algoritmes. Terwijl als je alle aannames, keuzes en uitgangspunten maar goed vastlegt (zie antwoord hier links naast), dan</i>	<i>Problemen: soms is er weldegelijk een democratische discussie (in de Raad of de Tweede Kamer) vooraf gegaan aan ontwikkeling van een algoritme, of zijn er andere democratische waarborgen geweest die ook wel ergens vastliggen. In principe is er dus transparantie, want veel van die documenten worden actief openbaar gemaakt (bijv. website tweede kamer). Maar voor een burger, journalist of politicus is bijna niet te achterhalen welke besluitvormingsdocu</i>	<i>Problemen: ...</i>

	<i>werkt en waarop het zijn keuzes/adviezen/besluiten baseert. En dat maakt uitlegbaarheid van algoritmes bijna onmogelijk</i>	<i>is goed te achterhalen hoe het patroon wordt samengesteld.</i>	<i>menten bij welk algoritme horen. Laat staan dat hij/zij op bij een WOBverzoek zou weten waar hij naar moet vragen.</i>	
<i>Voorbeelden : ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden : ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden : ...</i>
<i>Oorzaken: er wordt bij het aanleggen van de dataset geen onderscheid gemaakt tussen openbare en niet-openbare informatie</i>	<i>Oorzaken: kost tijd en moeite, en er is ook niet altijd een eenduidig systeem om dit in te doen.</i>	<i>Oorzaken: men laat zich afschrikken door techniek. Techniek is vaak prima uit te leggen; en zo niet, dan moet de techniek zich aanpassen, niet de mens. We hebben als overheid de plicht doelmatig en rechtmatig te werken en moeten kunnen laten zien dat we dat gedaan hebben.</i>	<i>Oorzaken: het zit nog niet in ons systeem om gestructureerd informatie over algoritmes te verzamelen.</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

	<i>trainen algoritme</i>			
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Zie mijn eerdere voorbeeld van parkeercontroleauto's. Alleen als de regels volledig objectief uitgevoerd kunnen worden – maar dat door een mensen niet objectief uitgevoerd worden – is er sprake van verminderen/verhelpen van discriminatie e.d.

In alle andere gevallen ben je met AI juist aan het proberen om menselijke afwegingen, inschattingen, keuzes, etc. te automatiseren en optimaliseren.

Dat is overigens niet per definitie iets erg: onze menselijke oordelen en vooroordelen gebruiken we nu ook, om situaties in te schatten en besluiten te nemen. Alleen zodra je dat soort afwegingen als regels vastlegt in een algoritme, moet je wel zeker weten dat ze leiden tot het gewenste resultaat.

Wel zou je AI op metaniveau kunnen gebruiken om patronen van discriminatie en bias te herkennen (met daarna dus een check op causaliteit).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Zie hieronder (past niet in dit tekstvak). Het is veel, maar dit zijn de bronnen die wij hebben gebruikt bij het uitwerken van de instrumenten voor publieke controle op algoritmes (eerste vier bullits). Ik ben met name erg blij met / trots op de inkoopvoorwaarden omdat het de richtlijnen van EC en Rijksoverheid operationaliseert tot werkbare definities en afspraken

Brondocumenten

Tools om een algoritme mee te beoordelen:

Inkoopvoorwaarden (modelcontracten) voor inkoop van algoritmes:

<https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/grip-op-algoritmes/> - nu nog een Amsterdams voorbeeld, binnenkort samen met VNG, gemeenten, provincies, waterschappen, Politie en Rijkswaterstaat doorontwikkelen

Algoritmeregister Amsterdam – register met gebruiksvriendelijke (met burgers geteste) front end, waarin burgers/journalisten/raadsleden/etc. kunnen zien welke algoritmes gebruikt worden. Niet alleen uitleg van de algoritmes en hun werking maar ook technische info zoals datasets, source code en modelarchitectuur: binnenkort op <https://algoritmeregister.amsterdam.nl> (sneak preview te krijgen via l.fliert@amsterdam.nl) - nu nog een Amsterdams voorbeeld, binnenkort samen met VNG, gemeenten, provincies, waterschappen, Politie en Rijkswaterstaat doorontwikkelen

Binnenkort: bezwaarprocedures doorlichten; werken bezwaarprocedures als een algoritme betrokken was bij besluitvorming of heeft burger een informatieachterstand? (najaar in Amsterdam, later met hierboven genoemde overheden)

Binnenkort: beleidskader en beschrijving governance voor goede interne controle op algoritmes (najaar voor Amsterdam, later met hierboven genoemde overheden)

DEDA, vragenset voor ethische kwesties gemaakt door de Utrechtse data school: <https://dataschool.nl/deda/>

De AI impact assessment van ECP: <https://ecp.nl/actueel/artificial-intelligence-impact-assessment/>

Aanbevelingen van De Waag n.a.v. richtlijnen van min J&V: <https://waag.org/nl/article/algoritme-de-mens-de-machine>

Specifieke richtlijnen voor algoritmes:

Richtlijnen van de High Level Expert Group van de Europese Commissie: <https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>

Richtlijnen van de Raad voor Europa: <https://www.coe.int/en/web/freedom-expression/-/algorithms-and-automation-council-of-europe-issues-guidelines-to-prevent-human-rights-breaches>

Motie Verhoeven/Van der Molen over meldplicht voor algoritmes: <https://www.parlementairemonitor.nl/9353000/1/j9vvij5epmj1ey0/vl1tbgcusrqw>

De richtlijnen voor data-analyse van de overheid door het ministerie van Justitie en Veiligheid: <https://www.rijksoverheid.nl/documenten/rapporten/2019/10/08/tk-bijlage-over-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid>

VNG deed een impactanalyse, om te bepalen of gemeentes deze richtlijnen nu kunnen uitvoeren (belangrijkste conclusies: te vaag voor implementatie + geef gemeentes geld als we dit moeten uitvoeren)

Algemene uitgangspunten over digitale rechten, die bestuurlijk vastgesteld zijn die in Amsterdam als uitgangspunt gebruikt worden. Deze overlappen elkaar (groten)deels:

Declaration of Cities Coalition for Digital Rights: <https://citiesfordigitalrights.org/#declaration>

Vijf principes voor de digitale samenleving van VNG (geïnitieerd door A'dam en Eindhoven): <https://vng.nl/brieven/principes-voor-de-digitale-samenleving>

De TADA principes (over data): <https://tada.city/> (scroll naar beneden)

Voor Amsterdam: Agenda Digitale Stad: <https://magazinedigitalestad.nl>

Het AI magazine van Amsterdam met voorbeelden: <https://www.amsterdam.nl/wonen-leefomgeving/innovatie/digitale-stad/amsterdamse-intelligentie/>

Gerelateerde uitvoeringsagenda's (an sich geen waarden, uitgangspunten of richtlijnen dus);
Amsterdams- en Rijksbeleid voor AI:

AI Actieplan Amsterdam > binnenkort beschikbaar

Strategisch Actieplan voor AI van Ministerie van BZK en EZK: <https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie>

Specifieke aanbevelingen voor inkoopprocedures:

Richtlijnen voor AI-inkoop van de World Economic Forum: http://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf

Richtlijnen van de High Level Expert Group (zie boven bij Specifieke richtlijnen voor algoritmes)

Richtlijnen van de Raad voor Europa (zie boven bij Specifieke richtlijnen voor algoritmes)

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Zie o.a. 'procedurele transparantie' in	Zie o.a. 'procedurele transparantie' in	Zie o.a. 'procedurele transparantie' in	Het beleidskader en	Het beleidskader en

inkoopvoorwaard en algoritmes en de toelichting erop (link naar toelichting)	inkoopvoorwaard en algoritmes en de toelichting erop (link naar toelichting)	inkoopvoorwaard en algoritmes en de toelichting erop (link naar toelichting)	beschrijving van de governance (zie hierboven onder ‘tools’) moet hiervoor de oplossing geven = onderdeel van goedgekeurde aanvraag voor Innovatiebudg et BZK	beschrijving van de governance (zie hierboven onder ‘tools’) moet hiervoor de oplossing geven = onderdeel van goedgekeurde aanvraag voor Innovatiebudg et BZK

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

-
- De verplichting tot het Uitlegbaar maken van de werking van een Algoritmische toepassing is mede gebaseerd op verschillende uitspraken van de Afdeling Bestuursrechtspraak van de Raad van State. Zie onder meer: ABRvS 17 mei 2017, ECLI:NL:RVS:2017:1259 en ABRvS 18 juli 2018, ECLI:NL:RVS:2018:2454.
 - De [SyRI-uitspraak](#)
 - De richtlijnen in mijn antwoord op 2.1 onder “Specifieke richtlijnen voor algoritmes”
-

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Voor ons voorstel van principes en standaarden die hiervoor van belang zijn & waarom, zie de toelichting op de modelcontracten/inkoopvoorwaarden .				

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Voor elk van deze onderdelen geldt dat de tijd rijp is dat we een gestructureerde aanpak met elkaar moeten afspreken. Wij doen daarvoor (met andere overheden) een voorstel met ons register, inkoopvoorwaarden, beleidskader en bezwaarprocedures en onderzoeken wat wel en niet werkt; zie hierboven bij tools. Ik kan evt. ook de aanvraag voor het innovatiebudget van BZK toesturen voor iets meer context.				

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☒ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit: hoewel ik binnen sommige organisaties verandering zie, werken veel ontwikkelaars taakgestuurd, zien zij het (soms expliciet) niet als hun taak om over de beleidscontext na te denken en worden ze niet actief betrokken bij de beleidscontext/besluitvorming. Ook het gebruik en beheer van AI-systemen wordt vaak belegd bij functioneel beheerders, die meestal in de organisatiestructuur geen onderdeel zijn van het beleidsteam dat de inhoudelijke afwegingen maakt.

- ☒ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
- Leg uit: Ik heb hier geen duidelijke mening over. Ik denk dat klopt dat het vaak in jargon geschreven is, maar ik weet niet of dat per se de reden is dat het niet door ontwikkelaars gevonden of gebruikt wordt. Er is bovendien genoeg dat heel toegankelijk is (bijv. DEDA)*

- ☒ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit: absoluut! Dit was voor ons precies de reden om de inkoopvoorwaarden op te stellen (namelijk: waar moet je preciezer zijn dan in een juridisch contract met een leverancier?) en waarom we nu een gebruiksvriendelijk register aan het maken zijn, dat ook voor de ontwikkelaar makkelijk en duidelijk in te vullen moet zijn. De richtlijnen van EC, Raad voor Europa en MinJ&V zijn duidelijk in de zin dat duidelijk is welke waarden en uitgangspunten we moeten hanteren; maar als je moet programmeren is ‘transparantie’ echt te vaag.
- ☐ ☒ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: Ik denk niet dat dat het probleem is. Natuurlijk is elke context en elk proces anders; maar volgens mij kunnen we prima universeel (in ieder geval voor Nederland) afspreken aan welke minimale voorwaarden bijv. transparantie moet voldoen.
- ☐ ☒ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Ik vind het een super goed teken dat er een wildgroei aan richtsnoeren en standaarden is ☺ dat betekent dat veel mensen een probleem voelen! En dat was een aantal jaar geleden echt nog niet zo op dit gebied. We zijn er ook nog lang niet uit met elkaar wat echt werkt. Maar de echt werkbare standaarden en richtsnoeren komen ‘vanzelf’ bovendrijven. Hoewel het natuurlijk wel helpt als er centraal een beetje gecoreerd wordt; het is prettig als bijvoorbeeld BZK op een gegeven moment overzicht geeft van wat er allemaal is en bepaalde tools, richtlijnen en standaarden aanbeveelt.
- ☒ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: precies, zie vorige antwoord. Dat betekent niet dat we maar moeten afwachten of doorgaan met stimuleren van eindeloos veel initiatieven. Probeer vanuit Rijksoverheid te zoeken naar de standaarden die echt potentie hebben en stimuleer de doorontwikkeling daarvan.
- ☒ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: dat wil zeggen, erg veel is afhankelijk van of een organisatie intern toezicht goed heeft ingericht. Ik zou het goed vinden als we in Nederland een soort

risicoprofielen op gaan stellen voor algoritmes (ik geloof dat Harvard met zoiets bezig is). En ik zou het dan logisch vinden dat we toezicht en handhaving inrichten langs die risicoprofielen: een algoritme met het laagste risicoprofiel kan door een ontwikkelaar alleen in de gaten gehouden worden, iets hoger bijvoorbeeld door een ontwikkelaar en beleidsverantwoordelijke; en een algoritme met het hoogste profiel wordt door een nationale toezichthouder gecontroleerd.

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.3.5 Respondent 5

Enquête

Naam: Wouter van Haaften

Organisatie: Hooghiemstra en partners

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

AI is het door middel van data en/of algoritmen nabootsen van menselijke intelligentie

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

*Profiling voor online marketing
Beslismachines voor regelgestuurde organisaties (Belastingdienst, Banken)
Automatische voertuigen*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Overschatting van de mogelijkheden
Onderschatten van de mate waarin de mens verantwoordelijk blijft in juridische zin.
Verlies van 'meaningful human control', de uitkomsten van het AI proces kunnen niet meer door mensen worden begrepen, cq voor hun rekening worden genomen.*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

*Ondersteuning grootschalige processen
Herwaardering van de menselijke intelligentie doordat de kunstmatige faalt*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, veel bureaucratische processen zijn goed te automatiseren.....

*Acceleratoren zijn: efficiency, besparing
...*

Nee want, bij de inzet van AI worden de mogelijkheden overschat, waardoor veel fout gaat

Obstakels zijn: Overschatting AI, gebrekkig inzicht in de verhouding mens/machine, zeker als het gaat om rechtsbescherming....

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: gebiasde dataset...	Problemen: alle gebreken van de ontwikkelaars worden meegeprogrammeerd. In de training wordt de bias versterkt ...	Problemen: Dubbele bias, van de data en het algoritme...	Problemen: Op basis van verkeerde data zijn geen goede beslissingen te nemen. Mensen zien en kunnen heel veel meer dan computers in de maatschappelijke context	Problemen: Moeilijk te beoordelen, alleen als het fout is gegaan is er kans op verbetering ...
Voorbeelden: Bankbestanden, GBA gegevens etc. ...	Voorbeelden: Als je een gecompliceerd bureaucratisch proces laat ontwerpen door IT-ers zonder kennis van de materie raakt het vol ongesanctioneerde interpretaties...	Voorbeelden: Toeslagaffaire en vergelijkbare, minder gepubliceerde misleunen...	Voorbeelden: Veel van de ellende van automatische besluitvorming kan worden voorkomen door bij twijfelgevallen te kiezen voor menselijke tussenkomst...	Voorbeelden: Alle cyclische systemen ...
Oorzaken: onvoldoende nadenken over het systeem. Dat doet het	Oorzaken: Overschatting AI Gebrek aan kennis en overzicht op het vakgebied bij	Oorzaken: Overschatting AI. Gebrek aan kennis en overzicht op het	Oorzaken: Te weinig geld en verstand van zaken	Oorzaken: Onvoldoende inzicht in de aard en omvang van de

systeem toch zelf? ...	opdrachtgever en opdrachtnemer...	vakgebied bij opdrachtgever en opdrachtnemer ...		geautomatiseer de verwerking t.o.v. de klant...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen / trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: big data a priori niet transparant...	Problemen: ontwikkelen en 'trainen?' algoritmen moeilijk transparant te krijgen, zeker als zelf leren wordt toegepast...	Problemen: Hoge kwaliteit data nodig en veel kennis van het veld om relevante patronen te kunnen onderscheiden	Problemen: Interpretatie van de combi algoritme data is cruciaal en mensenwerk	Problemen: transparantie neemt af naarmate de keten vordert door de verwerkingen in de keten
Voorbeelden: Toeslagen, politie automatisering automatische voertuigen	Voorbeelden: : toeslagen banksystemen automatische voertuigen	Voorbeelden: Toeslagen, compliance systemen banken, belastingdienst ,	Voorbeelden: Toeslagen, hypotheekverstrekking , Vergunningverlening, Automatische voertuigen ...	Voorbeelden: ...
Oorzaken: kwaliteit data t.o.v. het doel niet altijd goed verkrijgbaar	Oorzaken: Ontwikkelaar heeft onvoldoende kennis van content, klant onvoldoende van IT...	Oorzaken: Na twee minnen volgt niet altijd een plus...	Oorzaken: Onvoldoende transparante AI leidt tot slechte besluit/beleidsvorming	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

De bovenbeschreven aspecten zijn bij uitstek die waarvoor menselijke intelligentie is vereist.

Aanleggen dataset <i>Hoge kwaliteit data, (Basisreg. serieus nemen)</i>	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data <i>wordt makkelijker als data goed is</i>	(4) Besluit- en beleidsvorming <i>Bij hoge kwaliteit data en verwerking zou AI kunnen helpen</i>	(5) Feedbackloop
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

nee

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Nee

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

What Computers Still Can't do, Hubert Dreyfus, MIT-press 1992

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit: Kan ik niet overzien.....

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit: Zou kunnen maar met name juridische taal is bepalend voor het resultaat. Je hebt IT-ers nodig die het juridische begrijpen (zeldzaam) of je moet rulebased, vanuit het juridisch kader, gaan programmeren. Dit vergt juristen met overzicht over de materie en aardigheid in IT.

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: Het automatiseren van open normen valt niet mee. Toch moeten voor succes de juridische en IT-technische wereld bij elkaar worden gebracht. Rule based werken kan een gemeenschappelijkheid creëren.

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: Kan ik niet overzien.

☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Zou kunnen, standaardkeuze is in principe aan de klant.

☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Sluit ik niet uit.

☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Dat is zeker waar.

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.3.6 Respondent 6

Enquête

Naam: Nadia Benaissa
Organisatie: Bits of Freedom

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ... onvolledige data, verzameld</i>	<i>Problemen: ...</i>	<i>Problemen: ... data geeft op zichzelf niet</i>	<i>Problemen: ... ntransparante beslissingen die</i>	<i>Problemen: ...</i>

vanuit een bepaald en beperkt perspectief	(onbewust) sturende algoritmen	weer welke (sociaal maatschappelijk ke) problemen hebben geleid tot de data	personen raken, maar niet goed uitlegbaar zijn en waartegen verweer of bezwaar moeilijk is	
Voorbeelden: ... data over succesvolle medewerkers die alleen of met name bestaan uit profielen van witte mannen	Voorbeelden: ... algoritmen waarbij de oplossingsrichting voor een probleem (obewust) al is bedacht	Voorbeelden: ... kans op recidive na een veroordeling	Voorbeelden: ... de Belastingdienst	Voorbeelden: ...
Oorzaken: ... beperkt perspectief, maar ook sociaal maatschappelijk ke problemen	Oorzaken: ... bevooroordeeld	Oorzaken: ... een onvolledig beeld van de complexe maatschappij die zich moeilijk laat vertalen tot contextloze data	Oorzaken: ... intransparantie , bevooroordeeld zijn, incompetentie	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: ... grote hoeveelheden data die voor mensen moeilijk te	Problemen: ... complexiteit in de stappen, met name wanneer omstandigheden een rol spelen	Problemen: ... partonen die niet per definitie relevant hoeven te zijn	Problemen: ... moeilijk of niet traceerbare stappen die hebben geleid tot een besluit	Problemen: ...

<i>onderzoeken zijn</i>	<i>bij de besluitvorming</i>			
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ... data is op zichzelf niet contextueel en vertelt niet wat voor maatschappelijke problemen tot de data geleid hebben.</i>	<i>Problemen: ... De stappen in een algoritme kunnen gestuurd worden op een wijze die discriminatie in de hand werkt.</i>	<i>Problemen: ... gedestilleerde patronen uit data zijn niet altijd relevant voor een uitkomst of oplossing.</i>	<i>Problemen: ... door discriminatoire stappen of ‘gekleurde’ data kunnen bepaalde mensen positief of negatief geraakt worden in besluit- en beleidsvorming.</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ... als uit data blijkt dat witte mannen vaker fulltime werken en meer verdienen dan andere groepen, vertelt dat niet dat een mogelijke oorzaak discriminatie van andere groepen is.</i>	<i>Voorbeelden: ... bijvoorbeeld door onderscheid te maken op basis van persoonlijke kenmerken of kenmerken van groepen en dat bij de besluitvorming mee te laten spelen.</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Bestaande problemen: ... onvolledige data, data zonder context	Bestaande problemen: ... algoritmen sturen op basis van vooroordelen	Bestaande problemen: ... irrelevante patronen koppelen aan een oplossing of van invloed laten zijn op besluitvorming	Bestaande problemen: ... intransparantie	Bestaande problemen: ...
AI gedreven oplossingen: ... correcties aanbrengen, de context onderzoeken	AI gedreven oplossingen: ... Een oplossing is niet simpel omdat men zich vaak niet bewust is van eigen vooroordelen of beperkte perspectief. Diversiteit in ervaringen, standpunten, belangen, etc. kan een perspectief verbreden	AI gedreven oplossingen: ... menselijke tussenkomst, wederom vanuit een divers perspectief	AI gedreven oplossingen: ... uitlegbare algoritmen	AI gedreven oplossingen: ...

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
			An ecosystem of trust, waarbij vanuit richtlijnen aandacht wordt gevraagd voor mogelijke risico's van AI en getracht wordt deze risico's te beperken.	

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
Dataminimalisatie, vereiste grondslag en doelbinding voor	Verplichte DPIA		Menselijke tussenkomst	

dataverwerking, onderscheid bij bijzondere persoonsgegevens en andere gevoelige persoonsgegevens				

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
			Er is specifiekere wetgeving nodig om het gebruik van	

			algoritmen en AI te reguleren en risico's te beperken.	

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars,

mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit:

De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit:

De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
Hangt er van af.

Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:

De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:

Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:

Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

7.3.7 Respondent 7

Enquête

Naam: Maranke Wieringa, MA

Organisatie: Datafied Society, Utrecht University

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

'Intelligentie' gedemonstreerd door machines/applicaties, welke tot uiting komt doordat ze hun omgeving waarnemen en hun 'gedrag' daarop kunnen aanpassen.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Automatische vertalingen van teksten, predictive policing, frauderisicodetectie, facial recognition.

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

-
- 1) Het ontwerp/gebruik van algoritmen/AI gaat gepaard met vertalingen van wet/regelgeving naar beleid, naar implementatie, naar algoritmen. Die keuzes zijn niet neutraal.*
 - 2) De keuzes omvatten een tijds- en plaatsgebonden afweging van publieke waarden. Die aannames over normen/waarden verandert in de loop van de tijd in de samenleving, maar het systeem is rigide. Systemen gaan dus functioneren als tijdschapsules, die op een gegeven moment mogelijk uit de pas lopen met de samenleving.*
 - 3) Het herorienteren/herontwerpen van (de aannames van de waarden en normen) in systemen is heel erg moeilijk (bijv. vendor lock-in, systemen raken vervlochten met elkaar, complex/groot/ondoorgroendelijk systeem) en kostbaar (tijd/geld).*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

- 1) *Mogelijkheden om (mits adequaat gebruikt) publieke waarden te realiseren.*
- 2) *Efficiëntie.*
- 3) *Automatisering.*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, (lokale) overheden zijn er steeds meer mee aan het experimenteren

Acceleratoren zijn: hype, idee om kosten te besparen.

Nee want, kennis ontbreekt nog vaak.

Obstakels zijn: ontbreken kennis, in moeten kopen, zelf geen eigenaar zijn. Gebrek aan literacy/begrip/overzicht van implicaties.

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overall iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: Data is altijd capta (zie Drucker 2011). Data is niet een</i>	<i>Problemen: Trainen op hond/wolf herkenning of trainen op</i>	<i>Problemen: Correlatie is geen causatie. Een patroon</i>	<i>Problemen: Vaak weinig ruimte voor grijze gebieden/discreti</i>	<i>Problemen: Self-fulfilling prophecies/vicieuze cirkel.</i>

gegeven iets maar een geconstrueerd iets. Niet in de werkelijkheid maar onttrokken daaraan. De keuze om iets te meten in plaats van iets anders is niet neutraal. Datasets zijn daarmee ook nooit neutraal. Versterking van bestaande bias doordat trainingsdata skewed is.	sneeuw? Het is vaak lastig om te controleren waar het systeem precies op optimaliseert (zeker bij bijv. deep learning).	betekent nog niet perse dat er ook daadwerkelijk een verband is. Versterking van bestaande ongelijkheden.	e in besluiten. Wie is verantwoordelijk? Wat is echte menselijke tussenkomst?	
Voorbeelden: datasets van CBS uitgesplitst naar migratie-achtergrond	Voorbeelden: Husky/wolf onderscheid leren aan algoritme. Algoritme herkent aanwezigheid sneeuw..	Voorbeelden: Het aantal doden door verdrinking in zwembaden correleert met het aantal films waarin Nicolas Cage verscheen. Dat wil niet zeggen dat dit ook een zinnig patroon is. Amazon die een hiring algorithm opzette die alleen maar mannen aannam (want die werden in het verleden het meeste aangenomen)	Voorbeelden: SyRI uitspraak	Voorbeelden: predictive policing

Oorzaken: <i>Onvoldoende literacy mbt dataverzameling</i>	Oorzaken: <i>Technische ontransparantie/computationale logica en optimalisatie is niet vergelijkbaar met menselijke logica</i>	Oorzaken:	Oorzaken: ...	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Waarom is AI überhaupt nodig? Wat waren afwegingen tussen alternatieven (zowel tussen AI en niet-AI/algorithmische oplossingen)?

Transparantie over algorithmic presence

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>Trainingsdata kan niet altijd vrijgegeven worden</i>	Problemen: <i>Begrijpbaarheid van de systemen</i>	Problemen: <i>Begrijpbaarheid, meet je wat je denkt te meten</i>	Problemen: <i>Traceerbaarheid</i>	Problemen: <i>Resultaten verifiëren</i>
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

In het algemeen geldt voor algoritmen:

Algoritmen (dus ook AI) zijn vertaalslagen. Van wet/regelgeving → beleid → beleidsimplementatie → algoritmische implementatie. In elk van die vertaalslagen worden keuzes en afwegingen gemaakt.

Die keuzes/afwegingen zijn plaats- en tijdsgebonden. Dat maakt algoritmes/AI dus ook niet neutraal, omdat die keuzes/afwegingen/aannames ingebouwd worden in het systeem.

Systemen zijn vaak moeilijk te heroriënteren/herontwerpen/hertrainen, omdat het kostbaar (tijd/geld) is, of omdat systemen vervlochten raken met andere systemen.

Algemene problemen:

Gebrek aan literacy;

AI benaderen als alleen een technisch iets ipv socio-technisch.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>Data is altijd incomplete afspiegeling van de werkelijkheid</i>	Problemen: <i>Publieke waarden afwegen</i>	Problemen: <i>Versterking van bestaande ongelijkheid</i>	Problemen: <i>Systeem en haar besluiten kunnen uit de pas gaan lopen met de werkelijkheid</i>	Problemen: <i>Onderliggende aannames worden vaak niet getoetst</i>
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Ja, je kan het gebruiken om bijv. bias aan te tonen. (Zie Amazon hiring algorithm) Met die resultaten zou je bijv. weer nieuw en beter beleid kunnen voeren. Ook zou het positief zijn als besluiten beter getraceerd kunnen worden, MAAR: 1) het moet in een begrijpelijk format, 2) we moeten de besluiten over het besluitsysteem niet vergeten hierin, anders ben je het probleem alleen aan het verschuiven.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

	<i>trainen algoritme</i>			
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Ja, die ken ik.

Los van algemene richtlijnen (e.g. Grondwet, ABBB), zijn de meeste specifiek op AI-gerichte richtlijnen/etc. vaag/hoog over (e.g. EU HLEG AI), en/of focussen zich voornamelijk op het strikt technische (e.g. conceptrichtlijnen BZK/J&V) zijn er veel guidelines/richtsnoeren (e.g. EU HLEG AI, CoE, richtlijn van BZK/J&V in ontwikkeling).

Een goed startpunt vind ik Principles for Accountable Algorithms

(<https://www.fatml.org/resources/principles-for-accountable-algorithms>)

En het literatuuronderzoek wat ik op dit thema heb uitgevoerd bundelt nog meer aspecten waar men rekening mee moet houden: <https://dl.acm.org/doi/abs/10.1145/3351095.3372833>

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
Door het hele proces heen (en om het proces heen).				

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Ja. GDPR, SyRI/SUWI casus, EHCR, Grondwet etc.

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
Blijft door het hele proces heen cruciaal.				

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Zie de uitvoerige literatuurstudie die ik heb uitgevoerd:

<https://dl.acm.org/doi/abs/10.1145/3351095.3372833>

Het moet niet alleen draaien om technische standaarden/principes. Want AI is niet alleen technisch. Het is socio-technisch.

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Zie o.a. <https://www.nature.com/articles/s42256-019-0088-2> voor een overzicht

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan

waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

We missen nog de focus op het hele proces, ipv onderdelen.

We missen de aandacht voor het socio-technische ipv alleen het technische

We missen enforceability. Het is te vaak vrijblijvend.

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

Dat ligt enorm aan de organisatie.

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

Niet per se. Dat verschilt per document/richtsnoer.

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit:

Ja, dit is een probleem. Veel organisaties willen graag maar weten niet hoe.

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
Ja.
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
Dat is een risico. Anderzijds, denk ik dat verschillende situaties misschien ook voor verschillende aandachtspunten zorgen (een AI voor VRI besturing is nu eenmaal iets anders en heeft andere gevoeligheden dan een AI die over uitkeringsfraude gaat).
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
Ja, maar ook: het is nu te veel vbrijblijvend.
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
Ja, zeker. Maar ook: het ontbreekt ons aan bestaande/enforceable/auditable standaarden.
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Het opbreken in stappen/delen van het proces is logisch op papier, maar in de praktijk is het veel rommeliger. Stappen vloeien in elkaar over, en er hoeft geen lineaire ontwikkeling te zijn.

Neem het socio-technische aspect mee. Het systeem bestaat niet in een vacuum. Als er als aan het project als iets schort kun je je AI nog zo non-discriminatoire maken, maar dan schiet ie er als noa niets mee op.

7.3.8 Respondent 8

Enquête

Naam: Max Vetzo

Organisatie: Universiteit Utrecht

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Systemen die met een hoge mate van autonomie intelligent handelen om specifieke doelen bereiken.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Gezichtsherkenning – Contentmoderatie – Medische diagnostisering

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Ongelijke behandeling (discriminatie)
Rechtsbescherming (toegang tot de rechter en eerlijk proces)
Gegevensbescherming*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Verhelpen discriminatie – Efficiëntie en in besluitvorming – Domeinspecifieke voordelen (e.g. misdaadbestrijding, ziektepreventie etc.)

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

Ja want, ...duidelijke maatschappelijke en economische voordelen, geopolitieke incentives.

Acceleratoren zijn: ... grootschalige investeringen, uitgebreide onderzoeks aandacht

Nee want, potentiële gevaren zijn duidelijk aanwezig. Belemmert grootschalige implementatie

Obstakels zijn: te rigide interpretatie juridische kaders, houding van ‘enkel denken en (nog) niet doen’, i.p.v. ‘trial and error’

1. Probleemanalyse

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: ... Niet-representatieve of actuele dataset	Problemen: ... Bij supervised learning o.m. biases in labelen data	Problemen: ... Discriminatoire proxy's	Problemen: Gebrekkig corrigerend vermogen van betrokken menselijke actoren	Problemen: ... Discriminatoire output (door oorzaken hiervoor) gebruikt als input versterkt discriminatie
Voorbeelden: Werken met datasets waarin (door historisch gegroeide ongelijkheden) specifieke groepen niet (voldoende) of juist overmatig zijn meegenomen.	Voorbeelden: ... Zie ter illustratie Barocas & Selbst 2016, p. 681-684	Voorbeelden: ... Betrekken verboden gronden van onderscheid (bijv. etniciteit of geslacht) als relevante variabele in model	Voorbeelden:	Voorbeelden: ... E.g. predictive policing waarbij hogere misdaadcijfers (ten gevolge van intensievere surveillance in door AI aangewezen gebieden) worden gebruikt als input voor verdere training AI-systeem.
Oorzaken: ... Variabel, maar mogelijk	Oorzaken: ... Onbewuste vooroordelen die	Oorzaken: ... Onvoldoende waarborgen	Oorzaken: ... Afwezigheid technologische	Oorzaken: ... Gebrekkige reflectie op

gerelateerd aan complexiteit verkrijgen accurate, representatieve en actuele dataset	doorwerken in labelen van data	tegen ‘oppikken’ van dergelijke verboden gronden van onderscheid door algoritme.	kennis i.c.m. gebrekkige uitlegbaarheid AI-systemen	biases die zich eerder in proces voordoen.

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: Sterk variabele, constant geactualiseerde (dynamische) datasets kunnen ertoe leiden dat moeilijk achterhaald kan worden op welke data een besluit is gebaseerd.	Problemen: ... Inherente complexiteit van deep learning algoritmes. Hoge mate van autonomie en complexiteit gevonden correlaties bemoeilijkt transparantie en daarmee begrijpelijkheid.	Problemen: ... Ibid. 2.	Problemen: ... Motivering van besluit- en beleidsvorming waarbij geen aandacht wordt besteed aan uitleg, of sterk de nadruk ligt op (voor burgers niet begrijpelijke) technische transparantie	Problemen: ...
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ... Niet-registratie van data op basis waarvan op een specifiek moment een besluit is genomen.	Oorzaken: ... Inherente complexiteit en ondoorzichtigheid.	Oorzaken: ... Ibid 2.	Oorzaken: ... Gebrekkige technologische kennis of intrinsieke complexiteit en ondoorzichtigheid van systemen.	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: Gegevens- bescherming (dataminimalisatie en reikwijdte)	Problemen:	Problemen: ... Gegevensbescherming: analysefase nagenoeg ongereguleerd en in deze fase (vaak, in ieder geval bij Big Dataverzamelingen) met hoog geaggregeerde data.	Problemen: ...	Problemen: ...
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ... Streven naar dataminimalisatie i.c.m. behoefte aan rijke datasets met veel informatie over voorliggende gevallen. Data niet zijnde persoonsgegevens niet onder reikwijdte AVG	Oorzaken: ...	Oorzaken: ... Aandacht van meest geavanceerde rechtsinstrument ter zake (AVG) niet gericht op analysefase, maar op fase van dataverzameling.	Oorzaken: ...	Oorzaken: ...

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- beleidsvorming en	(5) Feedbackloop
Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: ... 1. Biases zijn inherent aan menselijke besluitvorming. 2. Menselijke besluit- en beleidsvorming is eveneens ondoorzichtig. Het wantrouwen tegenover moeilijk uitlegbare AI-toepassing getuigt deels van overdadig vertrouwen in de menselijke mogelijkheid tot uitlegging van zijn eigen oordeels- en besluitvorming. 3. Menselijke besluit- en beleidsvormig is niet altijd consistent. Ook dat kan leiden tot vormen van ongelijkheid.	Bestaande problemen: ...
AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ... 1. Goed geprogrammeerde en geteste AI-toepassingen zijn potentieel objectiever dan mensen. Daarbij kunnen AI-systemen grote hoeveelheden informatie over het individu meenemen. Dat leidt tot potentieel sterk	AI gedreven oplossingen: ...

			geïndividualiseerde besluitvorming, die – als discriminatierisico’s in acht worden genomen – bijdraagt aan behandeling op basis van relevante, individuele kenmerken. Dat valt te prefereren boven bevooroordeelde menselijke besluitvorming. 2. Ook AI-toepassingen waarbij uitlegbaarheid reeds in de programmeerfase aandacht heeft gekregen en besluit of beleid <i>ex post</i> wordt gemotiveerd en uitgelegd, vormen een oplossing voor problemen met de ondoorzichtigheid van menselijke besluitvormingsprocessen. 3. AI-systemen kunnen bijdragen aan de consistentie van besluitvorming.	

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Op Europees niveau springen de ethische richtsnoeren in het oog. Zie bijvoorbeeld de 'Ethics Guidelines for Trustworthy AI'. Voor literatuur over een veelheid aan ethische richtsnoeren, zie: 'The global landscape of AI ethics guidelines' (<https://www.nature.com/articles/s42256-019-0088-2>). Deze zijn evenwel niet bindend en worden gekenmerkt door een hoog abstractieniveau. Dat geldt in veel mindere mate voor de op nationaal niveau tot stand gekomen 'Richtlijnen voor het toepassen van algoritmes door overheden' (bijlage bij Kamerstukken II 2019/20, 26643 en 32761, nr. 641), die relatief gespecificeerd en daarmee richtinggevender zijn. Ook deze richtsnoeren zijn niet bindend. Zij beperken zich bovendien tot de publieke sector. Daarbij is wat mij betreft wel sterk de vraag in hoeverre deze toch wat beleidsmatige stukken, daadwerkelijk technologisch richting geven bij het vormgeven van AI-

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

De hierboven aangeduide richtsnoeren zien op het gehele proces dat AI-systemen doorlopen.

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Nederlandse wetgeving: niet, anders dan 'algemene' wetgeving die ook op AI-systemen van toepassing kan zijn. Ik denk primair aan de AVG, de Awb en – belangrijk – grondrechtenstandaarden, die – hoewel abstract – een potentieel belangrijke rol kan spelen bij het opvangen van de negatieve gevolgen van de inzet van AI-systemen.

Rechtspraak (over algoritmes, waarbij veelal geen sprake is van AI, maar van regelgebaseerde algoritmes), o.m:

- ABRvS 17 mei 2017, ECLI:NL:RVS:2017:1259 (AERIUS I)
- ABRvS 18 juli 2018, ECLI:NL:RVS:2018:2454 (AERIUS II)
- CRvB 15 mei 2019, ECLI:NL:CRVB:2019:1737
- HR 17 augustus 2018, ECLI: NL:HR: 2018:1316
- Rb. Amsterdam 4 juli 2019, :NL:RBAMS:2019:4799
- Rb. Den Haag 5 februari 2020, ECLI:NL:RBDHA:2020:865 (SyRI)

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
			De rechtspraak ziet primair op de inzichtelijkheid van het algoritmisch besluitvormingsproces. Daaraan worden uitlegbaarheidseisen gesteld. De genoemde algemene kaders strekken zich potentieel uit over de volle breedte van de ontwikkeling van AI-systemen. Dat laat zich niet eenvoudig in deze tabel vatten.	

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

In vele studies worden aanvullende standaarden neergelegd. Belangrijk, en recent, vind ik daarbij de publicatie van Heleen Janssen, 'An approach for a fundamental rights impact assessment to automated decision-making' (International Data Privacy Law, 2020).

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

	<i>trainen algoritme</i>			
			De publicatie van Janssen richt zich op DPIA's en geeft daaraan een specifieke mensenrechtelijke invulling, die over de volle breedte van het AI-proces van belang kan zijn.	

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

De mij bekende standaarden hebben een algemeen karakter. Dat kan behulpzaam zijn en biedt algemene richting. Tegelijkertijd kan ik mij voorstellen dat er behoefte is aan domeinspecifieke instructies, met een meer technische insteek. De algemene tekortkomingen die ik in dit document heb aangegeven worden ongetwijfeld 'afgedekt' door abstract geformuleerde principes, maar de vraag is of deze principes in de praktijk richtinggevend werken.

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit: ongetwijfeld zal sterkere bewustwording een positieve rol spelen bij het implementeren van de richtlijnen. Of dat op dit moment onvoldoende het geval is, kan ik lastig beoordelen. Daar waar wettelijke kaders specifiek én bindend zijn (zoals de AVG), is van onbekendheid hoogstwaarschijnlijk geen sprake.

- x ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit: Zie opmerking onder 2.1.

- x ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals

‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: Zie opmerking onder 2.1.

- x ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is

Leg uit: De ontwikkeling en werking van AI-systemen is sterk afhankelijk van de context waarbinnen zij opereren. Algemene standaarden vangen de bijzonderheden van deze context niet altijd, al leggen zij wel een abstracte ondergrens neer.

- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)

Leg uit: Zie antwoord op eerste vraag. Dat is mij niet bekend, al lijkt mij dat een proliferatie van standaarden weinig behulpzaam is en ‘standard-shopping’ in de hand kan werken, zeker zolang concrete, juridische bindende eisen zich nog uit moeten kristalliseren.

- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt

Leg uit: Zie antwoord op eerste vraag. Dat is mij niet bekend,

- x ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes

Leg uit: Toezicht en handhaving richt zich op bindende kaders (zoals de AVG en Awb). Ik kan mij goed inbeelden dat, dat voor niet bindende standaarden (waarvan er bovendien meerdere zijn) in het geheel niet het geval is, althans niet van overheidswege.

- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Neen, niet anders dan hierboven aangegeven.

7.3.9 Respondent 9

Enquête

Naam: Rob van den Hoven van Genderen

Organisatie: Vrije Universiteit (Universiteit van Lapland)/ Switchlegal advocaten

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Een systeem van dataverwerking , analyse en handeling of besluitvorming dat op een of meerdere vlakken kan worden vergeleken met de analyse en handelingsoefening van menselijke intelligentie

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Surveillance en beveiliging door AI s; Autonome voertuigen, Autonome medische toepassingen (chirurg)

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Te veel varen op de resultaten van AI bij menselijke besluitvorming
Inbreuk door de overheid op de fundamentele rechten
Misbruik van AI voor criminele acties*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

*Oplossing armoede en honger door agrarische toepassingen
Oplossingen op het gebied van voorkoming vervuiling en duurzaamheid
Oplossing medische problemen door onderzoek en toepassing in de*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, efficiëntie en effectiviteit zal bevorderd worden.....

Acceleratoren zijn: ...opleiding, voorlichting , vertrouwen

Nee want, opleiding van natuurlijke personen zal achter blijven....

Obstakels zijn: opleiding, voorlichting, vertrouwen....

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset:	(2) Ontwikkelen/ trainen algoritme:	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: ... indien data zijn gebaseerd op bestaande statistieken en selectief gekozen data	Problemen: ... indien door natuurlijke personen geselecteerde data	Problemen: ...Indien geprogrammeerd door vooringenomen natuurlijke personen	Problemen: ...Indien uitgevoerd door vooringenomen natuurlijke personen	Problemen: ... idem
Voorbeelden: ... Besluitvorming overheid bij subsidies, uitkeringen, rechterlijke uitspraken etc	Voorbeelden: ... Programmeren en selectie door vooringenomen programmeurs of keuze selectieve data	Voorbeelden: ... Keuze door AI of menselijke controle	Voorbeelden: ... Zie 1	Voorbeelden: ...
Oorzaken: ... Statistieken dragen bias	Oorzaken: ... idem	Oorzaken: ... Verkeerde keuze	Oorzaken: ... Politieke orientatie of desinteresse, gemakzucht etc.	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo

specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluiten beleidsvorming	(5) Feedbackloop
<i>Problemen: zie boven: afhankelijk van selectiemechanisme door betrokken natuurlijke actoren...</i>	<i>Problemen: ...transparantie is in dit geval waarschijnlijk niet nuttig of noodzakelijk</i>	<i>Problemen: kwestie van geloofwaardige voorlichting over de selectie ...</i>	<i>Problemen: ... Alleen transparantie als noodzakelijk, transparante klachten systeem</i>	<i>Problemen: ... Garbage in garbage out and in again 😊</i>
<i>Voorbeelden: ... Selectie doelgroep; geografische/functionele locatie</i>	<i>Voorbeelden: ... Inzage in selectiemechanisme door ondeskundigen/betrokkenen kan leiden tot verkeerde conclusie en/of vertraging</i>	<i>Voorbeelden: ...ten behoeve van waar aan besluiten in 4</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ... Onbekendheid of bewuste keuze Gebrek aan mogelijkheden toegang door betrokkenen of juist bewust vermijden, terecht of onterecht van transparantie</i>	<i>Oorzaken: ... Ondeskundigheid, slechte voorlichting aan betrokkenen</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Ik zie weinig verschil met voorgaande vragen. Het gaat om uitsluiten maatschappelijke problemen door objectieve criteria en vermijden van vooringenomenheid. Daarnaast is goede voorlichting en opleiding bij het gebruik van AI zeer relevant.

(1) Aanleggen dataset	(2) Ontwikkelen / trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>Gezien bovenstaande is het nooit een technisch probleem maar het is aan de programmeur/selectiemechanisme ho de dataset wordt vormgegeven....</i>	Problemen: ... <i>Dit geldt evenzo voor het trainen van het algoritme</i>	Problemen: ...	Problemen: ...	Problemen: ...
Voorbeelden: ...	Voorbeelden: : ...	Voorbeelden: : ...	Voorbeelden: : ...	Voorbeelden: : ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Bestaande problemen: <i>wederom vooringenomenheid ligt bij de aanlevering door natuurlijke personen, aanleg data set. Kan eventueel door AI gebeuren maar cirkel bias moet dan voorafgaand worden vermeden...</i>	Bestaande problemen: ... Idem door objectieve AI systeem	Bestaande problemen: ... Idem	Bestaande problemen: ... Hier wel samenwerking AI en natuurlijke besluitnemers	Bestaande problemen: ... Indien objectieve informatie

<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen / trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluiten beleidsvorming</i>	<i>(5) Feedback</i>
https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_well_being.pdf https://ethicsstandards.org/wp-content/uploads/2019/12/OCEANIS-Think_Piece-12052019.pdf etc.	idem	idem	OECD/ EU ethical commit tees, Raad van Europa	

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de

belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

x ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
☐ *Leg uit: Veel van hetzelfde zonder concrete handvatten maar overigens onbekend bij de ontwerpers.....*

x ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
☐ *Leg uit: Inderdaad.....*

x ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
☐ *Leg uit: Inderdaad.....*

- x ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
☐ *Leg uit: Inderdaad*
- x ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
☐ *Leg uit: ...Zou kunnen, hoewel de meeste richtlijnen op hetzelfde neerkomen..*
- ☐ x☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Als hierboven geconstateerd, te algemeen en een goede zeef gecombineerd met per gebied/functie praktische handleiding zou zijn vruchten af kunnen werpen.....
- ☐ x☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: ...Een brug te ver, niet te handhaven, te vaag , niet afdwingbaar , geen reële wetgeving etc..
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

***Nogmaals, goede voorlichting, onderricht op multidisciplinaire basis.
 Functioneel selectief, geen algemene dooddoeners zoals bij de meeste voorstellen voor richtlijnen.***

7.3.10 Respondent 10

Enquête

Naam: Sander Klous

Organisatie: UvA/KPMG

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Iedere geautomatiseerde vorm van data analyse.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Zelfrijdende auto's, slim aansturen van stoplichten, aankoopadvies bij webwinkels.

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Verkeerde interpretatie van uitkomsten die leid tot foute beslissingen.
Ongewenste bijeffecten, zoals uitsluiting van bepaalde groepen omdat bijvoorbeeld gezondheidsrisico's te hoog zijn.
Het leggen van verantwoordelijkheden bij verkeerde groepen, bijvoorbeeld Data Scientists laten beslissen over wat ethisch is of niet.*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

*Lagere Kosten, bijvoorbeeld door voorspellend onderhoud.
Betere/Nieuwe producten/diensten zoals gepersonaliseerde gezondheidszorg
Reduceren van risico's, bijvoorbeeld door beter inzicht in netwerkeffecten*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

Ja want, we zien dat dit in een aantal sectoren al plaatsvindt. De functionaliteit is er al, de toepassingen moeten daar vaak nog bij gezocht worden alsook hoe dit op een manier vorm te geven dat dit voor de maatschappij en individuen acceptabel is. Dat zijn in mijn ogen praktische en geen fundamentele problemen en dus binnen afzienbare tijd oplosbaar.

Acceleratoren zijn: BigTech die hier veel in investeert. Overheid en Regulators die bezig zijn de randvoorwaarden te scheppen (e.g. inzake privacy en uitlegbaarheid). Onderzoekers, media en betrokken burgers die zorgen voor bewustwording bij het grote publiek.

Nee want, het blijkt erg lastig om people/process/technology zo in te richten dat een organisatie hier ook echt in het bedrijfsproces gebruik van kan maken. Bovendien blijken er keer op keer haken en ogen te zitten aan de inzet van AI die niet waren voorzien en waardoor het vertrouwen van de samenleving wordt beschadigd.

Obstakels zijn: genoeg mensen met de juiste skills. Het introduceren van een werkwijze die recht doet aan de complexiteit van het onderwerp en basale zaken als data kwaliteit, borging van privacy, schaalbaarheid en de investeringen die hiervoor nodig zijn. Ook inrichting van risicomanagement en controleprocessen staat nog in de kinderschoenen.

1. Probleemanalyse

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset Brondata wordt bij AI vaak gebruikt als definitie van succes.	(2) Ontwikkelen/trainen algoritme Geen individuele oorzaak – gevolg relatie	(3) Patronen destilleren uit data Het zogenaamde “look elsewhere effect”	(4) Besluit- en beleidsvorming Verkeerd interpreteren van resultaten	(5) Feedbackloop <i>Self-fulfilling prophecies</i>
Problemen: Als de labels niet kloppen sluipt discriminatie via de data het proces in.	Problemen: Algoritmes worden statisch getraind op reduceren van false positives/false negatives. Daarbij wordt dus bewust geaccepteerd dat er op individueel niveau fouten kunnen voorkomen.	Problemen: Als maar naar genoeg verschillende variabelen wordt gekeken vind je altijd wel een correlatie die puur veroorzaakt wordt door een statistisch effect.	Problemen: Resultaten van analyse zijn vaak alleen bruikbaar als de context van de analyse goed is gedefinieerd. Anders leidt het al snel tot verkeerde conclusies.	Problemen: Door een bias in de resultaten wordt bij gebruik van het algoritme het effect ook aangetoond, wat vervolgens leidt tot versterking van de bias in het algoritme.
Voorbeelden: De beste kandidaten bij sollicitaties zijn in historische data gelabeld door recruiters en vaak niet bias free.	Voorbeelden: Apple pay die credit limit lager instelt voor vrouw dan man terwijl ze kredietwaardiger is	Voorbeelden: correlaties tussen het uitbrengen van films met een bepaalde acteur en de beurskoers.	Voorbeelden: Jongeren met een Marokkaanse achtergrond zijn criminelere dan jongeren met een Nederlandse achtergrond.	Voorbeelden: Een algoritme voorspelt hogere criminaliteit in wijk X, waardoor in wijk X meer gesurveilleerd wordt en dus ook meer criminaliteit wordt geconstateerd. Die informatie wordt meegenomen bij de hertraining, waardoor nog meer nadruk op wijk X komt te liggen.
Oorzaken: Recruiters hebben zelf	Oorzaken: Een incidentele false positive of	Oorzaken: Iedere analyse wordt	Oorzaken: Als criminaliteit van jongeren	Oorzaken: Definitie van succes is niet

<i>bewust of onbewust gediscrimineerd op e.g. ras, leeftijd of geslacht.</i>	<i>negative die waarschijnlijk ook andersom optreedt.</i>	<i>opzichzelf beschouwd. Stel dat er een kans is van 1 op 1.000.000 is dat een uitkomst het gevolg is van een statistische fluctuatie. Dan lijkt de kans heel klein dat zich dat toevallig voordoet, behalve als je ook een miljoen verschillende soorten analyses doet.</i>	<i>met verschillende etnische achtergrond wordt afgezet tegen economische toestand van het huishouden blijkt dit verschil te verdwijnen. Etniciteit is hier dus een proxy voor gebrek aan welvaart.</i>	<i>goed opgesteld. Die zou zo moeten worden opgesteld dat objectief kan worden vastgesteld of het algoritme voorspellende waarde heeft. In genoemde voorbeeld kan dat wellicht door surveillance tijd mee te modeleren.</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: data is vaak onderdeel van een aggregatie moet die dan ook weer verwijderd kunnen worden?</i>	<i>Problemen: wat is de transparantiebehoefte van een bepaalde stakeholder t.a.v. trainen en patronen destilleren. Die verschilt enorm.</i>		<i>Problemen: Algoritmes zorgen er vaak voor dat meer uitleg mogelijk is, maar die uitleg leidt ook tot meer vragen.</i>	<i>Problemen: De feedback loop kan voor polarisatie zorgen waar de consument zich niet van bewust is.</i>
<i>Voorbeelden: Er leiden in NL 500 mensen aan ziekte X. Eén persoon trekt</i>	<i>Voorbeelden: Een datascientist zal technische transparantie willen over het trainingsproces, een rechter vooral transparantie over</i>		<i>Voorbeelden: Klassieke variant – U krijgt geen lening, want uw salaris is</i>	<i>Voorbeelden: De filter bubble is een bekend fenomeen waarbij op maat geleverde zoek resultaten er voor zorgen dat men niet meer wordt</i>

opt-in in, moeten we dan 499 mensen melden? En alle resultaten gebaseerd op de 500 ook aanpassen?	procesmatige borging van uitkomstkwaliteit. Een burger zal zich willen kunnen vergelijken met anderen.	minder dan X. Vraag: waarom X? Antwoord: dat is de regel (geen verdere uitleg mogelijk). - U krijgt geen lening want ons algoritme schat in dat u de aflossing niet kan betalen met Y% zekerheid. Vraag: waarom Y? Antwoord: onderbouwing vanuit het algoritme (maar daar zitten aannames in die vervolgens ter discussie kunnen worden gesteld)	uitgedaagd op denkbeelden. Op die manier ontstaat voortdurende bevestiging van een bestaand beeld en mist in feite de transparantie in de informatievoorziening die er voor kan/moet zorgen dat er voldoende nuance wordt aangebracht.
Oorzaken: Transparantie over welke gegevens allemaal zijn afgeleid van brondata is heel ingewikkeld. Zeker over organisaties heen.	Oorzaken: Verschillende stakeholders hebben verschillende kennisniveaus en verschillende belangen. Transparantie op zichzelf is geen doel. Door middel van transparantie willen we stakeholders in staat stellen om iets te doen, of bijvoorbeeld om meer vertrouwen te hebben. Afhankelijk van dat doel is ook het soort benodigde transparantie anders.	Oorzaken: ook X komt voort uit een risico-inschatting, maar die is inmiddels vastgelegd in regelgeving en kan niet meer ter discussie worden gesteld. Algoritmes bieden veel meer flexibiliteit en mogelijkheden en dus mag er meer verwacht	Oorzaken: Volledig geautomatiseerde feedback loop, zonder interventiemogelijkheden voor de consument/gebruiker.

		worden qua uitleg.	

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: Gevoelige data mag niet worden vastgelegd, maar is wel nodig om fairness van algoritme te toetsen.	Problemen: ...	Problemen: ...	Problemen: wet en regelgeving stelt eisen aan besluitvorming door algoritmes waar menselijke besluitvorming nu niet aan voldoet, alleen is het bij menselijke besluitvorming lastiger hard te maken.	Problemen: ...
Voorbeelden: ik wil toetsen of mijn algoritme net zo veel false positives / negatives heeft voor verschillende etnische groepen, maar mag die gegevens niet verzamelen.	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: Recruiters discrimineren waarschijnlijk meer dan algoritmes, maar bij algoritmes is het beter inzichtelijk te maken, e.g. d.m.v. simulaties en training.	Voorbeelden: ...
Oorzaken: GDPR	Oorzaken: ...	Oorzaken: ...	Oorzaken: De eis dat discriminatie bij sollicitaties niet is toegestaan, is logisch, maar bij algoritmes	Oorzaken: ...

			<i>zal je altijd m.b.v. statistische middelen kunnen aantonen dat er nog wat discriminatie plaatsvindt. Bij menselijke besluitvorming is de discriminatie waarschijnlijk veel groter, maar onmogelijk te herleiden naar een individuele beslissing.</i>	

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Is bij mijn eerdere antwoorden al een aantal keer ter sprake gekomen.

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

IEEE 7010-2020 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
EPRS The impact of the General Data Protection Regulation (GDPR) on AI
OpenAI Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis,

interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:

☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:

☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:

☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.3.11 Respondent 11

Enquête

Naam: Marc van Lieshout
Organisatie: RU - iHub

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

AI staat voor de toepassing van op wiskundige principes gebaseerde protocollen in geautomatiseerde toepassingen die resultaten produceren die door mensen als 'intelligent' kunnen worden ervaren.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

*Zelfrijdende auto
Zelfdiagnostische systemen
Ondersteuning rechtspraak.*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Bias
Stigmatisering
Onterecht vertrouwen*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

*Ondersteuning professionals in patroonherkenning
Verbreiding inzetbaarheid patroonherkenning in uiteenlopende situaties
Consumentenapps*

Ja want, ...

Acceleratoren zijn: ...

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

Nee want, ...er is een gerede kans dat de algemene inzetbaarheid van AI zoals die nu gepropageerd wordt uiteindelijk tot veel basalere toepassingen leidt (die overigens op zich toegevoegde waarde zullen hebben).

Obstakels zijn: technische complexiteit om tot sluitende oplossingen te komen; weerstand van mensen om zonder meer op technische uitkomsten te vertrouwen; uiteindelijk beperkte toepasbaarheid in strikt afgebakende probleemgebieden.

1. Probleemanalyse

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: ... Bias	Problemen: ...al leen oog voor al bekende vraagstukken en problemen	Problemen: ...Stigmatisering	Problemen: ...	Problemen: ...
Voorbeelden: ... Benutting foto's voor gezichtsherkenning; Herkennen melanomen;	Voorbeelden: ... Sollicitatieprocedures; herkennen schaduwen	Voorbeelden: ... Predictive policing Herkennen kankersoorten	Voorbeelden: ... SyRI	Voorbeelden: ...

Voorspellen crimineel gedrag				
<i>Oorzaken: ...In de regel onvoldoende oog voor herkomst van dataset en de daarbij gebruikte parameters.</i>	<i>Oorzaken: ...Complexiteit van real world</i>	<i>Oorzaken: ...Op zoek naar herkenbare relaties; geen oog voor uitbijters</i>	<i>Oorzaken: ...Belidsmakers moeten ook scoren</i>	<i>Oorzaken : ...</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ... Herkomst van data</i>	<i>Problemen: Uitlegbaarheid</i>	<i>Problemen: ...</i>	<i>Problemen: Onmogelijke keuzes</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ... Kwaliteitsregistraties Biobanken</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: 'trolley probleem bij zelfrijdende auto's</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...grote datasets afkomstig van uiteenlopende bronnen</i>	<i>Oorzaken: wiskundige complexiteit van werking algoritme</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: verandering ethische gedragsregels (van ad hoc naar op voorhand)</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destillieren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>Kwaliteit, volledigheid, correctheid dataset</i>	Problemen: <i>Tunnelvisie; vasthouden aan bestaande aanpak</i>	Problemen: ...	Problemen: Vast blijven houden aan oude probleemdefiniering	Problemen: ...
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: <i>Predictive policing; bestrijding fraude</i>	Voorbeelden: ...
Oorzaken: <i>Bijhouden over jaren heen; wisselende standaarden</i>	Oorzaken: <i>conservatisme in innovatie</i>	Oorzaken: ...	Oorzaken: <i>conservatisme; willen scoren</i>	Oorzaken: ...

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destillieren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Bestaande problemen: te beperkte dataset om juiste inferenties te plegen	Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: noodzaak van serieuze mens-in-de-loop.	Bestaande problemen: ...
AI gedreven oplossingen: cureren en bijeenbrengen van gestandaardiseerde datasets;	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: Niet mogelijk; noodzaak om systeem zo in te richten dat	AI gedreven oplossingen: ...

<i>oppknippen bijeengebrachte datasets in kleinere toegesneden verzamelingen</i>			<i>resultaten uitlegbaar en transparant blijven.</i>	

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
Regulatory and Ethics toolkit Global Alliance for Genomics and Health: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/				
ALLAI response on EU AI Whitepaper: https://allai.nl/wp-content/uploads/2020/06/ALLAI-Final-Analysis-of-the-EU-Whitepaper-on-AI-consultation.pdf				

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

	<i>trainen algoritme</i>			
AVG				
Consumentenbescherming				
Aansprakelijkheidswetgeving				

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ x De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: Groot verschil tussen de technische expertise noodzakelijk om bepaalde systemen te bouwen (constructieve insteek) en de kennis die nodig is om vanuit standaarden, codes of conduct en wet- en regelgeving aan te geven welke voorwaarden aan deze systemen gebesteld zouden moeten worden (reflexief-constructieve insteek).
- x ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: Discussie rond wet- en regelgeving wordt in eerste instantie binnen de eigen beroepsgroep gevoerd. Binnen een beroepsgroep zal eerst de controverse overwonnen moeten worden alvorens beginselen algemeen aanvraagd kunnen worden en het mogelijk wordt deze beginselen in andere bwoordingen toegankelijk voor derden te maken.

- ☐ x De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit: Dat is niet meer het probleem. De principes worden inmiddels veel specifiekere gemaakt en uitgewerkt. Bijv. in de medische wereld.
- ☐ x De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: Ook dat kun je niet meer zeggen. Er wordt per vakgebied onderzocht welke principes steekhoudend zijn. Probleem blijft de vertaalslag van een juridische en allesomvattende benadering naar een minder stringente verwoording waarin de kern nog steeds overeind blijft.
- ☐ x Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Nee. Data-analysten en systeemontwikkelaars werken niet in een vacuum. Waar wel behoefte aan is, is om technisch geschoolde, juridisch geschoolde en beleidsmatig geschoolde experts bij elkaar te brengen om gezamenlijk vraagstukken rond bias, discriminatie, uitsluiting etc door te akkeren en met elkaar de grenzen en voorwaarden van de toepassing te bespreken. .
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ x Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Toezicht is nog niet toegerust om tot adequaat en effectief toezicht te komen. Daarvoor ontbreekt de expertise. Met name de expertise om te zien wat in de praktijk daadwerkelijk gebeurt vraagt om een combi van technische, juridische en beleidsmatige achtergrondkennis. Dat is op dit moment moeilijk te organiseren, onder meer doordat technische specialisten eerder te vinden zullen zijn bij de partijen die de toepassingen aan het maken zijn.
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Het lijkt me goed om ook te kijken naar vormen van discriminatie die nieuw zijn en die minder snel in het oog zullen lopen, juist omdat ze zich onttrekken aan al bestaande patronen (discriminatie op huidskleur, etc.). Voorbeeld kan zijn discrimination by device: het apparaat maakt uit wat je krijgt aangeboden. Bekend voorbeeld is de iPhone en de iPad die tot andere voorgeschotelde keuzes leiden dan een 'gewone' Android telefoon of tablet. Uitgangspunt is dat een iPhone een proxy is voor bestedingsgedrag en dergelijke maar dit ontstijgt traditionele sociaaldemografische categorisering (en introduceert nieuwe). Zie ook het werk van Frederik Borgesius Zuiderveen hierover (vast en zeker bij jullie bekend).

7.3.12 Respondent 12

Enquête

Naam: Bart W. Schermer

Organisatie: Considerati / Universiteit Leiden

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Besluitvorming door een computer die normaliter (menselijke) intelligentie zou vereisen.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

-
- *Machine learning voor het herkennen van afbeeldingen*
 - *Machine learning voor detectie van afwijkende patronen (fraude detectie)*
 - *Software agents / slimme assistenten die taken uitvoeren in opdracht*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

-
- *Discriminatie (vooringenomenheid van het model op basis van trainingsdata)*
 - *Incorrecte besluitvorming (fouten in de voorspelling)*
 - *Informatie asymmetrieën (machtige partijen worden machtiger)*
-

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

-
- *Efficiency verbeteringen*
 - *Ontdekken nieuwe medicijnen / behandelingen*
 - *Inzicht in complexe problemen (o.a. klimaat)*
-

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

Ja want, de wens is er om data te verzamelen en daar inzichten uit te verwerven.

Nee want, ...

Obstakels zijn:

Acceleratoren zijn: hype (veel investeringen, hoge verwachtingen); beschikbaarheid grote hoeveelheden gegevens.

1. Probleemanalyse

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>sample (selection) bias</i>	Problemen: ...	Problemen: ...	Problemen: <i>discriminatie wordt niet herkend</i>	Problemen: ...
Voorbeelden: <i>dataset voor gezichtsherkenning getraind op blanke mensen</i>	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden:	Voorbeelden: ...

<i>herkent gekleurde mensen niet goed.</i>				
<i>Oorzaken: over/onder vertegenwoordiging van een bepaalde feature in een data set (bijv. bepaalde betniciteit)</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: moeilijk te detecteren, sluit aan bij bestaande denkbeelden en stereotypen (confirmation bias)</i>	<i>Oorzaken: ...</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: grote hoeveelheden data</i>	<i>Problemen: geheim houden algoritme</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: datasets met honderden / variabelen</i>	<i>Voorbeelden: COMPAS</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: onmogelijk voor subject om te doorgronden welke data wat representeert</i>	<i>Oorzaken: Angst voor 'gaming the system' ; bedrijfsgeheimen .</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ... gebruik complexe modellen (machine learning / deeplearning)</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: COMPAS, Husky v. Wolf probleem</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: model is te complex om te begrijpen (niet globaal interpretabel)</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

AI zou in potentie objectiever moeten kunnen zijn dan mensen. Hiervoor is dan wel de data selectie van groot belang.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

EU Guidelines on trustworthy AI > gezien de autoriteit van het document heeft dit nut op zichzelf (los van de inhoud is dit het richtinggevende document in de EU, zie ook de EU whitepaper on AI).
ICO guidelines on AI explainability > praktische handleiding
ECP AIIA (AI Impact Assessment > praktische tool voor impact assessment
Singapore Model AI governance Framework
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems > nog werk in uitvoering, maar wel gezaghebbend

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Syri uitspraak uiteraard, zie verder proefschrift Marlies van Eck voor relevante jurisprudentie in de Awb context.

116th Congress, 2019. Algorithmic Accountability Act

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkel n/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming motiveringsbegin el	(5) Feedbackloop
Gelijkheidsbegin el		gelijkheidsbegin el		

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Zie literatuur selectie aan het einde document.

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Deze vraag is wat te omvangrijk voor me ben ik bang ☺

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

X ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: zijn primair documenten voor politici en juristen, niet voor data scientists / AI experts.

X ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: Eens.

X ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: Ja, het moet concreter (wat is transparantie?). Met name aan de voorkant (risico-inschatting) gebeurt denk ik te weinig. Data scientists zijn wel degelijk bezig overigens met het voorkomen en detecteren van bias, omdat het ook de voorspellende waarde van je model ondermijnt.

X ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:

X ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Ik denk niet dat dit het geval is. Ethische overwegingen e.d. zijn vaak niet aan data scientist, maar aan degene die geautomatiseerd besluiten wil gaan nemen. Vaak zit het probleem bij de opdrachtgever (business / beleid) niet zozeer bij de data scientist die de meest effectieve oplossing moet zoeken. Er zijn wel veel documenten, maar die hebben niet een gedeeld vocabulaire. Ook wordt het niet gezien als een probleem dat je multi-disciplinair moet oppakken.

X ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Zeker. Denken is nog relatief nieuw.

X ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Er zijn nog geen afdwingbare standaarden of wet- en regelgeving. Je kan veel AI vraagstukken onderbrengen in bestaande wet- en regelgeving (Awb, AVG, etc.) maar dat is op een hoog abstractieniveau. Aldus is er geen specifieke handhaving op de zorgvuldige toepassing van AI (audits van modellen e.d.).

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Een handleiding op zichzelf is nuttig, maar idealiter moet dit worden ingebed in een breder governance proces voor AI toepassing (model risk management).

Literatuur:

- Council of Europe, 2019a. Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes, s.l.: s.n.
- Council of Europe, 2019b. Unboxing Artificial Intelligence: 10 steps to protect Human Rights, s.l.: s.n.
- Council of Europe, 2019c. Governing the Game Changer - Impacts of Artificial Intelligence Development on Human Rights, Democracy and the Rule of Law - Conclusions from the Conference. Helsinki, s.n.
- Craglia, M., ed., 2018. Artificial Intelligence - A European Perspective. In: s.l.:Joint Research Centre for the European Commission.
- Crawford, K., West, S. M. & Whittaker, M., 2019. Discriminating Systems. Gender, Race, and Power in AI, s.l.: AI Now.
- European Commission, 2019. Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions. Building Trust in Human-Centric Artificial Intelligence. Brussels: s.n.
- European Group on Ethics in Science and New Technologies, 2018. Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems, s.l.: s.n.
- European Parliament, 2017. European Parliament resolution of 16 february 20167 with recommendations to the Commission on Civil Law Rules on Robotics (2014/2103(INL)). s.l.:s.n.
- European Parliamentary Research Service, 2019. A governance framework for algorithmic accountability and transparency, s.l.: European Parliament.
- Floridi, L. et al., 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, Volume 28, pp. 689-707.
- IEEE, 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2., s.l.: IEEE.
- Information Commissioner's Office, 2017. Big data, artificial intelligence, machine learning and data protection, s.l.: s.n.
- ISACA, 2019. Auditing Artificial Intelligence, s.l.: s.n.

Joint Research Centre, 2018. Artificial Intelligence. A European Perspective. s.l., European Commission.

Kemper, J. & Kolkman, D., 2019. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*.

Kim, B., Khanna, R. & Koyejo, O., 2016. Examples are not enough, learn to criticize! criticism for interpretability.. *Advances in Neural Information Processing Systems*.

Kroll, J. et al., 2016. Accountable Algorithms. *U. Pa. L. Rev.*, Issue 165, p. 633.

Kusner, M., Loftus, J., Russell, C. & Silva, R., 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems*.

Loi, M. & Ferrario, V. E., 2019. Transparency as design publicity: explaining and justifying inscrutable algorithms.. s.l.:s.n.

Miller, T., 2017. Explanation in artificial intelligence: Insights from the social sciences. s.l.:s.n.

Mittelstadt, B. D. et al., 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*.

OECD, 2019. Recommendation of the Council on Artificial Intelligence, s.l.: s.n.

PDPC, 2018. Discussion Paper on Artificial Intelligence (AI) and Personal Data - Fostering Responsible Development and Adoption of AI. Singapore, s.n.

Reisman, D., Schultz, J., Crawford, K. & Whittaker, M., 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. s.l.:AI Now.

Université de Montréal, 2018. Montreal Declaration for a Responsible Development of AI, s.l.: s.n.

USACM, 2017. Statement on Algorithmic Transparency and Accountability, s.l.: s.n.

UNI Global Union, 2018. Top 10 Principles for Ethical Artificial Intelligence. s.l.:s.n.

Wachter, S. & Mittelstadt, B., 2018. A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI.. *Columbia Business Law Review*.

Wachter, S., Mittelstadt, B. & Floridi, L., 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation.. *International Data Privacy Law*, Volume 7.2.

Wachter, S., Mittelstadt, B. & Russell, C., 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2).

Wallach, W. & Marchant, G., 2018. An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics. s.l.:s.n.

7.3.13 Respondent 13

Enquête

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Van AI is sprake als er een geautomatiseerde beslissing wordt genomen op basis van regels die niet vooraf door een mens zijn geprogrammeerd.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

*Gezichtsherkenning.
Zelfrijdende auto's.
Beoordelen kandidaten op geschiktheid voor een functie*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Te veel vertrouwen in de correctheid van de geautomatiseerde beslissingen.
Niet kunnen achterhalen op basis waarvan beslissingen genomen zijn.*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

*Snelheid waarmee beslissingen genomen kunnen worden.
Objectiviteit van beslissingen.
Transparantie van beslissingen.*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, kosten wegen steeds beter op tegen de baten.

Acceleratoren zijn: concurrentie. Als de concurrent sneller/goedkoper is, moet meegegaan worden in de ontwikkeling.

Nee want, er is nog steeds een groot wantrouwen jegens AI.

Obstakels zijn: negatieve ervaringen in de media.

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvormin g	(5) Feedbackloop
Problemen: te weinig diversiteit in trainingsdata	Problemen: keuze voor succescriteria/variabelen. Die kunnen (indirect) correleren met andere kenmerken van kandidaten (bijvoorbeeld geslacht)	Problemen: ...	Problemen: ...	Problemen: ...
Voorbeelden: systemen die getraind worden op “successen uit het verleden”.	Voorbeelden: Amazon is het bekendste voorbeeld. Daar kregen mannelijke kandidaten een hogere score dan vrouwelijke kandidaten.	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...

Oorzaken: weinig/geen alternatieven	Oorzaken: te weinig oog voor deze proxies	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Ik vind dit lastig te beantwoorden in de onderstaande matrix vandaar dat ik een citaat uit een rapport van TNO toevoeg:

“Het recruitmentproces is ook zonder gebruik van de besproken recruitment technologieën vaak niet volledig transparant (en dit is naar alle waarschijnlijkheid niet altijd haalbaar of wenselijk). De ‘nieuwe generatie’ recruitment technologieën kan echter nieuwe uitdagingen opleveren als het gaat om transparantie van het proces. Zo is het vaak niet duidelijk welke online gegevens gebruikt worden voor de zoektocht naar kandidaten die uitgenodigd worden om te solliciteren. Het is niet duidelijk welke conclusies op basis van de online gegevens worden getrokken en het is al helemaal niet mogelijk deze conclusies te weerleggen. Het kan bovendien lastig zijn een ‘digitaal imago’ te veranderen.

Ook het proces van selectie in latere fasen van het recruitment proces is voor kandidaten vaak niet transparant. Ze kunnen de besluitvorming m.b.v. online assessments beperkt controleren en hebben veelal geen toegang tot achterliggende modellen. Deze onduidelijkheid maakt dat discriminatie en biases moeilijk aan te tonen zijn.”

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: ...	Problemen: ...	Problemen: ...	Problemen: ...	Problemen: ...
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: mensen hebben vaak onbewust bepaalde vooroordelen.</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: mensen beslissen vaak op gevoel.</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: computers hebben geen last van vooroordelen.</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: beslissingen op basis van objectieve criteria.</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

<https://www.rijksoverheid.nl/documenten/rapporten/2019/10/08/tk-bijlage-over-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid>

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Lastig om dit te specificeren binnen het geschetste proces. Het heeft betrekking op alle fases.

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

De arbowet en de Waadi zullen gewijzigd worden. Onderdeel daarvan is dat gebruikers van geautomatiseerde systemen bij werving en selectie zich ervan moeten vergewissen dat die systemen geen discriminatie in de output met zich meebrengen. Waarschijnlijk gaat het wetsvoorstel begin juli naar de Raad van State.

Vindplaats internetconsultatie:

https://www.internetconsultatie.nl/wet_toezicht_discriminatievrije_werving_en_selectie

Daarnaast is er de AVG. Bij inzet van geavanceerdere digitale middelen kan art 22 AVG van belang zijn. Dit artikel geeft betrokkenen de mogelijkheid om menselijke tussenkomst te eisen indien er sprake is van geautomatiseerde besluitvorming (waaronder profilering). Een voorbeeld hiervan is een eerste selectie van kandidaten waarbij de ondervraging van de kandidaten volautomatisch verloopt en de resultaten daarvan volautomatisch tot een profiel van de kandidaat worden uitgewerkt. Als een volautomatisch gecreëerd profiel de grondslag vormt van de beoordeling van de kandidaat zal de AVG dit niet snel toestaan.

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Zie hierboven.

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Zover reikt mijn kennis niet

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit:

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit:

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

7.3.14 Respondent 14

Enquête

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Genereren van inzicht op een ongestructureerde wijze en op ongestructureerde input.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Zelfsturende auto, beursalgorithms, selectie medewerkers

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Het geloven van de uitkomst. De input (leermodules) zal altijd fouten bevatten. De mens verschuilt zich achter de machine.

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Inzicht krijgen in complexe materie/structuren

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, snel en goedkoop. In 80% van de gevallen slim genoeg.

Acceleratoren zijn: commerciële toepassing

Nee want, het systeem maakt fouten en blijft fouten maken

Obstakels zijn: acceptatie dat het systeem het fout kan hebben.

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>selectieve set</i>	Problemen: ...	Problemen: ...	Problemen: <i>geloof in de uitkomsten. Geen weerwoord mogelijk</i>	Problemen: <i>nieuwe input kan leiden tot nieuwe uitkomsten. Iedere keer weer opnieuw valideren</i>
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: AI is niet te begrijpen., Is een blackbox</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destillieren uit data</i>	<i>(4) Besluit- en beleidsvorming</i> Gebruik AI als second opinion.	<i>(5) Feedbackloop</i>
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

nee

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

7.3.15 Respondent 15

Enquête

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Zelflerende software

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Gezichtsherkenning, onderwerp herkenning op basis van tekst

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Er worden vooroordelen aangeleerd, beslissingen zijn niet meer te reconstrueren, derden kunnen het algoritme manipuleren

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Efficiencywinsten, minder fouten, details herkennen die een mens niet kan herkennen.

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, aantal fte neemt af, terwijl overheidstaken toenemen.....

Acceleratoren zijn: data komt steeds meer beschikbaar...

Nee want, beslissing en handelingen moeten herleidbaar zijn voor de (civiele en bestuurs)rechter

Obstakels zijn: het herleidbaar zijn van

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>wat is de kwaliteit van de data, wie kiest de data en op basis waarvan, wordt bestaande data gebruikt (en op basis waarvan is die tot stand gekomen).....</i>	Problemen: <i>degene die de AI programmeert kan zijn / haar eigen (tunnel)visie. De programmeertaal is zwart/wit, terwijl de werkelijkheid dat niet is.</i>	Problemen: ... <i>Er worden verkeerde relaties gelegd.</i>	Problemen: ... <i>Het model is slechts een beperkte weergave van de complexe buiten wereld en houdt geen rekening met andere gevolgen</i>	Problemen: <i>De juiste uitkomsten zijn niet bekend en onjuiste uitkomsten worden doorgegeven, zodat het algoritme onjuistheden leert</i>
Voorbeelden: Digitale Opkopers Register, welke fietsenhandelaar wordt op welk moment gecontroleerd	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden Op basis van de scores van de zelfredzaamheid matrix wordt geadviseerd over de zorg van een burger. Ten onrechte kan worden beslist dat deze burger alleen maar reguliere zorg krijgt en geen maatwerk, met als gevolg dat de burger verder	Voorbeelden: ...

			afglijdt, extra en langere zorg noodzakelijk is en de maatschappelijke kosten toenemen	
<i>Oorzaken: van fietshandelaren die nog nooit zijn gecontroleerd zijn ook geen gegevens bekend</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Als het om bestuursrechtelijke besluiten zit, zit dat de gehele keten. Een besluit moet wel overwogen zijn en goed gemotiveerd /onderbouwd.

Maar ook bij handelingen (zonder rechtsgevolg) speelt dit, zie bijv onderstaand artikel uit het parool.

de eigen toepassingen en algoritmes. "Zo vragen we ons af of de zandbureaus waarmee we zaken doen, voldoende transparant zijn bij de selectie van medewerkers."

Ook wordt kritischer gekeken naar de eigen algoritmes. "We onderzoeken of daar vooroordelen in verborgen zitten," zegt Baron. Dat bleek inderdaad het geval bij de verwerking van klachten. "Iedereen die een melding doet bij Amsterdam.nl wordt beoordeeld door een algoritme. We zien dat sommige mensen 'beter' klagen dan anderen. Dat heeft te maken met hoe het algoritme woorden weegt. 'Klacht' weegt in het algoritme zwaarder dan 'teringzooi'. Het ene woord wordt in Zuid vaker gebruikt, het andere in Noord. Sommige mensen weten nu welke woorden ze moeten gebruiken om het beste resultaat te halen. Dat zou niet moeten uitmaken."

Ook test de stad een algoritme waarmee de naleving van regels voor vakantieverblijf van woningen in kaart wordt gebracht. Zo is het in Amsterdam alleen toegestaan om woningen maximaal 30 dagen per jaar via sites als Airbnb

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

<http://www.stibbeblog.nl/all-blog-posts/environment-and-planning/ongevraagd-advies-raad-van-state-normering-van-geautomatiseerde-overheidsbesluitvorming/>

<http://www.stibbeblog.nl/all-blog-posts/environment-and-planning/faq-wat-zijn-de-gevolgen-van-de-avg-voor-geautomatiseerde-besluitvorming-door-de-overheid/>

<http://www.stibbeblog.nl/all-blog-posts/environment-and-planning/afdeling-bestuursrechtspraak-formuleert-toetsingskader-voor-geautomatiseerde->

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk.

Het lijkt mij verstandig dat de doelgroep veel groter wordt. Techneuten zijn geen beleidsmakers, en beleidsmakers zijn geen (bestuursrecht)juristen. Maar het (bestuurs)recht beschermt de burger wel tegen geautomatiseerde besluiten en tegen discriminatie. Mijn ervaring is dat deze drie beroepsgroepen weinig van elkaars normen en regels weten.

Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

[Ja, zie ook mijn opmerking hiervoor.](#)

Ja Nee

X ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: [het gaat niet alleen om bekend zijn met, maar vooral dat techneuten /programmeurs niet bekend zijn met de regels die voor overheden gelden. Ze hebben niet het gevoel voor een overheid te werken, maar aan software te werken.](#)

X ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: [zie hiervoor](#)

- X ☒ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit: Ik vraag me af of het helpt om alles in documenten vast te leggen. Het lijkt mij veel efficiënter om een multidisciplinair team samen te stellen, waarbij het (bestuurs)recht wordt geborgd. Het zou ook helpen als de programmeurs ondervinden wat de consequenties van hun werk is (bijv. kruip in de huid van een dakloze).
- X ☒ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- X ☒ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- X ☒ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- X ☒ ☐ Anders, namelijk: zie al mijn opmerkingen hiervoor.....

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Het is hele een uitdaging om met een handreiking te bereiken dat er niet meer wordt gediscrimineerd en dat geautomatiseerde besluiten voldoen aan de Awb. Ik wens u veel succes met die uitdaging en kijk uit naar het eindproduct

7.3.16 Respondent 16

Enquête

Naam: Daniel Kapitan
Organisatie: freelance data scientist, werkzaam voor o.a.
gemeenten, provincies, zorginstellingen en software leveranciers
zie <https://linkedin.com/in/dkapitan> voor details

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

In mijn werk hanteer ik de definitie van Russel & Norvig, A Modern Approach to AI. Zij maken hierin onderscheid tussen thinking - acting enerzijds, en humanly - rationally anderzijds. Wat in de praktijk vaak bedoeld wordt met AI is 'acting rationally' en dit is de scope die ik in het dagelijks werk hanteer. Daarbij vind ik het belangrijk om onderscheid te maken tussen Narrow AI vs Artificial General Intelligence (AGI). Dat laatste is mijns inziens hyped en het is maar zeer de vraag of we dat met huidige technieken kunnen. Dus samengevat: AI zie ik als narrow AI, waarbij we middels algoritmes agents kunnen bouwen die rationeel acteren om een bepaalde taak uit te voeren.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

- Autonomous vehicles (drones, autos)
- Automatische classificaties / triage / beoordeling van allerlei bedrijfsprocessen
- Textmining

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

- Te hoge verwachtingen: narrow AI vooral goed inzetbaar wanneer de taak die geautomatiseerd moet worden zo specifiek mogelijk is gedefinieerd
- Misverstand tussen causaliteit en associatie: eeuwenoude valkuil van statistiek, is juist nu relevant in het kader van explainable en responsible AI
- Te weinig observationele data die feitelijk vastlegt wat er is gebeurd: bijvoorbeeld in de zorg

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

- Kwaliteit van beslissingen verbeteren
- Verminderen van administratieve taken
- Verbeteren van dienstverlening

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

- Ja, voor commerciële sectoren zal het steeds meer op grote schaal worden toegepast. Acceleratoren zijn personaliseren van dienstverlening aan klanten, en realiseren van operationele efficiency
- Ik twijfel wat betreft non-profit instellingen (zorg, overheid), omdat er inmiddels zoveel obstakels zijn opgeworpen (wet- en regelgeving, publieke debat) die weinig ruimte laten voor experimenten

1. Probleemanalyse

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Aanleggen dataset

Problemen: te weinig observationele data van dezelfde populatie die een verschillende interventie hebben gehad

Voorbeelden: bias bij gebruik van landelijke kwaliteitsregistraties waardoor ongewenst discriminatie optreedt tussen verschillende patientgroepen. Van behandelde patienten relatief veel data beschikbaar, van niet behandelde patienten te weinig

Oorzaken: registratielast, beperkingen t.a.v. data governance

Ontwikkelen/trainen algoritme:

Problemen: te weinig inzicht in vraagstukken met zogenaamde imbalanced classes (je wilt uitspraken doen over groepen die relatief weinig voorkomen in de totale populatie)

Voorbeelden: nietszeggende krantenkoppen die zeggen dat we met '95% procent kunnen voorspellen dat iemand niet dood gaat als hij/zij in het ziekenhuis komt'.

Oorzaken: gemiddelde data scientist nog weinig ervaring om goed met bias en onzekerheden om te gaan. Grote druk (commercie, publicatie) om positieve resultaten te publiceren. Meer kennis en expertise nodig om b.v. met imbalanced classification..

Patronen destilleren uit data

Problemen: te weinig onderscheid c.q. inzicht in correlatie vs. causatie

Voorbeelden: arme mensen hebben hogere kans op slechte uitkomst bij een behandeling, dus dan maar niet behandelen

Oorzaken: meer geavanceerde technieken om causaliteit te modelleren zijn complexer, vergen meer tijd.

Besluit- en beleidsvorming

Problemen: kloof tussen taal van data scientist vs. beslis/beleidsmakers/eindverantwoordelijke

Voorbeelden: computer geeft bepaalt advies, arts begrijpt advies niet en handelt als vanouds

Oorzaken: te weinig bekendheid, vertrouwen in nieuwe techniek

Feedbackloop

Problemen: relatief weinig algoritmen in productie, en diegene die in productie zijn genomen hebben vaak geen voorziening om makkelijk geupdatet te worden

Voorbeelden: decision-support systemen in de zorg zijn soms 20 jaar oud, effectiviteit wordt niet meer gemeten

Oorzaken: kostbaar (mensen, techniek, kennis) om de feedbackloop goed in te richten

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: bewerkelijk om vast te leggen welke data precies gebruikt is voor het model	Problemen: ...	Problemen: zelfde als 1.1	Problemen: zelfde als 1.1	Problemen: zelfde als 1.1
Voorbeelden: burgers weten niet dat hun gezondheidsdat a in landelijke registraties staan	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: bestaande governance en structuren zijn verouderd	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: vertrouwen om data beschikbaar te stellen	Problemen: ...	Problemen: ...	Problemen: zelfde als 1.1	Problemen: zelfde als 1.1

Voorbeelden: zie <i>nieuwe initiatieven rondom data delen</i> (o.a. <i>hhdc.nl</i>)	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: <i>slechte ervaringen met Big Tech</i>	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: <i>mensen hebben zelf veel biasen, en weinig manieren om hun hier vanaf 'te helpen'...</i>	Bestaande problemen: <i>mensen zijn geen rational agents, en nemen lang niet altijd rationeel beslissingen</i>	Bestaande problemen: ...
AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: <i>bayesian belief networks, modeleren van couterfactuals</i> <i>Zie argumentatie the Book of Why, Judea Pearl</i>	AI gedreven oplossingen: <i>meer inzicht geven wat rationele keus zou zijn (op basis van AI model) vs. de werkelijke keuze</i>	AI gedreven oplossingen: ...

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

- nieuwe richtsnoeren vanuit VWS en EU. Ik vind deze zodanig generiek dat ik er weinig mee kan
- best bruikbare standaarden komen uit onderzoek en open source, bijvoorbeeld the Turing Institute (UK): <https://the-turing-way.netlify.app/welcome>

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

The Turing Way geeft richtlijnen op alle punten in het proces, zie de website

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

- niet van op de hoogte

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

- nvt

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub

bron

toe)?

- ik probeer vooral wetenschappelijke literatuur, met name toegepast onderzoek, te volgen
- op dit moment geen concrete suggesties

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
opheffen van data monopolies. Derde partijen toevang verschaffen tot dezelfde data zodat zij resultaten kunnen reproduceren	publiceren van de broncode	publiceren van evaluatie van resultaten, waarbij uitgebreide toelichting wordt gegeven op zaken als generaliseerbaarheid, bias-variance trade-off, onzekerheid in het model	publicatie c.q. voorlichting wanneer een algoritme wordt gebruikt. B.v. 'uw aanvraag wordt automatisch verwerkt. Ga naar www voor een toelichting hoe deze wordt behandeld'	in lijn met AVG proces om bezwaar te maken veel toegankelijker maken

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Nee: er is meer doorzetmacht nodig, b.v. vanuit de overheid om	Idem: meer actieve rol van de overheid om zaken af te dwingen	Idem	Idem	Idem

datamonopolies te doorbreken				

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- X ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: ik ken ze niet, en als ze er zijn lijkt het mij een goede zaak om deze breder bekend te maken. Wel een aandachtspunt qua bruikbaarheid (zie eerdere opmerking)
- X ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: dit ervaar ik zo. Ik zou b.v. graag zien dat naast beleidsdocument ook voorbeeld code wordt gedeeld. Dan weet ik beter hoe ik dat in mijn werk kan integreren.
- X ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid

tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit: helemaal eens, zie ook hierboven. Graag voorbeeldcode of concreet uitgewerkte casussen, en niet alleen 'vage richtlijnen' (niet disrespectvol bedoeld overigens).

- X ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: het ontwikkelen van een algoritme is heel domein-specifiek en vergt een veelheid aan keuzes (b.v. waardes van hyperparameters). Ook hier: graag voorbeelden of misschien zelfs referentieimplementaties.
- ☐ X Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: ik ken er niet zoveel
- X ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: wellicht kan de overheid een rol spelen door met 'special interests groups' casussen te ontwikkelen zodat het concreter wordt ingevuld
- ☐ X Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: ik zou in eerste instantie niet op handhaving inzetten, maar het concreter maken van wat we willen.
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

Ik ben een groot voorstander van een open source aanpak en open data aanpak in alles wat we doen op het gebied van AI in het publieke domein. Waarom zou elke gemeente hetzelfde fraude-detectie algoritme ontwikkelen? Door data te delen en algoritmes te delen kunnen we veel verder komen. Ook in de aanpak van non-discrimination.

7.3.17 Respondent 17

Enquête

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Een digitaal systeem wat het menselijk brein tracht te evenaren of overtreffen. In het systeem moet een zelflerende component zitten en het moet autonoom kunnen opereren.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

-
- 1) Persoonlijke assistenten (siri, alexa etc)*
 - 2) Zelfrijdende auto's*
 - 3) Zorgrobots*
-

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

-
- 1) Discriminatie door vooringenomenheid (ook proxie discriminatie)*
 - 2) Gebrek aan transparantie door black box algoritmes. Niet kunnen uitleggen hoe je tot uitkomsten komt.*
 - 3) Privacyschendingen*
-

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

-
- 1) We worden productiever en efficiënter*
 - 2) 'saai' werk wordt overgenomen door AI*
 - 3) Er zijn minder ongelukken/het is veiliger (zelfrijdende auto's zijn veiliger dan menselijk bestuurders)*
-

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, toezichthouders zijn zich in hoge mate aan het focussen op 1) toezicht op AI en 2) toezicht met AI Acceleratoren zijn: media aandacht en kamervragen

Nee want, ...

Obstakels zijn:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: Er zit een bias in de data	Problemen: De programmeur heeft een bias	Problemen: de profielen/categorieën zijn discriminerend.	Problemen: De uitkomsten zijn discriminerend	Problemen: de feedback loop bevestigt de bias en vergroot daardoor het discriminatoire effect
Voorbeelden: een historische dataset van succesvolle nieuwe werknemers op	Voorbeelden: Twee chinese onderzoekers presenteerden een algoritme waar ze met	Voorbeelden: Obv postcodes en andere proxies komt iemand in een hoge	Voorbeelden: Oost-Europeanen krijgen gemiddeld sneller een	Voorbeelden: een algoritme dat moet beoordelen of gevangenen of verlof mogen

een HR afdeling heeft een vooringenomenheid tov blanke witte mannen (voorbeeld van Amazon)	89,5% zekerheid een crimineel konden identificeren obv de vorm van hun gezicht.	risicocategorie voor een lening	belastingconrole dan blanke Nederlanders	heeft een vooringenomenheid tegen zwarte personen. In de feedback loop zijn daarom onevenredig veel zwarte mensen vertegenwoordigd, waardoor het algoritme nog discriminerender wordt.
Oorzaken: Blanke witte mannen, nemen blanke witte mannen aan. We nemen graag mensen aan waarin we ons herkennen.	Oorzaken: Frenologie is een zeer discutabele methode. Hier kunnen dus mensen als crimineel worden aangemerkt op basis van de vorm van hun gezicht zonder dat er wetenschappelijk bewijs is dat frenologie echt werkt.	Oorzaken: Ligt bij problemen 1 en 2	Oorzaken: Ligt bij problemen 1 en 2	Oorzaken: Ook hier ligt de oorzaak in principe bij problemen 1 en 2.

Zo presenteerden twee Chinese onderzoekers in een studie³⁹⁹ een algoritme dat met 89,5% zekerheid een crimineel kon identificeren op basis van de vorm van zijn gezicht.

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

³⁹⁹ X. Wu, X. Zhang, 'Automated Inference on Criminality using Face Images, Computer Vision and Pattern Recognition, 2017.

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: niet transparant welke data wordt gebruikt	Problemen: Niet transparant welke beslissingen/aannames worden gedaan.	Problemen: Onduidelijk hoe algoritme werkt (black box)	Problemen: Onduidelijk hoe tot een besluit gekomen is	Problemen:
Voorbeelden: Google search engine	Voorbeelden: De keuze voor bepaalde variabelen en parameters. Waarom zegt een postcode iets over of iemand wel/niet een lening terug kan betalen?	Voorbeelden: SyRi	Voorbeelden: SyRi	Voorbeelden: ...
Oorzaken: Ontwikkelaars houden de kaarten liever tegen de borst. Geheimhouding voor bedrijfsbelang	Oorzaken: Vaak zijn mensen zich zelf niet bewust van wat hun aannames zijn.	Oorzaken: Het algoritme is te ingewikkeld. Of er zijn overwegingen rondom de IP	Oorzaken: met name door probleem 1 en 3	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: Privacyschending, doelbindingsloze informatieverzameling en doelbindingsloze	Problemen: ...	Problemen: Correlatie causaliteitsproblemen	Problemen: Een AI kan door gebreken niet goed functioneren	Problemen: Interactie met andere algoritmes kan tot onwenselijke

informatieverwerking			in het menselijke domein	uitkomsten leiden
Voorbeelden: Medische gegevens worden gebruikt om te voorspellen of iemand suïcide pleegt	Voorbeelden: ...	Voorbeelden:	Voorbeelden: ongelukken bij zelfrijdende auto's, verkeerde medicatie geven bij zorgrobots	Voorbeelden: Flash crash Collusie (twee algoritmes spannen samen in een kartel om de prijzen te verhogen)
Oorzaken: Door de grote hoeveelheden data waarmee gewerkt wordt is het moeilijk vast te stellen of je wel/niet compliant bent met de AVG/GDPR.	Oorzaken: ...	Oorzaken: Door de grote hoeveelheid data kunnen er 'toevallige' correlaties in de dataset zitten die niks zeggen over causaliteit.	Oorzaken: De AI is niet goed getraind. Er kunnen zich situaties voordoen die de AI toepassing nog nooit heeft gezien of kan overzien.	Oorzaken: We weten niet hoe algoritmes op elkaar inwerken. Soort van interactieve black box

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: Een toezichthouder besluit obv meldingen een inspectie te doen. Deze meldingen hoeven niet direct iets te zeggen over het (integere) handelen van deze organisatie.	Bestaande problemen: ...
AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: Een AI zou een neutrale analyse van de gedragingen van de onder toezicht staande instelling kunnen geven. De AI kan zeggen waar we wel/niet	AI gedreven oplossingen: ...

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

De ethische data assistent (<https://dataschool.nl/deda/>)

AI impact assessment (<https://ecp.nl/jaarcongres/artificial-intelligence-impact-assessment/>)

Ethische richtsnoeren vd Europese comissie (<https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>)

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Eerlijkheid	Eerlijkheid		Eerlijkheid	
			Uitlegbaar	
Controleerbaar	Controleerbaar	Controleerbaar	Controleerbaar	Controleerbaar
		Betrouwbaar		
Verantwoording	Verantwoording	Verantwoording	Verantwoording	
		Consistent		
Privacy				

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

We volgen de ontwikkelingen van de EU op de voet. We hopen op een GDPR achtig wetgevingskader voor AI. Ook zijn we aan het kijken hoe onze bestaande bevoegdheden zich verhouden tot AI.

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
			Rechtvaardigheid	
			Redelijkheid	
			Billijkheid	

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

AFM/DNB <https://www.afm.nl/nl-nl/nieuws/2019/jul/verkenning-ai-verzekeringssector>

DNB

https://www.dnb.nl/binaries/General%20principles%20for%20the%20use%20of%20Artificial%20Intelligence%20in%20the%20financial%20sector_tcm46-385055.pdf

German Research institute for public administration:

https://www.vzbv.de/sites/default/files/downloads/2019/07/19/martini_regulatory_system_algorithm_based_processes.pdf

WRR [https://www.wrr.nl/onderwerpen/big-data-privacy-en-](https://www.wrr.nl/onderwerpen/big-data-privacy-en-veiligheid/documenten/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving)

[veiligheid/documenten/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving](https://www.wrr.nl/onderwerpen/big-data-privacy-en-veiligheid/documenten/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving)

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
	Ethiek			
Skills	Skills			

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan

waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
			Controle	Onafhankelijkheid
			Betekenisgeving	

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: vaak zijn de data/computer scientists niet zo van het lezen, maar van het doen.
- ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit:
- ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: klopt, hoe vertaal je een principe tot een regel, die weer in een programmeertaal in het algoritme geprogrammeerd moet worden.

- ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:

- ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: absoluut het grootste probleem. Dit zie je al ontstaan binnen het zorgdomein.

- ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:

- ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Toezichthouders beginnen zich nu pas te verhouden tot AI. Pas als duidelijk is waarop toezicht en handhaving zich zal richten zullen onder toezicht gestelden zich tot AI principes gaan verhouden.

- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Een GDPR achtige AI wetgeving van de EC is wenselijk. Verder is het belangrijk per domein te kijken welke principes belangrijk zijn. In de zorg is privacy bijvoorbeeld belangrijker dan in de financiële sector. Het zou helpen om consolidatie te brengen; dus evt alle stukken door gaan en dan de gemeenschappelijkheden in 1 document presenteren.

7.3.18 Respondent 18

Enquête

Naam: Evert F. Stamhuis

Organisatie: Erasmus School of Law, Jean Monnet Centre of Excellence Digital Governance

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Elke vorm van algoritme-toepassing waarin leereffecten automatisch worden ingelijfd in het functioneren van het algoritme.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Fraudebestrijding, autonome mobiliteit, predictive public administration

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Bugs in de code; data corruption (in training data of use data); te groot vertrouwen in de eigenaren (tech giants)

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Meer detaillering mogelijk ten opzichte van rule-based systemen voor beslissingen; voorspellende kracht voor netwerkeffecten (virussen, ander onheil): personalisering van services

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

*Ja want, er is een krachtige PR, het is al op gang en veronderstelde kosteneffecten zullen de komende tijd in goede aarde vallen.
Acceleratoren zijn: NL/EU wil achterstand inhalen.*

Nee want, ...

Obstakels zijn:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Het invullen van de matrices kostte me te veel tijd. Excuses.

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen:</i>	<i>Problemen: vooroordelen worden versterkt</i>	<i>Problemen: data patroon wordt als intentioneel gedrag gezien</i>	<i>Problemen: onaanvaardbare factoren tellen mee</i>	<i>Problemen: systeem kan feedbackloop manipuleren</i>
<i>Voorbeelden:</i>	<i>Voorbeelden: parole prediction US</i>	<i>Voorbeelden: toeslagen affaire</i>	<i>Voorbeelden: sociale status, gender</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken:</i>	<i>Oorzaken: garbage in, garbage out</i>	<i>Oorzaken: systeem kijkt niet achter de data</i>	<i>Oorzaken: model ziet correlaties zonder ethiek/recht</i>	<i>Oorzaken: machine leert</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: systeem weet</i>	<i>Problemen:</i>	<i>Problemen: ...</i>

		<i>zelf niet wat wij willen weten</i>		
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Brief Minister van rechtsbescherming; HLEG guidelines for responsible AI; DNB richtlijnen; beleidsbrief BZK, AI, publieke waarden en mensenrechten, zie ook [Acatech.de/themen/digital-selbstlernend](https://acatech.de/themen/digital-selbstlernend)

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Aerius jurisprudentie Raad van State;

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit: dit kan ik niet overzien, maar mijn indruk is dat de techneuten wel interesse hebben, maar ook focus op hun eigen opdracht.

x ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: Het bekende eigentaal fenomeen, dat verder gaat dan jargon! Er zijn concepten die per discipline van inhoud verschillen en zelfs binnen een discipline variëren. 'Vertalen' zou wel eens ontoereikend kunnen blijken.

x ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: Zonder context zijn high-over principes holle frasen. Die moeten nog inhoud krijgen en eventuele spanningen tussen principes moeten nog opgelost worden, met policy keuzes in de specifieke context. Daar moeten de effecten van de keuzes op de relevante belangen voorzien en afgewogen worden.

- X ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: zie vorige item
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Wees erop bedacht dat onderscheid maken vaak juist wel moet om een juiste toepassing van wet- en regelgeving te realiseren. Essentieel is dat gelijkheid soms wel en soms niet rechtvaardig is. Daar zit een waardeoordeel achter, per sector of wet variërend. Het heeft de voorkeur dat een brede consensus over welke factoren mee mogen tellen en welke niet in een systeem wordt ingebouwd. De gedachte dat we dit even in een handreiking voor de systeembouwers kunnen aieten is techno-optimisme.

7.3.19 Respondent 19

Enquête

Naam: Bram Klievink

Organisatie: Universiteit Leiden

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Software, systemen of machines die in staat zijn taken uit te voeren en daarbij op een of andere wijze opgaat met of leert van de omgeving.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Gezichtsherkenning; fraude-detectie; patroonherkenning

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Te groot vertrouwen van menselijke actor op de output van AI;
(onopgemerkte) biases; gebrek aan transparantie*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Automatiseren van taken; ondersteuning van wetenschappelijk onderzoek met grote datasets; bevorderen effectiviteit van organisaties

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, ...

Acceleratoren zijn: ...

*Nee want, we zitten nu tegen de top van de hype cycle. De verwachtingen van AI zijn nu overspannen en het werkelijk waardevol gebruik blijft achter.
Obstakels zijn: data kwaliteit, gebrek aan geschikte toepassingen.*

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

[Ik denk dat de feedbackloop ook in eerdere stappen al voorkomt]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: <i>dataset biased by default</i>	Problemen: <i>onvermijdbare keuzes in selectie algoritmen, of in training set</i>	Problemen: <i>veronderstellen causale relatie waar alleen correlatie is</i>	Problemen: <i>selectief shoppen in resultaten</i>	Problemen: ?
Voorbeelden: <i>alleen probleemgevallen worden geregistreerd</i>	Voorbeelden: <i>filteren en opschonen van data</i>	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: <i>we verzamelen en bewaren data met bepaald doel, wat het niet perse geschikt maakt voor AI</i>	Oorzaken: <i>vaak onvermijdelijke stap met vele kleine beslispunten</i>	Oorzaken: <i>dit is de aard van AI systemen</i>	Oorzaken: <i>logica van besluitvorming (zie ons GIQ paper voor uitgebreid argument)</i>	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: gevoelige of vertrouwelijke data</i>	<i>Problemen: onbegrijpelijkheid van algoritme voor non-expert (en soms ook voor experts)</i>	<i>Problemen: We weten vaak niet wat de machine precies doet om tot uitkomst te komen</i>	<i>Problemen: besluitvorming kan transparant, maar als de input ervan deels black box is, redt je het in dit stadium niet meer dat open te krijgen</i>	<i>Problemen: ...</i>
<i>Voorbeelden: patroonherkenning bij misdaad</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: fraudedetectiesystemen</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: data kan niet worden gedeeld</i>	<i>Oorzaken: technische expertise nodig om ook van transparant algoritme te begrijpen wat het doet</i>	<i>Oorzaken: zit in de aard van sommige AI systemen</i>	<i>Oorzaken: ibid.</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: datasets te groot om behapbaar te zijn voor niet AI</i>	<i>Bestaande problemen: menselijke actoren kunnen beperkt aantal factoren/variabelen meenemen</i>	<i>Bestaande problemen: ook menselijke actor kent bias</i>	<i>Bestaande problemen: veel ruimte voor 'politiek'</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: meer data kan worden meegenomen</i>	<i>AI gedreven oplossingen: kan veel meer meenemen en nog onbekende factoren helpen identificeren</i>	<i>AI gedreven oplossingen: niet perse een oplossing maar je bent dus niet perse slechter af.</i>	<i>AI gedreven oplossingen: zou theoretisch kunnen bijdragen aan meer evidence-based beleid (al ben ik sceptisch)</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Er zijn er van de high level expert group. Binnen Nederland merk ik dat de richtlijnen van MinJ&V door best wel wat organisaties worden geadopteerd (evt na aanpassing). Ik zit niet zo goed in de verschillende standaarden en richtlijnen, veel ervan is nog in ontwikkeling

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- beleidsvorming	en (5) Feedbackloop
FAIR (findable, accessible, interoperable, reusable)	Betrekken domein- experts	Moet aanvechtbaar zijn als besluit erop is gebaseerd	Volgens mij vaak al in bestaande kaders vervat, maar extra uitdagingen op gebied van verantwoording/motivering beslissingen	

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destillieren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destillieren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destillieren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: dit weet ik niet

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: of het ze weinig zegt weet ik niet, maar het zal weinig handelingsperspectief bieden, maar dat is eigenlijk het volgende punt.

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- ☒ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: denk dat de veelheid vooral het beeld geeft dat er geen overeenstemming over is
- ☒ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: de discipline is zich nog aan het vormen
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.3.20 Respondent 20

Enquête

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Software die in staat is zichzelf aan te passen, om taken optimaal uit te kunnen voeren.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Recommender systemen, prijsoptimalisatie, facial recognition, zelfrijdende auto's

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Datadiscriminatie, verlies van solidariteit, onbegrijpelijke uitkomsten, financieel bedrijfsrisico ("garbage in garbage out")

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Producten en diensten die beter aansluiten op klantbehoefte (preventie, schadeafwikkeling), mogelijkheid om klanten laagdrempelig te adviseren, automatisering van allerlei administratieve taken voor bedrijven en consumenten. Kostenreductie zowel voor de klant als voor het bedrijf, snelheid en zekerheid. Emotie speelt geen rol

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, er is veel waarde mee te genereren.

Acceleratoren zijn: hoge internetdichtheid NL, goed algemeen opleidingsniveau, NL AI coalitie en de bereidheid daaraan deel te nemen

Nee want, er is veel wantrouwen

Obstakels zijn: niet optimale samenwerking publiek-privaat, uitlegbaarheid

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvormi ng</i>	<i>(5) Feedbackloo p</i>
<i>Problemen: historische data ziet meestal op grote groepen consumenten</i>	<i>Problemen: algoritme benadeelt niet- hoogopgeleide/nie t-standaard mensen...</i>	<i>Problemen: te weinig data voorhanden ...</i>	<i>Problemen: ... besluitvormin g beïnvloedt fundamenteel de levenssfeer van klanten o.b.v. onvoldoende kwalitatieve data.</i>	<i>Problemen: ... er ontstaat een 'kromme' of verkeerde voorstelling van de werkelijkheid</i>
<i>Voorbeelden: klanten betalen meer, omdat anderen in hun</i>	<i>Voorbeelden: systeem ontwikkeld om paspoorten te</i>	<i>Voorbeelde n: profielen van klanten o.b.v.</i>	<i>Voorbeelden: ... zij worden mogelijk onverzekerd</i>	<i>Voorbeelden: ... grote groepen klanten</i>

<i>straat/postcode vaak te laat betalen</i>	<i>herkennen voor KYC proces: het systeem herkende alleen NL paspoorten...</i>	<i>willekeurig e data denken iets zinnigs over hen te zeggen (fraudeur, big spender, geïnteresseerd in bepaalde producten). ...</i>	<i>ar, onterecht bestempelt als fraudeur, bestookt met reclame voor dingen die ze niet echt willen, etc.</i>	<i>worden bestempeld als lastige klanten, slechte betalers, risicovol en roekeloos in hun doen en laten</i>
<i>Oorzaken: onvoldoende differentiatie/onvoldoende actuele data</i>	<i>Oorzaken: onvoldoende divers team van engineers/scientists ...</i>	<i>Oorzaken: ... data = magie, panacee voor oplossingen die vooral efficiëntie realiseren.</i>	<i>Oorzaken: ... makkelijker voor de bedrijfsvoering om profielen / labels / mensen in simpele risico categorieën te plotten.</i>	<i>Oorzaken: ... Wens om zicht te hebben op de 'betere' en minder goede verzekerden in de markt (laag risico portefeuille).</i>

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: opsporen en voorkomen van discriminatie is lastig, omdat veel beschermde gronden (etniciteit, geloof etc.) onbekend zijn</i>	<i>Problemen: ...</i>	<i>Problemen: klanten krijgen niet te horen waarom zij anders beoordeeld zijn...</i>

		<i>voor bedrijven: ze kunnen er dan ook moeilijk op corrigeren ...</i>		
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ... “Buurman krijgt een andere premie dan ik”</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: niemand wil discrimineren, maar er zijn niet altijd methoden voorhanden om te checken of er sprake is van onbedoelde discriminatie. Oplossing kan zijn om in de AVG een mogelijkheid te creëren om deze data wel aan te leggen, enkel met het doel om discriminatie op te sporen ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: bedrijf wil geen bedrijfsgeheime info delen & datascientists en klantcontactcentru m hebben onvoldoende contact ... Statistiek die er achter zit is voor de gemiddelde Nederlander niet te begrijpen.</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Algemeen: maatschappelijk wantrouwen jegens nieuwe techniek.

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: AVG gaat uit van dataminimalisati e. Daardoor is het</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>

<i>niet eenvoudig om voldoende trainingsdata te verzamelen, wat grote spelers zoals BigTech bevoordeelt t.o.v. kleinere spelers</i>				
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkel en/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: verdergaande differentiatie kan solidariteit in de verzekeringssector ondergraven...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: Een algoritme zou op de achtergrond discriminatoire patronen kunnen opsporen, door</i>	<i>AI gedreven oplossingen: ... Emotie speelt geen rol – uitsluiten van onbewuste beïnvloeding van besluitvor</i>	<i>AI gedreven oplossingen: het is niet AI gedreven, maar als Verbond organiseren we reeds 4 jaar lang de Solidariteitsmonitor, om te zien in hoeverre premies uit elkaar bewegen (of niet):</i> https://www.verzekeraars.nl/media/6498/solidariteitsmonitor-2019-definitief.pdf ...

		<i>uitkomsten te vergelijken met een synthetische dataset</i>	<i>ming door emotie</i>	

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

High level expert group on AI Ethics Guidelines for Thrustworthy AI
Ethisch Kader Verbond van Verzekeraars en daarop gebaseerde Ethisch Kader a.s.r.
Toezicht AP op AI e.d.; dit geeft een goede beschrijving van de mogelijkheden m.b.t. AI maar vooral ook de risico's

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
Technische robuustheid en veiligheid	Diversiteit non discriminatie en billijkheid	Technische robuustheid en veiligheid	Maatschappelijk en ecologisch welzijn	transparantie
Privacy en gegevensbeheer		fairness	Privacy en gegevensbeheer	verantwoording
Fairness			Respect voor menselijke autonomie	Menselijke tussenkomst en toezicht

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Er is voldoende wetgeving relevant voor AI in de financiële sector: wet gelijke behandeling, artikel 1 Grondwet, Wet Financieel Toezicht over 'adviseren' en 'beheerste bedrijfsvoering', Solvency II over datakwaliteit en stabiliteit van gebruikte modellen, AVG (artikel 22 profiling), Burgerlijk Wetboek

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) <i>Aanleggen dataset</i>	(2) <i>Ontwikkelen / trainen algoritme</i>	(3) <i>Patronen destilleren uit data</i>	(4) <i>Besluit- en beleidsvorming</i>	(5) <i>Feedbackloop</i>
Beschermen van persoonsgegevens (AVG)	Verbod op profilering		Zorgplicht	Informatie en mededelingsplicht
Juistheid van de data (AVG)			WFT: ☐ Tegengaan belangenverstrengeling ☐ Voorkomen van overtreden van wet, .o.a. strafbare feiten ☐ Preventie van verlies in vertrouwen financiële markten	Recht op menselijke tussenkomst
				Recht op uitleg

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

High level expert group on AI

Advies van ECP: <https://ecp.nl/publicatie/artificial-intelligence-impact-assessment-english-version/>

Advies Geneva Association over verantwoorde AI:

https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/ai_in_insurance_web_0.pdf

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
			inclusiviteit	

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☐ x ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: In de verzekeringsbranche is het gebruik van een ethisch kader per 1-1-2021 verplicht en het gebruik hiervan gaat ook getoetst worden.

☐ x De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: In het ethisch kader voor verzekeraars is de taal van de high level expert group vertaald naar voor Nederlandse verzekeraars relevante begrippen.

x ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de

principes zouden veel specifiekier moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: Niemand wil discrimineren, maar probleem is dat het lastig is om discriminatie op te sporen: bedrijven zijn niet op de hoogte van gegevens over etniciteit/seksuele geaardheid etc.. Wij overwegen een tool te ontwikkelen die de mate van discriminatie/differentiatie van verschillende modellen voor verzekeraars inzichtelijk maakt; zowel directe discriminatie (e.g. mannen betalen een hogere premie voor autoverzekering dan vrouwen, ceteris paribus) of indirecte discriminatie (e.g. men betaalt hetzelfde c.p., maar geslacht is gecorreleerd met kilometrage, daarom betaalt gemiddelde man meer dan gemiddelde vrouw). We zouden hiervoor een synthetische dataset van verzekerden kunnen aanleggen om dit te identificeren. Idee is dan dat je na het maken van een model je predict() functie meegeeft (of de vector van prijzen die uit het model rolt als je de synthetische dataset erin gooit) en dat daaruit een objectieve maatstaf van directe & indirecte discriminatie per variabele rolt. Door specifieke mutaties in de synthetische dataset kunnen we directe discriminatie identificeren: bijvoorbeeld door bij de helft van de rijen één variabele te muteren (bijv. geslacht van 'vrouw' naar 'man') – prijsverschil tussen deze twee rijen is dan een indicator voor directe discriminatie. Indirecte discriminatie is dan te bepalen middels het gemiddelde prijsverschil tussen groepen (onafhankelijk van andere variabelen) minus de eerder bepaalde directe discriminatie.

☐x ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: we zien dat de Europese verzekeringstoezichthouder nadenkt over ethische kaders voor verzekeraars en dat ook DNB/AFM hiernaar kijken: er zijn algemene richtlijnen denkbaar, maar het zal nodig zijn dit per sector te vertalen naar werkbare handreikingen. Alleen heel grote bedrijven kunnen misschien met de algemene principes werken, omdat zij voldoende capaciteit hebben om hier zelf wijs uit te worden.

☐x ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: De OECD werkt aan kaders en handreikingen, de Europese Commissie werkt hier aan, het Europees Parlement heeft een voorstel gedaan voor een verordening, toezichthouders werken hier op sectoraal niveau aan, op nationaal niveau wordt door de wetgever nagedacht: er is inderdaad erg veel. En dan heb je ook nog eens verschillende toezichthouders die vinden dat zij er over gaan: voor de financiële sector AFM/DNB enerzijds, en de AP anderzijds.

- ☒ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Bedrijven zullen geen toepassingen willen ontwikkelen die niet op maatschappelijke steun kunnen rekenen. Wij horen van leden nog niet dat zij vinden dat er te weinig regels zijn: vooralsnog vergt het implementeren en naleven van bestaande regels nog voldoende aandacht. Vraag is of meer regels helpen, of dat het gaat om het uitdokteren van hoe de bestaande regels verstaan moeten worden in het licht van de moderne technieken.
- ☒ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: De AVG stelt al regels rondom de uitlegbaarheid van automatische besluitvorming. Het zou helpen als er een toelichting kwam op artikel 22 en hoe dit zich verhoudt tot het gebruik van AI (idem voor informatieverplichtingen uit artikelen 12-14). Toch zal daarbij gewaakt moeten worden voor al te rigide uitleg, omdat hele strikte handhaving ook innovatie in de kiem kan smoren. Wij pleiten daarom voor meer gebruik van het concept 'sandboxes': plekken waarin toezichthouders en bedrijfsleven vrijelijk met elkaar van gedachten kunnen wisselen. AFM en DNB hanteren dit concept al enkele jaren, maar de AP ontbreekt in deze arena. Dat is jammer, nu veel innovaties rondom AI gepaard gaan met het verwerken van persoonsgegevens.
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Zorg voor voldoende specificiteit: verzekeraars maken onderscheid gebaseerd op risico. Dat lijkt soms net op discriminatie, maar is het niet: indirecte discriminatie is verboden, tenzij het doel legitiem is en de middelen passend en noodzakelijk. Zie bijvoorbeeld: ADVIES Aan Dazure B.V. over premiedifferentiatie op basis van postcode bij de Finvita overlijdensrisicoverzekering van het College voor de Rechten van de Mens van januari 2014

7.3.21 Respondent 21

Enquête

Naam: Ramon van den Akker

Organisatie: de Volksbank, Tilburg University (deelname enquête op persoonlijke titel)

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Vakgebied dat zich richt op het ontwikkelen van intelligente kunstmatige systemen

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Robotica, machine learning, NLP

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Te veel vertrouwen in systemen die eigenlijk niet zo intelligent staan
Effect op werkgelegenheid langere termijn
Afname van solidariteit in maatschappij agv verdere personalisering diensten

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Robotisering tbv ondersteuning zorg
Machine learning modellen ter ondersteuning van bijv. medisch onderzoek & artsen
Mogelijkheid om producten en diensten te personaliseren (denk bijv. aan medicijnen die tailor-made zijn voor een individu)

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want

- *Veel administratieve werkzaamheden lenen zich voor automatisering via robotica en/of machine learning*

Acceleratoren zijn:

- *Beschikbaarheid van tools waarmee je eenvoudig robots en AI systemen kan ontwikkelen zonder dat diepe data science en IT kennis benodigd is*

Nee want vanwege AVG en andere wetgeving alsook de onzekerheid omtrent de goede interpretatie blijven veel instellingen afwachten

Obstakels zijn:

- *data van goede kwaliteit*
- *onzekerheid over wetgeving (toekomst alsook interpretatie huidig)*
- *gedegen toepassing vraagt hooggekwalificeerde medewerkers die wellicht schaars zijn*

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: A <i>Geen aselecte steekproef</i>	Problemen: A Directe discriminatie door keuze van een variabele	Problemen: Bij gebruik van complexe modellen is het niet evident	Problemen: Te <i>veel vertrouwen in algoritmen/ AI</i>	Problemen: A <i>Indien er niet goed nagedacht wordt over</i>

uit populatie B subjectieve data	B Keuze van 'evaluation/performance' metrics C Niet goed corrigeren tbv representatieve sampling D Machine learning modellen identificeren gewoonlijk aan correlatie en geen causale relatie	welke patronen geïdentificeerd zijn. Hierdoor bestaat risico op onbedoelde indirecte discriminatie	Te weinig guidance aan data scientists voor wat wel/niet mag	feedbackloop en het model geüpdatet wordt obv nieuwe data kan er onbedoeld een negatieve spiraal ontstaan B Als er op subjectieve wijze labels worden vastgesteld welke in volgende ronde worden gebruikt voor hertrainen model
Voorbeelden: Beoordeling van medewerker door manager. Associaties in taal (bijv. dokter is in historische teksten gewoonlijk een man)	Voorbeelden: Indirecte discriminatie op geslacht middels bijv. de uitgaven aan kapper in een machine learning model voor kredietrisico obv transactiedata. Focus op het optimaliseren van de waarde voor het bedrijf wat kan leiden tot onjuiste keuze/advies voor een klant/gebruiker	Voorbeelden: Neural networks (deep learning) zijn vaak slecht uitlegbaar. In laatste jaren wel wat voortgang in academia. Maar gewoonlijk leg je alleen de geïdentificeerde correlatie uit.	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: Goede data is vaak niet beschikbaar. Voor goed herkennen van situaties is diepgaande	Oorzaken: Voor goed herkennen van situaties is diepgaande kennis van statistiek benodigd. Ook kennis van de hele keten nodig. Indien er vanuit silo's aan	Oorzaken: Geen genoeg nemen met black boxes. Gebruik van XAI.	Oorzaken: Onbewust onbekwaam Te lastig om te operationaliseren	Oorzaken: ...

<i>kennis van statistiek benodigd.</i>	<i>de schakels van de keten wordt gewerkt neemt risico vaak toe.</i>			

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ...</i>	<i>Problemen: Voor sommige toepassingen kan/wil je niet transparant zijn. Bijv. om gaming (denk aan fraude) te voorkomen, concurrentie</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen:</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>

<i>Toegang tot data AVG laat vastlegging van bijzondere persoonsgegevens niet toe zodat je deze niet kan gebruiken om te controleren/op te leggen dat een model niet onbedoeld discrimineert op deze dimensie (via een proxy effect)</i>				
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen:</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Naast bekende HLEG rapport en UU rapport 'algoritmen en grondrechten':

- ***SAFEST rapport DNB***
 - <https://ico.org.uk/about-the-ico/news-and-events/ai-auditing-framework/>
-

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Niet in EU. In VS heb je oa 'disparate impact' in arbeidswetgeving als ik het goed heb alsook de 'equal credit opportunity act'.

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
Transparantie	Bias, fairness	Bias, fairness	transparantie	Bias, fairness
Bias, fairness				

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja N
 ee



- ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit: Bij de grotere bedrijven die zich bezighouden met AI en data science zie ik dat er nagedacht wordt over het onderwerp. De Nederlandse Vereniging van Banken heeft bijv. een expert pool AI om over het onderwerp te discussiëren en de ontwikkelingen ogv wet- en regelgeving te monitoren.



- ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit: De juridische teksten zijn veel te generiek en het is niet duidelijk hoe je deze precies moet interpreteren. Veel juristen zijn zich bijv. helemaal niet bewust dat het primaire doel van veel modellen is om te discrimineren (in de zin van een onderscheid maken tussen objecten). De vervolgstap is dat gedefinieerd moet worden op welke dimensies niet gediscrimineerd mag worden en hoe we discriminatie willen meten (bijv. we kijken hoeveel vrouwen een hypotheek krijgen

vs. % voor mannen `of we kijken hoeveel vrouwen ten onrechte geen hypotheek krijgen tov mannen). In mijn ervaring lopen we in deze stap vast: onze wetgeving biedt nu te weinig houvast.

- ☒ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit:

- ☒ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:

Mijn ervaring is dat veel beleidsmakers (oa HLEG) van mening zijn dat zij geen guidance hoeven te geven over de vertaling van wetgeving naar algoritmen. Dit is echter verre van triviaal! Er zijn veel wiskundige definities van fairness en deze zijn in het algemeen strijdig. Dit speelt met name voor ‘discriminatie’ via proxy-variabelen. Zo kan het bijv. dat de ontwikkelaar van COMPAS kan zeggen dat het systeem deugt terwijl de waakhond ProPublica een fundamenteel probleem zag.

Vanuit die optiek ben ik benieuwd hoe de handreiking tot stand gaat komen en wat het detailniveau gaat zijn. Naar mijn mening zijn er al meer dan genoeg rapporten en raamwerken met hoog-over principes (uitlegbaarheid, fairness, transparantie etc.). Zaak is dus om een handreiking te ontwikkelen waarmee een data scientist uit de voeten kan. Mijn concrete vraag zou zijn: hoe kan ik als ontwikkelaar aantonen dat een model voldoet aan de wetgeving. In zekere zin vereist dit dus rule-based kaders als aanvulling op highly principle-based wet- en regelgeving.

-
- ☒ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)

Leg uit:

- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt

Leg uit:

- ☒ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes

Leg uit:

- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.3.22 Respondent 22

Survey

Name: Prof. Dr. Thomas Hoeren

Organization: Institute for Information, Telecommunication and Media Law (ITM), University of Münster, Germany

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Artificial intelligence today does not mean a conscious machine or human-like thinking and creating. Weak AI is defined as fulfilling a specific task without the direct control of a human in an automatic way. Strong AI can be considered as technically reconstructing human consciousness and human decision processes as well as human creativity.

Which three examples of (current or future) AI use come to mind first?

Smart home, parking assistant, face recognition on phones.

Which three dangers / negative sides of AI are you most concerned about?

Data protection, lack of explainability, growing dependence on machines.

Which three opportunities / positive sides of AI come to mind first?

Higher efficiency for work processes: greater work power, cost-saving; society without simple jobs can lead to more free time for everyone.

Do you think that AI applications and their use (for example by governmental organisations, companies and non-profit organizations) will increase in the near future? Why / why not - what are accelerators and what are obstacles?

1. Problem analysis

Yes because, it simplifies many processes, leads to higher efficiency and thereby greater growth expectancies and relieves society of burdensome jobs.

No because, ...

Obstacles are:

Accelerators are: ... Pandemics as the current coronavirus; Climate change; Geopolitical strategy towards the US and China.

Different phases / steps play a role in the development and use of AI systems, such as: (1) the creation of a data set, (2) the development and training of an algorithm, (3) the production of patterns, models and profiles through data analysis, (4) making decisions / policies based on these models and (5) a feedback loop in which the results are 'fed' to the self-learning algorithm. [If you envisage a different categorisation/other phases, please indicate so]

When using AI, (unintended) discrimination can occur. What do you think are the underlying causes and can you give concrete examples of these? Where (which sectors, applications, etc.) do you foresee that the problem of discrimination in AI systems will occur in particular? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: The algorithms can only work with the material and the coding, that it got. This infers structural racism inside the used training data sets. The algorithm is as discriminating as the programmer.</i>	<i>Problems: The algorithm itself advantages the majority group and disadvantages the minority group in many cases due to its inherent mathematical function.</i>	<i>Problems: If the algorithms are trained with discriminatory data, the patterns will be discriminatory too. The discrimination in step (1) will be found in every other</i>	<i>Problems: If a state/community allows the use of discriminatory AI software and does not regulate its use.</i>	<i>Problems: The learned decision model is founded on a structurally racist and sexist data set; the decisions made based on this data set are then fed in to new training data</i>

		<i>step; structural racism and sexism will be perpetuated by the model's decisions</i> <i>Furthermore, the functioning of the algorithm can lead to massive disadvantages for one group with a protected feature (such as ethnicity) without further fairness measures.</i>		<i>sets which are then again perpetuating the structural racism and sexism</i>
Examples: <i>Facial recognition software</i> (Buolamwini/Gebbru, Gender Shades (2018), Center on Privacy and Technology at Georgetown Law, The Perpetual Line-Up (2016).	Examples: <i>See the debate about Yan Le Cun's statement recently on twitter</i>	Examples: <i>COMPAS</i>	Examples: ...	Examples: ...
Causes: <i>The artificial intelligence is based on codes and data. If it only got structurally racist or sexist information, it will use this to train its decision model.</i>	Causes: <i>programing of the algorithm; lack of focus on discrimination in data science research</i>	Causes: ... <i>Lack of consideration of discriminatory behaviour of ML models</i>	Causes: ...	Causes: ... <i>Lack of non-discriminatory data sets which are really hard to establish; lack of consideration of fairness measures</i>

Transparency of AI-based policy and decision-making is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there with regard to transparency of AI processes and AI based policy- and decision-making? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: ...</i>	<i>Problems: ... Basic algorithm is often known, but hyperparameters are not; additionally weighting of neurons in deep learning models is hardly interpretable for humans.</i>	<i>Problems: Not easy to understand the decision.</i>	<i>Problems:</i>	<i>Problems:</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: Black box effect</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: The processes of AI are so complex, that they can't be reproduced and explained easily</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

What do you think are other bottlenecks with regard to AI, for example technical, social or societal obstacles/problems? Can you give concrete examples of this? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: The problems that exist in this phase will affect the whole process. If the dataset is created with discriminatory data, the whole AI model will be discriminatory.</i>	<i>Problems: A problem of ML models is their lack of precision. They are often only 70-80% precise which can be insufficient for many tasks.</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

Could AI also help reduce / remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how? Please indicate as specifically as possible where in the process these benefits can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Existing problems:</i>	<i>Existing problems: Decisions by humans always rest on their own values and expectations.</i>	<i>Existing problems:</i>	<i>Existing problems: Decisions made by human are always subjective. Even if they don't want to make discriminating</i>	<i>Existing problems:</i>

			decisions, it sometimes happens. Explainable AI solutions could lead to a better understanding of the decision than the “black box” of a human decision.	
AI-driven solutions: ...	AI-driven solutions: The AI is trained in a specific way. It will always decide the same way on the same kind of case. However, the foregoing training of the AI model can lead to different results with the same data set and the same hyper parameters (hence a heuristic and not an algorithm).	AI-driven solutions: ...	AI-driven solutions: Black Box models are highly controversial and lack explainability. However, white box models could lead a better understanding of decisions.	AI-driven solutions: ...

2. Existing Standards

2.1 Do you know of existing policy documents, standards, codes of conduct or guidelines with regard to AI? If so, which are most promising and why (please add source)?

Which principles or standards from those documents are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

- Codes of conduct by the expert group for AI, deployed by the European Commission, in March 2019
- ACM Principles for Algorithmic Transparency and Accountability
- Asilomar AI Principles
- Datenethikkommission
- FAT ML Principles for Accountable Algorithms and a Social Impact Statement for Algorithms
- IEEE P7000 series
- Tenets Partnership on AI

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
The dataset itself should not be discriminatory.	The algorithms have to be developed non-discriminating.	Accountability of the patterns.	The AI should be used for the social and political well-being.	The codes of conduct explicitly demand more fairness and transparency.

2.2 Are you aware of existing legislation or case law regarding AI? If so, which (please add source)?

Art. 15 – 22 GDPR, which are protecting the interests of the users (rights of data subjects), especially Right to Explanation and Art. 22 GDPR concerning automated decision making
District Court of the Hague, 06.03.2020, judgment on SyRI

Which principles or standards from that legislation or case law are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
Primary Use of an algorithm in only low income districts leads to unbalanced data sets.	Lack of Transparency			-Art. 15 GDPR (right to information)
				-Art. 20 GDPR (right to data portability) → concerning the transparency
				-Art. 17 GDPR (right to erasure)

Relevant case law (SyRI case, ruled by the District Court of the Hague):

The Dutch government has been secretly using an algorithms (called SyRI), to detect possible social welfare fraud

The lack of transparency is a big issue, SyRI pseudonymizes the data sources → black box effect, the exact process is not traceable

SyRI has been primarily used in low-income neighbourhoods → the discrimination is “pre-programmed”

2.3 Are you aware of literature laying down additional standards / principles for AI? If so, which studies do you find most important and why (please add source)?

-
- The Ethical Algorithm by Michael Kearns and Aaron Roth: Differential Privacy and Fairness Measures
 - Differential Privacy can solve the upcoming problem of lack of real anonymity
 - Fairness Measures can be an (additional) solution to ease algorithmic discrimination
-

Which principles or standards from that literature are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
Differential Privacy	Fairness Measures	Fairness Measures	Fairness Measures	Fairness Measures
Fairness Measures				

2.4 Do you believe that these standards (identified in questions 2.1, 2.2 and 2.3) adequately address the problems you have identified (question group 1)? If not, what are the main shortcomings / which principles are missing? Please indicate as specifically as possible where in the process these additional principles could be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
The standards set in the GDPR mostly apply only when the discrimination or decision has already taken place. There should be standards and principles that also rule the set up process of the ML model.		Algorithmic Fairness measures are an interesting idea to operationalize fairness in AI; however, it is highly doubtful if fairness can be implemented as a fix formula that only regards a result distribution; the risk of violation of individual rights becomes virulent.		

3. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☒
☐

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: The developers are programmers or IT-experts but not legal experts. They are not familiar with the existing guidelines.

☐
☒

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: The existing documents may be hard to understand for a layperson but that doesn't lead to the conclusion, that they mean little to them. Maybe there is the problem, that they think it means little to them, just because they don't

understand everything. Nevertheless, the principles are very important to consider.

- ☒ ☐ The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain: I agree, because even for a legal expert these principles are sometimes hard to understand. However, in the daily life the developers/users of the AI systems are the ones who have to observe the rules. This is not or just limited possible, if the principles are too general to understand for not legal informed persons.

- ☒ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain: The last question already implies that the principles are phrased very generally. The principles can't consider every little case, so for every case the general principles have to be interpreted.

- ☐ ☒ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain: There are not too many guidelines and standards. They all have their importance and shouldn't be removed. In fact, there are too few guidelines. However, I agree with the fact that the developers and users are not familiar enough with the guidelines to decide which standards they should adhere to.

- ☒ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain: It needs time until there is a clear legal application, which can be used as a base for later cases.

- ☐ ☒ There is a lack of supervision and enforcement of the existing standards and principles

Explanation: The lack of supervision and enforcement concerning the existing standards, especially those set by the GDPR, decreases. This can be seen in the changing behaviour of media providers' cookie usage and information. However, for AI many standards needed do not even exist yet. It is therefore before all a

question

of

legislation.

☐ ☒ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

It should be easy to understand and to use for the programmers and users of AI Systems. It does not have a big effect when only legal experts understand the guide because they are not the main operators or affected ones. The principles must not only include broad principles, but explicit and detailed guidelines.

This is particularly the case if the guide is non-binding and is intended to help and guideline for developers and operators of AI systems.

7.3.23 Respondent 23

Enquête

Naam: Rik Helwegen
Organisatie: N/A

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Artificial Intelligence is het door machines laten uitvoeren van taken die menselijke intelligentie vereisen.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

*Zelfrijdende auto's
Google zoekmachine
AlphaGO*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

*Ongewenste discriminatie
Laagdrempelig opschalen van schadelijke behandeling/uitkomsten
Filterbubbel*

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

*Efficiëntie in de zorg
Standaardisatie van behandeling
Toegang tot meer flexibele modellen, bijvoorbeeld voor klimaat*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

Ja want, ...

Acceleratoren zijn: ...

Nee want, ...

In het bedrijfsleven is dat al gebeurd, en ik verwacht dat er daar groei zal blijven maar minder explosief. Bij andere instellingen, zoals bij de overheid, zal de toename van AI naar mijn verwachting geleidelijk gaan door de afhankelijkheid van bestaande systemen, organisatie structuur en de publieke opinie.

Obstakels zijn:(naast bovengenoemde) Uitvoering; AI toepassingen vereisen een sterke technische infrastructuur en gestructureerde data opslag.

1. Probleemanalyse

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Voor het **ontwikkelen en gebruik maken van** een AI-systeem stel ik de volgende alternatieve stappen voor.

Definiëren van een doel

Verzamelen van data

Ontwikkelen/trainen van een algoritme

Het maken van besluiten/beleid op basis van voorspellingen

Evaluatie en onderhoud van het model in gebruik

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Definiëren van een doel</i>	<i>(2) Verzamelen van data</i>	<i>(3) Ontwikkelen / trainen algoritme</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Evaluatie en onderhoud tijdens gebruik</i>

Problemen: 1) Onvoldoende gedefinieerd wanneer het maken van onderscheid ongewenst is 2) Ethische aspecten worden onvoldoende opgenomen in het doel, in aanvulling op bijvoorbeeld voorspelkracht	Problemen: 1) Bevooroordeling in het verzamelen van de data 2) Onnauwkeurigheid in de dataverzameling 3) Tekort aan data van bepaalde groepen 4) Onbewuste bevooroordeling in meetmethoden of kenmerkdefinities	Problemen: 1) Een model wordt ergens voor geoptimaliseerd, bijvoorbeeld voor voorspelkracht op een testset. Als het tegengaan van bepaalde vormen van discriminatie is opgenomen in het doel, dan moet dit ook worden meegenomen in de technische optimalisatie. Probleem hierbij is de vertaalslag van een 'intuïtieve' eis naar een statistisch doel.	Problemen: 1) Onvoldoende inzicht in werking en impliciete gevolgen van het algoritme	Problemen: ... 1) Feedback loop waarbij de voorspellingen invloed hebben op welke data wordt verzameld in een volgende iteratie 2) Feedback loop waarbij de voorspellingen invloed hebben op de daadwerkelijke situatie, waardoor de data anders is bij een volgende iteratie 3) Verandering van de situatie die niet ligt opgeslagen in de historische data waar het model op getraind is.
Voorbeelden: 1) Stel; bij een algoritme voor fraudeherkenning wil je voorkomen dat migratieachtergrond een rol speelt. Er blijkt een (indirecte) samenhang te bestaan tussen	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: 1) Wanneer een algoritme de focus op een bepaalde wijk legt, dan zal hierdoor meer data over die wijk binnenkomen in de toekomst. Dit

migratieachtergrond en inkomen. Tegelijkertijd is de definitie van fraude expliciet afhankelijk van het inkomen. Is het dan ongewenst om een ongecorrigeerde versie van het inkomen mee te nemen? De meningen hierover zullen waarschijnlijk verschillen, waaruit blijkt dat een precieze bepaling van welk onderscheid ongewenst is moet worden opgenomen in het doel.				kan zorgen voor een vergrote of juist een kleinere focus op die wijk in een volgende iteratie, omdat er nu meer observaties zijn die laten zien wat er in die wijk gebeurd. 2) Wanneer een algoritme de focus op een bepaalde wijk legt kan dit er ook voor zorgen dat de mensen in die wijk zich anders gaan gedragen. Dit heeft niet perse invloed op welke data (/personen) worden geobserveerd, maar op de waarde van de observaties.
Oorzaken: 1) Het hoofddoel van een algoritme is om onderscheid te maken. Het algoritme moet bijvoorbeeld een fraudeur onderscheiden van een niet-fraudeur.	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Voor het maken van een implementatie is het nodig om precies te weten welke vormen van (indirecte) discriminatie binnen de gegeven context ongewenst zijn. Dit komt niet op een natuurlijke manier naar boven uit het publieke debat of uit de bestaande wetgeving. 2) Ethische doelen zijn in veel gevallen moeilijk te formuleren op een manier die ethisch begrijpelijk en aanvaardbaar is, en tegelijkertijd technisch implementeerbaar				

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Definieren van een doel	(2) Verzamelen van data	(3) Ontwikkelen / trainen algoritme	(4) Besluit- en beleidsvorming	(5) Evaluatie en onderhoud tijdens gebruik
Problemen: ...	Problemen: Mogelijk onbewuste sample bias	Problemen: Er is een afweging tussen voorspelkracht en transparantie/ begrijpelijkheid van modellen.	Problemen: ...	Problemen: ...

Voorbeelden: ...	Voorbeelden: In een bepaalde wijk meer controles uitvoeren, en op basis hiervan conclusies trekken voor een hele stad	Voorbeelden: Een neurale netwerk voorspelt voor sommige toepassingen beter dan een logistische regressie, daardoor vermindert de willekeur in het beleid als gevolg van de voorspellingen. Echter is een voorspelling van een neurale netwerk niet in begrijpelijke taal te linken aan de input variabelen, en dat is wel zo bij een regressie.	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef ook zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Definiëren van een doel	(2) Verzamelen van data	(3) Ontwikkelen / trainen algoritme	(4) Besluit- en beleidsvorming	(5) Evaluatie en onderhoud tijdens gebruik
Problemen:	Problemen: ...	Problemen: ...	Problemen: ...	Problemen:
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Definiëren van een doel	(2) Verzamelen van data	(3) Ontwikkelen / trainen algoritme	(4) Besluit- en beleidsvorming	(5) Evaluatie en onderhoud tijdens gebruik
Bestaande problemen: Het niet in staat zijn om goed te definiëren wat gewenste en ongewenste discriminatie is (zie voorbeeld 1.1); is een probleem dat los staat van AI, wat dus ook fout zal gaan in bestaande oplossingen	Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: ... Behandeling kan nu afhankelijk zijn van de dienstdoende persoon, en daarmee enige willekeur bevatten. AI kan dit standaardiseren, waardoor een meer gelijke behandeling mogelijk wordt.	Bestaande problemen: ...
AI gedreven oplossingen: Doordat er voor het trainen van een algoritme zo'n expliciete definitie nodig is van ongewenste discriminatie ontstaat er debat, en misschien consensus, over wat het is.	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Definiëren van een doel</i>	<i>(2) Verzamelen van data</i>	<i>(3) Ontwikkelen / trainen algoritme</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Evaluatie en onderhoud tijdens gebruik</i>

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Definiëren van een doel</i>	<i>(2) Verzamelen van data</i>	<i>(3) Ontwikkelen / trainen algoritme</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Evaluatie en onderhoud tijdens gebruik</i>

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

paper: Counterfactual Fairness, Kusner et al. (2017)

<https://arxiv.org/abs/1703.06856>

Bovenstaande studie werkt toe naar een definitie van fairness die het maatschappelijke doel dient en ook technisch inzetbaar is.

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Definieren van een doel	(2) Verzamelen van data	(3) Ontwikkelen / trainen algoritme	(4) Besluit- en beleidsvorming	(5) Evaluatie en onderhoud tijdens gebruik

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

	(2) Verzamelen van data	(3) Ontwikkelen / trainen algoritme	(4) Besluit- en beleidsvorming	(5) Evaluatie en onderhoud tijdens gebruik

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

- ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

- ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: bovengenoemde principes zijn op tal van manieren te interpreteren, dat geeft te weinig houvast voor het maken een systeem.

- ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis,

interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:

☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:

☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:

☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Er is (naar mijn weten) geen instantie ter bescherming tegen discriminatie door algoritmen. Enigszins paradoxaal lijkt de huidige controle ter bescherming van de privacy de 'eerlijkheid' in algoritmen tegen te werken. Voor het controleren of corrigeren van discriminatie tussen bepaalde groepen is het nodig om te weten wie tot welke groep behoort. Dit betekent al snel dat er met 'bijzondere persoonsgegevens' gewerkt moet worden, en dat levert meer werk en weerstand op door de privacywetgeving. Het is daardoor minder aantrekkelijk geworden voor instanties om te controleren op, en te corrigeren voor, ongewenst discriminatie.

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

7.3.24 Respondent 24

Enquête

Naam:
Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Kunstmatige intelligentie die taken van mensen overneemt.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Zelfrijdende auto, beeldherkenning, workflowoptimalisatie

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Selectie bias in traintdata, werkloosheid, fout-positieven

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Automatisering, kostenbesparing, gebruikersgemak

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, ...

Nee want, teveel praktische problemen

Acceleratoren zijn: ...

Obstakels zijn: kwaliteit trainingsdata

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: selectiebias	Problemen: techniek niet optimaal	Problemen: ...	Problemen: ...	Problemen: ...
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: wet- en regelgeving	Problemen: ...	Problemen: ...	Problemen: ...	Problemen: ...
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...

<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: Aansprakelijkheid</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: Wie is verantwoordelijk voor verkeerd besluit?</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: Selectiebias</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: Aanleggen brede dataset</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

FAIR data

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit:

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit:

☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:
- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:
- ☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:
- ☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Enquête

Naam: Frank van Praat

Organisatie: KPMG

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Technologische systemen die in staat zijn handelingen, taken en acties van mensen die intelligentie vereisen over te nemen.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

-
- *Fraudedetectie in de financiële sector*
 - *Zelf-rijdende auto's*
 - *Tekstherkenning in verschillende vormen/varianten*

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

-
- *Ethische risico's: kans op etnisch profileren (vooringenomenheid)*
 - *Overschatting van de mogelijkheden (bijv. generaliseerbaarheid)*
 - *Fouten in algemene zin door onvoldoende kwaliteit*
-

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

-
- *Taken sneller, beter, efficiënter uitvoeren*
 - *Nieuw vak/werkgebied voor professionals*
 - *Nieuwe oplossingen voor bestaande problemen*

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, er vindt al heel veel onderzoek plaats op dit gebied. Daarnaast zijn er use cases waaruit blijkt dat er sprake is van échte meerwaarde.

Acceleratoren zijn: veel investeringsbereidheid, voorbeelden van waar het werkt.

Nee want, ...

Obstakels zijn:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) <i>Aanleggen dataset</i>	(2) <i>Ontwikkelen/ trainen algoritme</i>	(3) <i>Patronen destilleren uit data</i>	(4) <i>Besluit- en beleidsvorming</i>	(5) <i>Feedbackloop</i>
Problemen: - historische data niet representatief voor werkelijkheid - data kwaliteit is vaak onvoldoende om goede analyses te draaien - legacy IT systemen maakt ontsluiten van data moeilijk door verschillende vormen van opslag	Problemen: - geen eenduidige definitie (manier van meten) van discriminatie. - onvoldoende kennis/vaardigheden van de ontwikkelaar	Problemen: - bestaande patronen worden “gekopieerd” door algoritmes - niet alle data om een algoritme op eerlijkheid te testen is beschikbaar (bijv. door AVG)	Problemen: - vaak wordt “human in the loop” als oplossing gezien, maar als deze “human” beslissing neemt op basis van uitkomsten van algoritme, is kwaliteit van die uitkomsten net zo belangrijk als anders - teveel vertrouwen in de juiste uitkomsten van een algoritme	Problemen: - bestaande fouten in een algoritme worden versterkt - onvoldoende monitoring op de loop - geen eenduidige metrieken die helpen bij het monitoren
Voorbeelden :	Voorbeelden: - sommige metrieken sluiten elkaar uit (één zegt: eerlijk, ander zegt: niet eerlijk, en andersom) - iedereen noemt zich tegenwoordig	Voorbeelden: - om een algoritme op eerlijkheid te kunnen toetsen moeten de sensitieve attributen	Voorbeelden: ...	Voorbeelden: ...

	“data scientist” zonder de échte kennis/ervaring daarvoor te hebben	(bijv. geslacht) worden gebruikt. Maar dit mag niet vanuit wet- en regelgeving.		
Oorzaken: - data kwaliteit is iets wat pas in de laatste jaren belangrijk is geworden, terwijl ook eerdere data noodzakelijk zijn om te gebruiken	Oorzaken: - onvoldoende eenduidigheid over te volgen aanpak: er zijn geen duidelijke afspraken wat onder discriminatie wordt verstaan in het licht van algoritmes - geen “beschermd” beroep (bijv. dmv certificering)	Oorzaken: - bestaande wet- en regelgeving, onvoldoende richtinggevend hoe moet worden omgegaan met discriminatie en gebruik van data om dit tegen te gaan - historische data is vaak het resultaat van menselijk handelen	Oorzaken: ...	Oorzaken: - er wordt bij de ontwikkeling onvoldoende nagedacht over de negatieve gevolgen op de langere termijn - ontwikkelaar is alleen bezig met de ontwikkeling, het “draaien” daarna wordt niet meegewogen in de ontwerpkeuzes

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: ...	Problemen: - Dit kan dermate complex worden dat transparantie geen uitkomst biedt want burgers begrijpen het toch niet - Daarnaast kan	Problemen: - met bepaalde technologieën is het gewoonweg niet mogelijk om te begrijpen welke patronen ontdekt zijn, omdat dit te complex is.	Problemen: - er zijn geen duidelijke “objection” procedures - teveel vertrouwen in “de mens” achter het algoritme die uiteindelijk het besluit neemt	Problemen: ...

	het zijn dat er juist geen transparantie gewenst is, omdat er bijv. fraudeopsporing plaatsvindt			
<i>Voorbeelden: ...</i>	Voorbeelden: - een gemiddelde burger begrijpt niet hoe bepaalde wiskundige en statistische methoden werken. Transparantie daarover zal daarom niet veel bijdragen	Voorbeelden: - meest bekende en gemakkelijke voorbeeld is “deep learning” – dit is technologie die, ondanks onderzoek wat hoopgevend is, vooralsnog niet laat zien welke patronen leiden tot een beslissing	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	Oorzaken: - Complexiteit is inherent aan het ontwikkelproces van geavanceerde data-analyse	Oorzaken: - technologie nog niet zo ver	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Los van het voorgestelde schema zijn er, naast ongetwijfeld nog een hoop andere, nog drie andere knelpunten te herkennen:
 Beschikbaarheid van kwaliteit goede data om tot goede AI systemen te komen, met name voor organisaties met veel “legacy” vormt dit een uitdaging;
 De kennis van het bouwen van goede AI systemen concentreert zich bij grotere spelers die aantrekkelijk zijn voor ontwikkelaars (denk aan de Googles, Microsoft, etc.)
 Het vertrouwen in AI systemen bij burgers is gemixt. Veel burgers hebben niet het idee het te begrijpen waardoor zij een van nature sceptische houding aannemen.

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde

beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) <i>Aanleggen dataset</i>	(2) <i>Ontwikkelen/ trainen algoritme</i>	(3) <i>Patronen destilleren uit data</i>	(4) <i>Besluit- beleidsvorming</i> en	(5) <i>Feedbackloop</i>
Bestaande problemen: ...	Bestaande problemen: - Op dit moment wordt er, als er sprake is van menselijke besluitvorming, vaak niet structureel onderzocht in hoeverre er sprake is van discriminatie. Er is veel vertrouwen in de mens.	Bestaande problemen: - Op dit moment wordt er nog onvoldoende aandacht besteed aan hoe de patronen in de historische data (met discriminatie) kunnen worden doorbroken	Bestaande problemen: - In bestaande besluiten en beleidsvormingsprocessen is er ook al sprake van discriminatie, alleen wordt deze minder herkend.	Bestaande problemen: ...
AI gedreven oplossingen: ...	AI gedreven oplossingen: - Bij het ontwikkelen van een algoritme wordt het cruciaal om te gaan toetsen of het algoritme “eerlijk” is, dit geeft veel meer inzicht in wat er speelt/gaande is.	AI gedreven oplossingen: - Er wordt steeds meer geredeneerd vanuit een causaliteitsprincipe, waarbij “echte wereldverbanden” worden gebruikt om de eerlijke werking van een algoritme te toetsen	AI gedreven oplossingen: - Mits goed uitgevoerd, geeft AI juist de mogelijkheid tot écht eerlijke (en niet vooringenomen) behandeling	AI gedreven oplossingen: ...

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

- Er zijn erg veel standaarden/guidelines etc. in omloop. Individuele organisaties stellen deze op, en ook de overheid zowel nationaal als internationaal komt met steeds meer richtlijnen. Al deze richtlijnen hebben hun waarde, maar het probleem is vaak wel de stap naar het operationaliseren. Dat ontbreekt vaak, en is niet concreet. Daar ligt nog een wereld te winnen. De m Om te spreken van de meest "veelbelovende" vind ik daarom wat te voorbarig, de meest impactvolle zijn wat mij betreft:
- EU guidelines for trustworthy AI (<https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>): erg uitgebreid en allesomvattend, daarom volledig en interessant.
- Overheidsrichtlijnen voor verantwoord algoritme gebruik (<https://www.rijksoverheid.nl/documenten/rapporten/2019/10/08/tk-bijlage-over-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid>): wordt serieus opgepakt en in de praktijk onderzocht om te concretiseren, daarom de moeite waard

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Deze vraag gaat uit van een (erg) uitgebreide analyse van de specifieke richtlijnen (+- paar uur werk). In het kader van de tijd sla ik dat over.

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

- De AVG is op dit moment het meest verregaand als het gaat om geautomatiseerde besluitvorming specifiek voor AI.
- Daarnaast is natuurlijk alle geldende wet- en regelgeving van toepassing die gelden in de specifieke domeinen waar AI voor wordt ingezet.

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
- AVG beperkt het gebruik van sommige databronnen zonder expliciete toestemming		- AVG beperkt het gebruik van beschermde data attributen die juist nodig zijn om de eerlijkheid van	- AVG vraagt om uitlegbaarheid - AVG verplicht om een andere mogelijkheid aan te bieden van	

		een algoritme te testen.	geautomatiseerde besluitvorming	
			- Overige wet- en regelgeving gaat met name hierover. Als een algoritme in een bepaald domein/context wordt gebruikt, is de geldende wet- en regelgeving in dat specifieke domein van toepassing	

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Zie 2.1.

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

Deze vraag gaat uit van een (erg) uitgebreide analyse van de specifieke richtlijnen (+- paar uur werk). In het kader van de tijd sla ik dat over.

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

- In algemeenheid geldt voor vrijwel alle richtlijnen dat deze erg "high level" zijn opgesteld en daarom niet erg richtinggevend om de vraag "hoe dan" te beantwoorden. Er is ondertussen veel literatuur te vinden over hoe de onderstaande specifieke stappen kunnen worden uitgevoerd, echter is dit zo fragmentarisch dat het niet mogelijk is dit nu in één overzicht op te nemen.

(1) Aanleggen dataset	(2) Ontwikkelen/	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

	<i>trainen algoritme</i>			
		- AVG werkt hier juist beperkend.	- AVG stelt hier vrij concrete eisen die goed zijn om de burger te beschermen. Daarnaast zijn de eisen principle-based, waardoor de invulling aan de organisatie in kwestie zelf wordt gelaten.	

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ X De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit: De bekendheid van de richtsnoeren, etc. is wel bekend. Alleen voor een ontwikkelaar zijn deze onvoldoende concreet om daadwerkelijk mee aan de slag te gaan of ze zijn te omvattend om overal rekening mee te houden. Het is uiteindelijk aan de organisaties waarin de ontwikkelaars werken zelf om te zorgen dat de richtlijnen worden vertaald naar concrete werkwijzen die voor de ontwikkelaars gelden. Maar deze stap wordt vaak nog niet gezet.

- ☐ X De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
- Leg uit: Het probleem zit denk ik niet in het begrijpen van de richtlijnen, maar vooral in het operationaliseren daarvan (zie bovenstaand).*

- X ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: Eens – vaak zijn de richtlijnen zo geformuleerd dat je er niet mee oneens kan zijn. Daarnaast kunnen AI-toepassingen (en de technologie daaronder) zo specifiek zijn, dat generieke normen geen oplossing bieden die concreet genoeg is om mee aan de slag te gaan.

- X ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: Zie bovenstaand. AI technologie is zo divers, dat het moeilijk is dit te generaliseren naar principes. Dat kan wel, maar dan kom je op het niveau uit waarvoor nu wordt gekozen, en dan blijft het vaak bij het functionele (het domein waarin de toepassing wordt gebruikt). Het technische wordt vaak overgeslagen omdat dit dus niet te generaliseren valt.
- ☐ X Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Het is niet erg dat er veel richtlijnen, etc. zijn. Dit geeft organisaties de ruimte om keuzes te maken wat ze wel/niet willen volgen. Het probleem zit in de concreetheid en de handrekeningen die uitblijven om dit te operationaliseren.
- ☐ X De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Zie eerdere antwoorden. Een aantal richtlijnen bestaan al best heel lang.
- X ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Het is zo dat de richtlijnen en standaarden nog geen dwingend karakter hebben. Echter zijn ze nu zo “open” dat dit ook vrij lastig zal zijn om te doen.
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

Wat mij betreft is het belangrijk dat de uitwerking voldoende concreet is, en niet alleen op functioneel niveau blijft steken. In de praktijk zou dit kunnen betekenen dat er voor verschillende technieken die gebruikt worden binnen de ontwikkeling van AI systemen (bijv. supervised, unsupervised, constraint satisfaction, expert systemen, etc.) verschillende designprincipes van toepassing zijn omdat het een andere manier van “design” vraagt. Het lijkt me verstandig om dit besef te hebben, en hiermee rekening te houden bij de uitwerking van de handreiking.

Survey

Name:

Organization:

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Data-driven computational techniques deployed to accomplish tasks which require certain degree of human cognitive capacity.

Which three examples of (current or future) AI use come to mind first?

Advertising
Policing
Insurance
Logistics

Which three dangers / negative sides of AI are you most concerned about?

-
- *Bringing the end of “political” as humans lose any chance of conscious intervention*
 - *Amplifying inequalities*
 - *Promoting a false sense of “solutionism”*
-

Which three opportunities / positive sides of AI come to mind first?

Personalised medicine
Efficient scrutiny of public spending and resource allocation
Elderly care

Do you think that AI applications and their use (for example by governmental organisations, companies and non-profit organizations) will increase in the near future? Why / why not - what are accelerators and what are obstacles?

1. Problem analysis

Yes because, it is cheap, efficient and with little accountability

No because, ...

Accelerators are: vast amount of data, processing capacity and the political in capacitation of the data subjects

Obstacles are :lack of clarity as to the availability accessibility and sharing of data. Also lack of regulatory framework addressing issues when thing things run amok.

Different phases / steps play a role in the development and use of AI systems, such as: (1) the creation of a data set, (2) the development and training of an algorithm, (3) the production of patterns, models and profiles through data analysis, (4) making decisions / policies based on these models and (5) a feedback loop in which the results are 'fed' to the self-learning algorithm. [If you envisage a different categorisation/other phases, please indicate so]

When using AI, (unintended) discrimination can occur. What do you think are the underlying causes and can you give concrete examples of these? Where (which sectors, applications, etc.) do you foresee that the problem of discrimination in AI systems will occur in particular? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) Creating a dataset	(2) Developing / training algorithm and	(3) Distilling patterns from data	(4) Decision and policy making	(5) Feedback loop
Problems: lack of access	Problems: Under or over representation of certain groups, places etc.	Problems: Obscured causality	Problems: who decides what is to be decided? Public functions over taken by private interests	Problems: ...
Examples: Who owns the data in IoT devices manufacturer, service provdier, the user?	Examples: People over certain age create less mobile phone data.	Examples: One's not purchasing a deep-frier does not prove that he/she is not	Examples: Why is one's social media nick name choice relevant for work performance	Examples: ...

		eating deep-fried.		
<i>Causes: no proper data access regime but mostly</i>	<i>Causes: basically social inequalities are reflected in data</i>	<i>Causes: no totally objective, abstract method for separating correlation from causality. Lack of Contextual awareness</i>	<i>Causes: Difficulties in conceptualising AI-based decisions in a normative outlook.</i>	<i>Causes: ...</i>

Transparency of AI-based policy and decision making is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there with regard to transparency of AI processes and AI based policy- and decision-making? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: One's lack of access to data and difficulties in discovering the extent of data use.</i>	<i>Problems: Which examples are chosen to comprise the training dataset</i>	<i>Problems: Obscured causality</i>	<i>Problems: ... Obscured normativity</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples:</i>	<i>Examples: see above</i>	<i>Examples: see above</i>	<i>Examples: ...</i>
<i>Causes: commercial interests and issues with the</i>	<i>Causes: This may disclose the commercially valuable ML model,</i>	<i>Causes: This may disclose the commercially valuable ML model,</i>	<i>Causes: see above</i>	<i>Causes: ...</i>

<i>integrity of the system</i>	<i>assumptions of underlying the system etc.</i>	<i>assumptions underlying the system etc.</i>		

What do you think are other bottlenecks with regard to AI, for example technical, social or societal obstacles/problems? Can you give concrete examples of this? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

Could AI also help reduce / remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how? Please indicate as specifically as possible where in the process these benefits can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>
<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>	<i>Existing problems: ...</i>

<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>	<i>AI-driven solutions: ...</i>

2. Existing Standards

2.1 Do you know of existing policy documents, standards, codes of conduct or guidelines with regard to AI? If so, which are most promising and why (please add source)?

Which principles or standards from those documents are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>

2.2 Are you aware of existing legislation or case law regarding AI? If so, which (please add source)?

Which principles or standards from that legislation or case law are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

UK Uber drivers are taking the algorithm to court
<https://techcrunch.com>

(1) Creating a dataset	(2) Developing / training algorithm	(3) Distilling patterns from data	(4) Decision and policy making	(5) Feedback loop

2.3 Are you aware of literature laying down additional standards / principles for AI? If so, which studies do you find most important and why (please add source)?

Stuart Russell, *Human Compatible_ Artificial Intelligence and the Problem of Control*, 2019
Thomas Wischmeyer, Timo Rademacher, *Regulating Artificial Intelligence*, 2020
Philip A.E. Brey et al., *Current Issues in Computing and Philosophy*, 2008
Franco Berardi, *And the phenomenology of the end*, 2014

Which principles or standards from that literature are relevant in the context of discrimination, bias, fairness and transparency? Please indicate as specifically as possible where in the process these can be situated (of course, you do not have to write down something in every box).

(1) Creating a dataset	(2) Developing / training algorithm	(3) Distilling patterns from data	(4) Decision and policy making	(5) Feedback loop

2.4 Do you believe that these standards (identified in questions 2.1, 2.2 and 2.3) adequately address the problems you have identified (question group 1)? If not, what are the main shortcomings / which principles are missing? Please indicate as specifically as possible where

in the process these additional principles could be situated (of course, you do not have to write down something in every box).

<i>(1) Creating a dataset</i>	<i>(2) Developing / training algorithm</i>	<i>(3) Distilling patterns from data</i>	<i>(4) Decision and policy making</i>	<i>(5) Feedback loop</i>

3. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☐ ☒ The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: way too many diverse, disorganised

☒ ☐ The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: they are mostly abstract principles

☒ ☐ The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain: perhaps legal rules, principles are not adequate to regulate AI

- ☒ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain:

- ☒ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain:

.....

- ☒ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain:

.....

- ☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles

Explain:

.....

☐ ☐

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

Enquête

Naam: Miek Wijnberg

Organisatie: Burgerrechtenvereniging Vrijbit

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

Kunstmatige computer slimheid waarmee softwareprogramma's, middels zelflerende rekenkundige modellen, elektronische gegevens kunnen combineren, analyseren en daar gevolgtrekkingen aan kunnen verbinden.

NB We duiden het begrip principieel nooit alleen in het engels aan: dus als Kunstmatige intelligentie(AI)aangezien het uitsluitend gebruik van engels bevordert dat mensen niet eens durven te vragen wat het is... net als met LoT.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Algoritmische risico profilering: A- t.b.v. fraude kansberekening door overheid via SYRi (en nieuwe wetsvoorstel SYRI-PLUS) en via de Belastingdienst, B-financiële afwegingen dienstverleningsvoorwaarden door particuliere banken en verzekeringsmaatschappijen & C-programma's als in UMC waarbij van patiënten agressief gedrag voorspeld wordt m.b.v. inzet KI.

Algoritmisch gedetermineerde persoonsprofielen samenstelling waarmee internetbedrijven persoonsgerichte reclame, gefilterde toegang tot informatie en politieke beïnvloeding realiseren Inzet van drones als autonome vernietigingswapens en zelfbeslissende grensbewakingsapparatuur

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Doorontwikkeling van grootschalige registratie- en controledrift vanuit de illusie dat overall technologische oplossingen voor mogelijk zijn. In combinatie met de toename van geïnstitutionaliseerde onverantwoordelijkheid. Teloorgang van persoonlijke verantwoordelijkheid in ruil voor het excuus dat de inzet van systemen/ computers minder foute beslissingen zouden maken dan mensen.) Gedragen door een toenemende ontmenselijking waarbij degenen die verantwoordelijk zijn voor beleidsbeslissingen/ besluiten over hoe mensen bejegend/beoordeeld worden geen direct contact meer hebben met betrokkenen waardoor emotionele besluiten over anderen vergemakkelijkt worden.

Afname van sociaal gedrag en solidariteit in de maatschappij door een voortgaande ontwikkeling dat individuen niet meer als autonome personen worden beschouwd maar als equivalent van het datapakket van hun persoonsgegevens,

Het gevaar dat wegens onkunde en onwil om naar bezwaren te luisteren de politiek, media en wetenschappers hun oren laten hangen naar de (deels geheime) bedrijfsbelanghebbenden bij invoeren van KI systemen, zonder dat de uitwerking daarvan te overzien is en zonder dat dergelijke ontwikkelingen teruggedraaid kunnen worden. Met als gevolg: nog grotere macht van bedrijfsleven- met name grote defensie& security bedrijven en tech giganten- ten koste van publieke belangen, het functioneren van democratie en het respecteren van fundamentele mensenrechten.

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Technische oplossingen voor technische problemen qua ontwikkeling van hulpmiddelen voor invaliden, of ontwikkeling van medicijnen.

Ruimteonderzoek

Berekeningen kunnen maken voor mondiale problemen als klimaatvervuiling en t/m voorspellen van natuurgeweld als aardbevingen, tornado's ed.

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, er zijn gigantische belangen mee gemoeid en de mogelijkheden zijn schier onbegrensd. En omdat de overheid kiest voor economische belangen en hiermee Nederland rijk en machtig verwacht te kunnen houden. Bedrijfsleven denkt er een geweldig verdienmodel mee in handen te hebben.

Nee want, ...

Obstakels zijn:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: foute data, zoals oude/vervulde gegevens, vermenging van harde en beoordelingsdata, Data los van context, fouten bij invoer waar bij één cijfer verkeerd opslaan al verregaande consequenties, gemanipuleerde data door vooringenomen opzet of hacking, ...	Problemen: algoritme is veelal geheim en niemand begrijpt de rekenkundige modellen zelfs meer zodra deze zich zelflerend doorontwikkelen. Begin data gaan gekleurd als ze zijn als selffulfilling professie de ontwikkeling van de systemen sturen. De techniek zelf wordt bepalend voor wat er wordt onderzocht, geanalyseerde en geconcludeerd bij onderzoeken of beheer datasystemen. de resultaten en Waardoor de techniek en de ICTers die dit ontwerpen en onderhouden sturend worden voor sociale processen en gaan bepalen hoe de samenleving wordt vormgegeven.	Problemen: Foute data en foute probleemstelling leveren onzin t/m gevaarlijke gevolgtrekkingen op. Waarbij in computersystemen onvermijdelijk patronen worden gedestilleerd die niet alleen op foutieve invoer data kan zijn gebaseerd maar ook op fouten die bij gebruik van de data ontstaan door overschrijving, computerstoring en tot koppeling aan andere systemen. Waarbij inherent is aan computersystemen dat fouten zich snel vermenigvuldigen, de oorsprong vaak niet te achterhalen is, niemand verantwoordelijk blijkt en herstel vaak onmogelijk. Daarbij is het een groot gevaar dat data altijd als harde gegevens worden beschouwd terwijl ze dat niet(hoeven te)zijn. Datasets die(deels) bestaan de uit observaties beoordelen met rekenkundige modellen is recept voor foute beoordeling. Wat	Problemen: Politieke besluiten die direct invloed hebben op de levens van het grootste gedeelte van de bevolking muntten vaak niet uit in een op kennis gebaseerde, effectieve en proportionele oplossing voor daadwerkelijke problemen. De maatschappij is te gecompliceerd, partijpolitiek domineert boven goed beleid, historisch besef mankeert, en meer.. In die situatie is het heel verleidelijk om gebruik te maken van de technologische mogelijkheden om zoveel mogelijk data te verzamelen en het gebruik van Big data te presenteren als moderne oplossing om beleidsbesluiten op te funderen. Besluiten op grond van data suggereren daarbij een neutrale en noodzakelijke gang van zaken, die iedere discussie overbodig lijkt te maken. De ervaring leert dat het overhaast in te zetten op technologische oplossingen waarvan men de gevolgen niet kan dan wel wil overzien doorgaans geen probleem	Problemen: ... Evaluatie is doorgaans geen recept voor het onderzoeken van de gebezigde methode op bewezen nut en analyse of bij aanvang gestelde doelstellingen daadwerkelijk danwel nog relevantie toekomt. Doorgaans betekent evaluatie bij de inzet van computersystemen slecht het marginaal toetsen of de inzet nog formeel voldoet aan bijvoorbeeld veiligheidseisen en het beoordelen of het systeem aangepast dient te worden. Afschaffen van systemen die niet blijken te werken, mensenrechten blijken te schenden, verboden worden door de rechter, of achteraf op grond van onjuiste feitenweergave blijken te zijn ingevoerd, is doorgaans niet aan de orde. Waar systemen falen wordt liever doorgemodderd aan systemen waar al in geïnvesteerd is, ook als alle deskundigen waarschuwen dat dat qua opzet en doelstelling niet goed kan gaan functioneren. Gezichtsverlies speelt hierbij een rol, niet willen dat reeds gespendeerd geld als

		<i>eveneens geldt voor het feit dat data niet los van de context kunnen worden beoordeeld als men er recht aan wil doen, en die context – hetzij in actuele situatie danwel in historisch perspectief, laat zich niet vangen in puur technische systemen.</i>	vormt voor het invoeren van nieuwe wet-en regelgeving zolang gehamerd wordt op economisch belang, en noodzaak tot beschermen van staatsveiligheid en openbare orde. Waarbij het de overheid prima uitkomt dat de uitvoering van beleid op grond van foute data analyse steeds vaker wordt uitgevoerd door de (semi) private sector(die zich in rechte kan beroepen op geheimhouding wegens bedrijfsbelang) zodat de wetgevende macht zelf buiten schot blijft als het beleid faalt en de uitvoerende partij de verantwoordelijkheid af kan schuiven op de wetgevende macht. Waarbij eveneens van belang is dat sinds het dualisme is ingevoerd als grondslag voor de politiek gemeenteraden en de Staten Generaal een marginale rol zijn gaan spelen in de controle van autocratisch regentenbestuur door B&W's en kabinet en steeds makkelijker zelfs de onderliggende data onthouden worden.	vermoet worden gekwalificeerd. Maar ook de rol van de technische bedrijven die gebaat zijn bij voortzetting van gebruik en onderhoud van technologie speelt een rol. Evenals het belang van onderzoeks-afdelingen en instituten en toezichhouders op publieke belangen, die er niet zitten te wachten dat zij achteraf het verwijt krijgen dat evaluatie hun foute beoordeling/advies of gebrek aan uitwijst dat zij gefaald hebben. Termijn stelling voor looptijd van een bepaalde maatregel/wetgevin gl met aansluitend verplichte inhoudelijke evaluatie ontbreekt standaard.
Voorbeelden: Medische patientgegevens zijn doorgaans onvolledig, slecht genoteerd, niet in	Voorbeelden: ... Wetenschappers en politici hebben er sinds men KI/AI heeft ontdekt als nabij	Voorbeelden: UWV Belastingdienst sociale recherche en zorgverzekeraars	Voorbeelden: Tasers voor politie worden ingevoerd op grond van valse data.	Voorbeelden: ... Evalutie zoals van de Wet op de uitgebreide ID-plicht- waarmee

context ingebed, verouderd, gewoon fout, ...	toekomstscenario de mond vol over ETHIEK die de oplossing zal moeten bieden maar verder dan dit wensdenken als tegenwicht voor doordenderende technologische ontwikkeling zie ik doorgaans geen concrete andere oplossingsrichting en dan de roep om algoritmen hun openbaar te maken.	destileren uit data van onschuldige personen van profielen van mensen die zij verdenken dan wel beschuldigen van crimineel gedrag...	Wetsvoorstel lichaamseigen materiaal wordt gepresenteerd als noodzakelijk om over goede dataregistratie te beschikken, zonder de daarmee gerealiseerde registratie van ieders DNA te benoemen. Jeugdzorg, zorg voor ouderen, onderwijs worden opgezaaid met een registratie- en controle drift die het goed kunnen functioneren van de sector kapotmaakt maar wel veel data oplevert om beleidsbeslissingen op te nemen. Zo bleek de paspoortwet die vingerafdrukken verplichtte uit te gaan van een probleem met ID-fraude die in werkelijkheid niet voorkwam, maar wel als zodanig werd gepresenteerd. Ites wat we ook zagen bij de herhaaldelijke wetsvoorstellen 'bewaarplicht telecomgegevens waarbij via een Ubocht de wetsvoorstellen als noodzakelijk werden voorgesteld vanwege data inzake criminaliteit uit de EU, die bij nader inzien een eigen leven waren gaan leiden terwijl ze door Nederland waren aangeleverd zonder dat het om daadwerkelijke	'geen onomkeerbare stappen zouden worden gezet' bleek eerst na de gestelde termijn van 3 jaar steeds uitgesteld te worden. Vervolgens door een niet onafhankelijk instituut te worden gedaan en niet te gaan over nut en noodzaak, maar over de mogelijke aanpassingen om de ingeslagen weg te consolideren. Als de overheid kritiek blijft krijgen op bepaalde wet-en regelgeving(inclusief inzet van bepaalde systemen) is de reactie standaard niet 'hoe tegemoetkomen aan de bezwaren', maar 'hoe leggen we het beter uit'. Als de rechter een wet/maatregel/systeem verbiedt omdat een wettelijke grondslag ontbreekt, wordt dat bijvoorbeeld omdat het te veel mensen betreft of vanwege bedrijfsbelangen die meespelen niet stopgezet, maar wordt ingezet om de wet zodanig te wijzigen dat er alsnog een wettelijke grondslag wordt gecreëerd. Of er wordt minimale aanpassing gedaan. Of de instantie die een systeem beheert wordt opgeheven en het systeem krijgt via een nieuwe beheerder een doorstart, waarna de hele juridische strijd opnieuw moet worden gevoerd...

			gevallen bleek te gaan. Poging de 'slimme'energiemet ers te verplichten werd gebaseerd op cijfers van energiebesparing uit een TNO onderzoek wat bij nadere beschouwing wel het cijfer van 1% noemde maar dat zelf als natte vingerwerk typeerde. Invoering van ID-plicht bij verkiezingen werd ook als van groot maatschappelijk belang getypeerd, terwijl de data van daadwerkelijke fraude over een andersoortige fraude bij verkiezingen gingen(namelijk ronselen).	en bijvoorbeeld de nieuwe beheerder de verantwoordelijkheid d voor niet verwijderde(en wel uitgewaaierde datasets) ontkent.
<i>Oorzaken: systemen bepalen dat data niet in context worden opgeslagen, worden gehackd, bestaan uit allerlei systemen die niet op elkaar aansluiten(afhankelijk zijn van bepaalde leveranciers, niet goed beveiligd, niet geactualiseerd enz) niemand is verantwoordelijk voor geheel; betrokkenen kunnen onmogelijk zelf zicht houden op wat op wat er over hen wordt vastgelegd door wie en voor welk doel(laat staan zorgen voor accurate data), mensen houden gegevens voor zichzelf(zoals eigen medicijnen die niet</i>	<i>Oorzaken: De overheid heeft de taak om de belangen van de burgers te behartigen tegen machten en krachten die hen hun autonomie en vrijheidproberen te ontnemen... maar die overheid ontbeert het aan kennis en macht om tegenwicht te bieden aan de bedrijfsbelangen en kiest ook zelf voor voorrang geven aan economische belangen.</i>	<i>Oorzaken: Je destilleert data over frauduleus gedrag, ongezonde levenswijze e.d op grond van zwakke economische positie, achterstandswijk en, etnische achtergrond e.d. De uitkomsten voeren het zelflerende systeem waardoor stigmatisering zichzelf gaat bevestigen en sterker wordt.</i> Overheid beschouwd de burgerprincipeel als tegenstander en potentieel crimineel.	<i>Oorzaken: Een maatschappij als bedrijf runnen is vast recept voor foute besluit- en beleidsvorming. Mensen worden ondergeschikt behandeld aan economische belangen. Kapitalisme, consumptiegedreven maatschappij, overschatting van maakbare samenleving... allemaal factoren die aan de basis liggen van principieel foute inzet voor technologie. Data als grondstof beschouwen van een neutrale technologie staat garant voor besluit- en beleidsvorming die niet geent is op fundamentele</i>	<i>Oorzaken: ...</i> <u>Macht</u> van zorgverzekeraars(zoals m.b.t. de gedragscode inzake de verwerking van persoonsgegevens die niet aan EVRM voldoet) <u>Macht</u> van NZa- die de bevoegdheid van haar publieke taak overschrijdt met het diagnose Informatie systeem DIS. <u>Macht</u> van toezichthouders die geen gevolg geven aan onpartijdige advisering en handhaving (zoals AP inzake Persoonsgegevens en DNB m.b.t. gebruik financiële data waarvoor betrokkenen geen toestemming gaven betaalregeling via PSD2)

<i>op doktersrecept worden gebruikt)</i>			<i>rechten en rechtszekerheid die de peilers van onze rechtstaat behoren te vormen. ...</i>	<u>Mogelijkheid</u> van manipulatie van onderzoek door ministerie(zoals WODC en bij universiteiten wegens hun afhankelijkheid van sponsors en subsidies. <u>Kortzichtigheid</u> mbt politieke besluitvorming Mogelijkheid doorstart systemen op andere noemer: private doorstart EPD-LSP door VZVZ, opheffen van stichting Benchmark ROM(routine outcome Monitoring) met doorstart via nieu instituut Akwa

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>éProblemen: Niemand kan meer nagaan waar en waarvoor alle</i>	<i>Problemen: Welke data gebruikt worden en volgens welke</i>	<i>Problemen: Los van alle problemen die samenhangen met bovenstaande</i>	<i>Problemen: De inzet van KI wordt als onontkoombaar gezien. Als</i>	<i>Problemen: ...</i>

persoonlijke data over hem of haar staan geregistreerd en worden gebruikt...	programma's kunnen bedrijven als bedrijfsgeheim afschermen, waardoor onafhankelijk controle onmogelijk wordt. De overheid maakt zich hier ook schuldig aan op grond van 'in belang van de veiligheid' te omvangrijke id van WOB verzoeken e.d. Datasets EN software om data te bewerken kunnen gemanipuleerd worden(ook zonder dat de beheerder zelf daar weet van heeft). Mensen kunnen steeds moeilijker aan het gewone leven deelnemen als ze niet meedoen aan het delen van hun data vis digitale systemen.	inzake problemen met destilleren om tot correcte data analyse te kunnen komen, lijkt hier ook van belang dat te vermelden dat het bij bijvoorbeeld risicoanalyse vaak gaat om kwetsbare groepen. Mensen die niet over voldoende opleiding/kennis, geld/advocatenhu lp ed beschikken om voor hun belangen op te komen. Maar ook om mensen die niet gestigmatiseerd willen worden als onderzoeksonderworpen omdat ze GGZhulp nodig hebben, in achterstands() gebied wonen, als niet westers beschouwd worden. Zaken die uiteraard aan het licht komen als men wil uitzoeken waarom en hoe KI op hen wordt losgelaten.	profeitelijk voor de kenniseconomie en handig als bezuinigingsmethode. De mantra 'KI inzetten mag alleen als het ethisch verantwoord is' lijkt als een soort toverformule te werken om te suggereren dat het daarmee al ethisch verantwoord is. Wie als burger hier kritiek op durft te uiten wordt weggehoond als dombo, neanderthaler, samenzweringsaanhanger, tegenstander van de vooruitgang, e.d. Dat weerhoudt velen -voor zover ze al in een omgeving verkeren- waar het onderwerp ter sprake komt. Kritiek wordt ook bij voorbaat gesmoord omdat 'iedereen vóór fraudebestrijding' is, en dus dan niet tegen het inzetten van middelen daarvoor mag zijn. Gewoon vragen stellen om beter begrip te krijgen over wat de inzet van KI praktisch inhoudt heeft vaak al hetzelfde gevolg. Waarbij het	

			doorgaans gevoerde discours in engels en met computerjargon ook niet helpt om transparantie te bewerkstelligen over wat er zich aan het ontwikkelen is.	
Voorbeelden: Via BSN beschikt de overheid over gigantisch veel data van burger, bedrijven leggen iedere digitale voetstap vast, van vervoerspatrone n tot financiële transacties, internet-en bel gebruik... 'Slimme 'energiemeters weigeren mag, maar is praktisch haast niet te doen waardoor energiemaatsch appijen en de overheid grondstof bemachtigen van het gedrag achter de voordeur geschikt om AI op los te laten voor ontwikkeling van het internet der dingen en beleid t.a.v. smart grit.	Voorbeelden: Waar dienstverlenin g afhankelijk wordt gemaakt van het aanleveren van data wordt de betrokkene al nauwelijks geïnformeerd door wie de data worden gebruikt... maar zeker niet HOE. Als je lichaamsmate riaal mag gebruikt worden voor wetenschappe lijke doeleinden zie dan maar eens uit te vogelen volgens welke programma's en voor welke doelen dat gebeurt/ kan gebeuren. Uitsluiten dat je dna wat	Voorbeelden: Als er iets is wat de kinderopvangtoes lag-affaire heeft aangetoond dan is het wel de onmacht van mensen om in het gelijk gesteld te worden (zelfs als je door de hoogste bestuursrechter dat gelijk krijgt). Met als huidige stand van zaken dat waar het financieel compenseren van al die duizenden gedupeerden al jaren gaat duren het evident is dat hun data tot in lengte van dagen door de systemen zullen blijven circuleren.... Schrikbarend gedetailleerd bleek het leven in kaart gebracht van de journalist in Twente die na ging wat de verwerking van al zijn bewegingen buitenshuis opleverde.	Voorbeelden: Vanuit Burgerrechte- nvereniging Vrijbit hebben we ruime ervaring in het worden weggehoond uit de tijd dat we opkwamen voor het recht op privacy enhet teolichten dat dit niet om een kwestie als 'niks te verbergen' ging. Diezelfde trant herhaalt zich nu we opkomen voor het belang van het behoud van de mogelijkheden voor on-digitale communicatie en kanttekeningen bij de wijze waarop met een gretigheid door de overheid en het bedrijfsleven wordt ingezet op het investeren in KI	Voorbeelde n: ...

	<i>bekend is vanwege medische behandeling of donorschap nooit gebruikt mag worden voor KI inzake ontwikkeling genetische manipulatie toepassingen is onmogelijk.</i>	<i>Onontkoombaar voor wie langs een camera loopt, in de auto stapt, OV gebruikt, elektronische betaald, met mobiele telefoon op zak loop is de onnavolgbare wijze waarop al die data nu al verknoopt kunnen worden laat staan wat daar via KI op grotere schaal of op langere termijn mee kan worden gedaan</i>		
<i>Oorzaken: Digitaliserings drang-en dwang in combi met doorgeschoten controledrift en nauwelijks besef en interesse van de meeste mensen waar dit toe kan leiden</i>	<i>Oorzaken: Mensen weten niet wie verantwoordelijk is voor het doorverwerken van hun data, doorgronden niet wat kunstmatige bewerking van hun data behels, interesseret het niet totdat men er negatieve gevolgen daadwerkelijk van ondervindt en het te laat is, hebben niet de macht om het juridisch op te nemen tegen machtige bedrijven en</i>	<i>Oorzaken: KI wordt gewoon gebruikt omdat het kan zonder dat de wet bescherming persoonsgegevens daar daadwerkelijk bescherming tegen kan bieden vanuit het recht van het individu om zijn gegevens op te vragen, ze enkel te ogen gebruiken na uitdrukkelijke toestemming vooraf e.d.</i>	<i>Oorzaken: Cognitieve dissonantie ten aanzien van het uitleveren van de samenleving aan technologische toepassingen waarvan de gevolgen niet overzien worden, de verantwoordelijkheden niet vastgelegd worden, wetgeving niet op is toegesneden, betrokkenen geen notie van hebben en alleen op de onomkeerbaarheid wek gerekend kan worden.</i>	<i>Oorzaken: ...</i>

	<i>overheidsinst anties.</i> <i>De NL Staat en EU hebben onvoldoende tegenmacht versus techgiganten die niet onder EU recht vallen danwel voor wie de VS Freedom Act prevaleert En hebben zich afhankelijk gemaakt van naties als China en India waar men totaal geen grip heeft op wat er gebeurt met data die daar terechtkomt via respectievelijk chips in computersystemen of medische data voor ontwikkeling van medicijnen.</i>			

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>

Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: ...	Bestaande problemen: ...
AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...	AI gedreven oplossingen: ...

Bekendheid met literatuur t.a.v problematiek aangaande ontwikkeling van standaarden, juridische kaders:

ECP- Artificial Intelligence Impact Assessment <https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf>

Gemeenschappelijke principes zoals de internationaal geverifieerde en geaccepteerde FAIR-principes (Findable, Accessible, Interoperable, Reusable als basisprincipes

De Personal Health Train (PHT) is een metafoor voor het afsprakenstelsel, de architectuur en de implementatie voor het verantwoord gebruik van gezondheidsdata in o.a. AI-toepassingen

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke

behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ JA: De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit: Ook op dit gebied blijkt dat mensenrechten doorgaans niet op de dagelijkse praktijk bij omgang met personen/persoonlijke data worden betrokken, maar beschouwd worden als iets wat vooral betrekking heeft op onderdrukking in verre landen. Daarnaast wordt het niet beschouwd als desam voor een vrije samenleving maar als rechten die uitsluitend betrekking hebben op de sector van de rechterlijke macht.

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

- ☐ ☐ JA: De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals 'wees transparant', 'vermijd het maken van ongerechtvaardigd onderscheid tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit:gevolg van te algemene randvoorwaarden is dat wie zich daaraan te houden weet denkt te voldoen, danwel weg tekomen, met het semantisch vermelden in aanbestedingen- projectopzetten e.d. dat men 'transparantie nastreeft', 'privacy respecteert', e.d.

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is

Leg uit:

- ☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)

Leg uit:

- ☐ ☐ Nee: Dat de standaarden pas recentelijk ontwikkeld zijn/worden en de tijd vanzelf liefst zelfregulerend tot adequate oplossingen zal leiden is een gevaarlijke doodoener om vooraf nadenken en grenzen stellen tegen te gaan.
- ☐ ☐ JA; Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Het toezicht op de handhaving van het respecteren van principiële mensenrechten
Laat sowieso veel te wensen over en zal m.b.t. de inzet van technologie waar de toezichthouders geen verstand van hebben er niet beter op worden.
Hetgeen eveneens te maken heeft met de praktijk dat toezichthouders maar zeer beperkt toezicht kunnen houden op de inzet van KI zodra de programma's als bedrijfsgeheim worden afgeschermd.
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Voor het ontwikkelen van een handreiking raad ik aan daarbij gebruik te maken van de opzet van JJ Hoepman voor Privacy-by-design. Namelijk dat men **vanaf de eerste stappen en in alle stappen van de ontwikkeling van digitale systemen** de principes van non discriminatie expliciet volgt. Waarbij het van belang is om mogelijke discriminatie van bepaalde mensen en groepen in de breedste zin van het begrip te beschouwen en niet voornamelijk te focussen op de actueel meest aansprekende vormen als discriminatie op etniciteit, geslacht en leeftijd.

Verder zou bij de opzet uitdrukkelijk dienen worden uitgegaan van het besef dat men in de digitale wereld niet anders met elkaar dient om te gaan dan volgens de fatsoensnormen in de 'gewone' fysieke wereld.

7.3.28 Respondent 28

Enquête

Naam: Dr. Jorgen Goossens
Organisatie: Tilburg University

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

AI is het probleemoplossend of lerend vermogen van een systeem om met behulp van gegevens specifieke taken te verwezenlijken.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Chatbots, fraudebestrijding en gezichtsherkenning

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

Gebrek aan transparantie, gebrek aan uitlegbaarheid en rigiditeit bij toepassing van regels

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

Verwerking van big data, automatisering, oplossingen op maat van de eindgebruiker

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja, want de rekenkracht van computers neemt exponentieel toe, wat enorme (toekomst)mogelijkheden creëert. Acceleratoren zijn: rekenkracht computers, investeringen in R&D, toegenomen bekendheid van AI

Nee want, ...

Obstakels zijn: inzicht en uitlegbaarheid

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destillieren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen:	Problemen:	Problemen: kans op bias, kwaliteit van data en historisering	Problemen: rigiditeit van automatische besluitvorming	Problemen: inbouwen van mogelijkheid tot menselijke tussenkoms soms vereist, historisering
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: opsporen fraude	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destillieren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

	<i>trainen algoritme</i>			
<i>Problemen: kwaliteit van data, privacybescherming</i>	<i>Problemen: transparantie by design</i>	<i>Problemen: ...</i>	<i>Problemen: gebrek aan toegankelijke uitleg; omzetten code naar begrijpelijke taal</i>	<i>Problemen: inbouwen van mogelijkheid tot menselijke tussenkomst soms vereist, historisering</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: SyRI- uitspraak: inzet van SyRI onvoldoende inzichtelijk en controleerbaar</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken:</i>	<i>Oorzaken: ...</i>

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destillieren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
<i>Problemen: ...</i>	<i>Problemen: bevoegdheid en verantwoordelijkheid voor ontwikkeling, omzetten gewone taal naar code</i>	<i>Problemen: ...</i>	<i>Problemen: de- individualisering</i>	<i>Problemen:</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden:</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: uitbesteding van ontwikkeling</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

	<i>algoritme: (sub)delegatie; Natural Language Processing is nog in volle ontwikkeling</i>			

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: menselijke fouten en beperkte capaciteit</i>	<i>Bestaande problemen: beperkt vermogen menselijk brein en traagheid van bestaande processen</i>	<i>Bestaande problemen: beperkt vermogen menselijk brein en traagheid van bestaande processen</i>	<i>Bestaande problemen: menselijke fouten en bias</i>	<i>Bestaande problemen: feedback gebeurt vaak slechts bij menselijke detectie van een probleem of bij het formuleren van een klacht</i>
<i>AI gedreven oplossingen: grote computer-rekenkracht, koppeling databases</i>	<i>AI gedreven oplossingen: grote computer-rekenkracht, automatisering van doorontwikkeling van regels</i>	<i>AI gedreven oplossingen: grote computer-rekenkracht en automatisering</i>	<i>AI gedreven oplossingen: automatisering neemt mogelijkheid van menselijke fraude en bias die kan ontstaan bij personen betrokken bij het nemen van concreet besluit weg</i>	<i>AI gedreven oplossingen: feedback kan op systematische wijze worden ingebouwd en door de analyse van vele beslissingen kan beter zicht ontstaan op het collectieve, geaggregeerde effect van beslissingen</i>

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Er bestaat een plethora aan beleidsdocumenten, standaarden, codes of conduct of richtsnoeren

- Bv. vanuit het bedrijfsleven:

IBM, 2018: [Ethics for Artificial Intelligence: a practical guide for designers & developers](#)

- Bv. vanuit een participatief samenwerkingsproces:

University of Montreal - Forum on the socially responsible Development of AI 2017: [The Montreal Declaration for a Responsible Development of Artificial Intelligence: a participatory process](#)

Recent en belangrijk zijn volgende vier documenten als deel van de Digital package van de Europese Commissie:

1) White Paper on Artificial Intelligence: a European approach to excellence and trust:

https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

2) Commission Report on safety and liability implications of AI, the Internet of Things and Robotics:

https://ec.europa.eu/info/files/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics_en

3) Communication: A European strategy for data:

https://ec.europa.eu/info/files/communication-european-strategy-data_en

4) Communication: Shaping Europe's digital future:

https://ec.europa.eu/info/files/communication-shaping-europes-digital-future_en

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

	<i>trainen algoritme</i>			

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Bv. SyRI-uitspraak, ECLI:NL:RBDHA:2020:865

Bestaande algemene en sectorspecifieke wetgeving blijft van toepassing ook als AI wordt ingezet.

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>
Privacy by design		Privacy by design	Compliance by design	

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Wanneer bestuursorganen AI gebruiken, dienen zij de algemene beginselen van behoorlijk bestuur na te leven, zoals bijvoorbeeld het zorgvuldigheids- en motiveringsbeginsel.

Zie bijvoorbeeld

de Poorter, J., & Goossens, J. (2019). Effectieve rechtsbescherming bij algoritmische besluitvorming in het bestuursrecht. Nederlands Juristenblad , 2019(44), 3303-3312.

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
	Non-discriminatiebeginsel	Non-discriminatiebeginsel	Non-discriminatiebeginsel	
Algemene beginselen van behoorlijk bestuur (bv. zorgvuldigheidsbeginsel)	Algemene beginselen van behoorlijk bestuur (bv. zorgvuldigheidsbeginsel)	Algemene beginselen van behoorlijk bestuur (bv. zorgvuldigheidsbeginsel)	Algemene beginselen van behoorlijk bestuur (bv. motiveringsbeginsel)	Recht op menselijke tussenkomst onder de AVG of in de bezwaarschriftprocedure bij bestuurlijke besluitvorming (recht op heroverweging)

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Betere regeling eigenaarschap van data, rechten en plichten	Invoeren specifieke vergewisplicht	Invoeren onderzoeksplicht inzake collectief effect van algoritmische besluitvorming	Nood aan betere operationalisering van een recht op nuttige uitleg en informatie	Mogelijke uitbreiding recht op menselijke tussenkomst

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☒ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: er zijn nog steeds aanzienlijke barrières tussen disciplines/vakgebieden

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: er is nog steeds een grote nood aan een interdisciplinaire dialoog, waarbij voldoende aandacht is voor inzichten uit verschillende vakgebieden die by design worden geïmplementeerd

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit: de kloof tussen theorie en praktijk is heel groot

- ☒ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: er is inderdaad nood aan meer sectorspecifieke analyses
- ☐ ☒ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: er is veel overlap en er zijn dus vrij gemakkelijk 'core standards' te onderscheiden
- ☒ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: de standaarden zijn ondertussen vrij duidelijk; het is nu inderdaad dringend tijd om er in de praktijk voldoende werk van te maken en best practices te ontwikkelen
- ☒ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: huidig toezicht en handhaving is versnipperd, gebrek aan inzicht en expertise
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Graag voldoende aandacht voor het eindgebruikersperspectief, het perspectief van de burger.

Justice must not only be done, but must also be seen to be done. Het system moet niet enkel feilloos werken, maar we moeten ook voldoende het gevoel krijgen dat het feilloos werkt om zo noodzakelijk vertrouwen in de techniek en dus legitimiteit voor de uitkomsten ervan te borgen.

7.3.29 Respondent 29

Enquête

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren?

AI is de wetenschap om machines dingen te laten doen die normalter menselijke intelligentie vereisen om uitgevoerd te worden.

Welke drie voorbeelden van (huidig of toekomstig) AI-gebruik schieten als eerst te binnen?

Legal: jurisprudentie analyse, contract analyse, Due Dilligence, voorspelling rechtspraak
HR: kandidaat selectie, CV scanning, wellbeing voorspelling
Finance: Know your customer processen, anti witwassen/fraude processen

Welke drie gevaren/negatieve kanten van AI staan bij u het scherpst op het netvlies?

AI kan (onopzettelijk) leiden tot discriminatie of ongelijkheid.
Beslissingen van AI zijn niet altijd transparant of uit te leggen.
Het is vaak onduidelijk wie er verantwoordelijk en aansprakelijk is voor de beslissingen van de AI.

Welke drie kansen/positieve kanten van AI schieten u als eerst te binnen?

AI kan "low value, high volume" taken automatiseren waardoor werknemers zich kunnen richten op de belangrijkste taken en meer waardevol worden.
AI werk sneller, efficiënter en met een kleinere foutmarge dan mensen en is bovendien 24/7 beschikbaar waardoor hogere kwaliteit geleverd kan worden tegen lagere kosten.
AI kan taken uitvoeren waartoe mensen mentaal of fysiek niet in staat zijn waardoor menselijke beperking overwonnen kunnen worden.

Denkt u dat AI-toepassingen en het gebruik daarvan (bijvoorbeeld door overheidsdiensten, bedrijven en non-profit instelling) in de nabije toekomst een hoge vlucht zal nemen? Waarom wel/waarom niet – wat zijn acceleratoren en wat zijn obstakels?

1. Probleemanalyse

Ja want, ook bij de overheid, kan er een enorme efficiëntie slag behaald worden m.b.v. AI.

Acceleratoren zijn partnership tussen de public en private sector

Nee want, ...

Obstakels zijn:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) het aanleggen van een dataset, (2) het ontwikkelen en trainen van een algoritme, (3) het vervaardigen van patronen, modellen en profielen door data-analyse, (4) het maken van besluiten/beleid op basis van deze modellen en (5) een feedbackloop waarbij de uitkomsten en resultaten terug worden gegeven aan het zelflerende algoritme. *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden. Wat zijn volgens u de onderliggende oorzaken en kunt u daar concrete voorbeelden van geven? Waar (welke sectoren, toepassingen, etc.) ziet u discriminatie in AI systemen zich met name voordoen? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: de data is geen juiste afspiegeling van de groep of de dataset bevat een bias	Problemen: ...	Problemen: de AI legt ongewenste verbanden tussen eigenschappen en kwaliteiten van subjecten. (correlatie zonder causaliteit)	Problemen:	Problemen: ...
Voorbeelden: een sollicitatie tool selecteert bijna geen vrouwen als geschikte kandidaat.	Voorbeelden: ...	Voorbeelden: Een AI besluit om subjecten die zwanger zijn een hogere hypotheek te geven	Voorbeelden:	Voorbeelden: ...

Oorzaken: Vrouwen waren slecht vertegenwoordigd in de trainingsdata waardoor de AI het verband legde dat vrouwen niet geschikt waren.	Oorzaken: ...	Oorzaken: Op basis van statistiek correleert zwangerschap wellicht met kredietwaardigheid maar er lijkt geen causaliteit te bestaan	Oorzaken:	Oorzaken: ...

Transparantie van AI-besluiten is een best practice; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
Problemen: individueel zijn zich niet bewust dat van het feit dat hun data gebruikt wordt in een data set	Problemen: ...	Problemen Het is niet inzichtelijk hoe de AI tot een conclusie komt	Problemen:	Problemen: ...
Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...	Voorbeelden: ...
Oorzaken: Gebrekkige of ontbrekende privacy statements of de data wordt niet direct van het data subject vergaard	Oorzaken: ...	Oorzaken: ... de AI opereert als een blackbox	Oorzaken: ...	Oorzaken: ...

Wat zijn volgens u andere knelpunten ten aanzien van AI, bijvoorbeeld technische of maatschappelijke obstakel/problemen? Kunt u daar concrete voorbeelden van noemen? Geef

aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>	<i>Problemen: ...</i>
<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>	<i>Voorbeelden: ...</i>
<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>	<i>Oorzaken: ...</i>

Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe? Geef aub zo specifiek mogelijk aan waar in het proces deze voordelen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop
<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: indirecte discriminatie door AI: een ogenschijnlijke neutrale regel kan indirect een bepaalde groep discrimineren</i>	<i>Bestaande problemen: ...</i>	<i>Bestaande problemen: ...</i>
<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen:</i>	<i>AI gedreven oplossingen: ...</i>	<i>AI gedreven oplossingen: ...</i>

		Statistische check op het resultaat van de regel om te controleren of de regel een bepaalde beschermde groep meer benadeeld dan een andere.		

2. Bestaande Standaarden

2.1 Kent u bestaande beleidsdocumenten, standaarden, codes of conduct of richtsnoeren ten aanzien van AI? Zo ja, welke zijn het veelbelovendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die documenten zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

2.2 Bent u op de hoogte van bestaande wetgeving of jurisprudentie ten aanzien van AI? Zo ja, welke (voeg aub bron toe)?

Welke principes of standaarden uit die wetgeving of jurisprudentie zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.3 Bent u op de hoogte van literatuur waarin additionele standaarden/principes voor AI worden neergelegd? Zo ja, welke studies zijn wat u betreft het belangwekkendst en waarom (voeg aub bron toe)?

Welke principes of standaarden uit die literatuur zijn relevant in het kader van discriminatie, bias, fairness en transparantie? Geef aub zo specifiek mogelijk aan waar in het proces deze te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

<i>(1) Aanleggen dataset</i>	<i>(2) Ontwikkelen/ trainen algoritme</i>	<i>(3) Patronen destilleren uit data</i>	<i>(4) Besluit- en beleidsvorming</i>	<i>(5) Feedbackloop</i>

2.4 Bent u van mening dat deze standaarden (geïdentificeerd in vraag 2.1, 2.2 en 2.3) de problemen die u heeft signaleerd (vraaggroep 1) adequaat adresseren? Zo nee, wat zijn de belangrijkste tekortkomingen/welke principes missen nog? Geef aub zo specifiek mogelijk aan waar in het proces deze aanvullende principes te situeren zouden kunnen zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) Aanleggen dataset	(2) Ontwikkelen/ trainen algoritme	(3) Patronen destilleren uit data	(4) Besluit- en beleidsvorming	(5) Feedbackloop

3. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analysten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analysten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

☒ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Ervaring leert dat developers vaak niet op de hoogte zijn van zulke documenten. In sommige gevallen wordt er wel gebruik gemaakt van development manifesten waarin bepaalde principes uit deze documenten verwerkt zijn.

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analysten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
De bron documenten zijn inderdaad slecht leesbaar voor de eind gebruiker van het systeem. Echter gebruiken sommige organisatie wel interne documentatie die geschreven is voor de eindgebruiker.

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid

tussen groepen' en 'vermijd bias in de dataset' te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
De principes zijn inderdaad te algemeen. De toepassing van principes is altijd context afhankelijk en toepassing vereist interpretatie van de principes en een analyse van het systeem in elk specifiek geval. Specifiekere uitwerking en het gebruik van verschillende voorbeelden zou de bruikbaarheid kunnen vergroten.

☒ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Zie antwoord hierboven

☒ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Er zijn per land veel verschillende richtsnoeren en standaarden. In de Verenigde Staten zijn er bijvoorbeeld al meer dan 20. Door deze wirwar aan documentatie is het voor organisaties niet duidelijk welke documenten te gebruiken.

☐ ☒ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:

☒ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Toezicht en handhaving staan op dit moment nog in de kinderschoenen.

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Zie de bijlage van het begeleidend schrijven bij deze enquête.

7.4 Ingevulde enquêtes (enquête 2)

7.4.1 Respondent 1

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificiële Intelligence (AI) definiëren?

Programmatuur die getraind is.

Wat is uw functie/rol (b.v. data scientist, product owner, project manager, data-analist, data engineer, machine learning engineer, manager, onderzoeker, leidinggevende)?

Data scientist

In welke industrie bent u werkzaam (b.v. gezondheidszorg, financiën, energie, online retail, wetenschap)?

Overheid

Binnen welke AI-domein(en) vinden uw werkzaamheden met name plaats (b.v. computer vision, natural language processing, recommender systems, (geautomatiseerde) besluitvorming)

Besluitvorming, recommender systems

Bij de ontwikkeling of het gebruik van welk soort AI-applicaties bent u betrokken (b.v. fraudedetectie, vraagvoorspelling, customer service)?

Fraudedetectie, vraagvoorspelling

1. Probleemanalyse

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden.

Is het optreden van discriminatie door het gebruik van AI **in het algemeen** volgens u:

Ja Nee

x ☐ Een **belangrijk** probleem
Leg uit: Macht en verantwoording moeten hand-in-hand gaan. Met het toenemende gebruik van AI, moeten de (ethische) standaarden waaraan AI moet voldoen ook beter en beter nageleefd worden.

x ☐ Een **urgent** probleem
Leg uit: Beslissingen die geautomatiseerd kunnen worden lopen niet meer tegen technische maar tegen ethische problemen aan. Derhalve moet daar nu eerst aandacht aan worden besteed.

x ☐ Een **veelvoorkomend** probleem
Leg uit: Het heeft te weinig aandacht gehad en er zijn te weinig handvatten voor.

1.2. Is het optreden van discriminatie door het gebruik van AI in **uw werkdomein** volgens u:

Ja Nee

x x ☒ Een **belangrijk** probleem
Leg uit: Ja en nee. Ja, want er wordt veel aandacht aangegeven en de politiek is er erg mee bezig. Nee, want er zijn in het beslisproces altijd mensen betrokken die ook het laatste woord hebben. Eventuele discriminatie komt dan ook niet door AI.

x x ☒ Een **urgent** probleem
Leg uit: Ook weer ja en nee. Ja, want er wordt veel aandacht aangegeven en de politiek is er erg mee bezig. Nee, want mede door die aandacht zal er nog wel een langere periode menselijke besluitvorming bij elk individueel geval blijven.

☐ x ☒ Een **veelvoorkomend** probleem
Leg uit: Zoals twee vragen eerder: Nee, want er zijn in het beslisproces altijd mensen betrokken die ook het laatste woord hebben. Eventuele discriminatie komt dan ook niet door AI.

In welke AI toepassingen (sectoren/domeinen) verwacht u dat ongewenste discriminatie/ongelijke behandeling zich met name zal voordoen?

Ja Nee

x ☐ Gezondheidszorg
Leg uit: Leeftijdsdiscriminatie is iets waar zelfs openlijk over gesproken wordt

☐ x Welzijn
Leg uit: Discriminatie is hier geen belangrijk onderdeel van het werk.

x ☐ Politie en justitie
Leg uit: De politie gaat bij geweldsdelicten bijvoorbeeld vaak op zoek naar mannen. Er is hier geen tegenbeweging voor, dus dat zal nog wel een tijd zo blijven.

x ☐ Onderwijs
Leg uit: Kijk naar de man-vrouw verhoudingen bij hogleraren. Discriminatie zit hier duidelijk diep ingebakken in het systeem.

Ja Nee

☐ x Huisvesting
Leg uit: Discriminatie is hier geen belangrijk onderdeel van het werk.

☐ x Retail
Leg uit: Pecunia non olet.

x ☐ Arbeidsmarkt
Leg uit: Zie ook hier bijvoorbeeld de man-vrouw verhoudingen bij hogleraren.

☐ ☐ Anders, namelijk...
Leg uit:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) Probleemdefinitie (identificeren bedrijfsprobleem en vertaling naar ML taak), (2) Initiële dataverzameling en analyse, (3) Datavoorbereiding, (4) Modelleren (verschillende modellen uitproberen en een model selecteren), (5) Evaluatie (geselecteerde model evalueren aan de hand van succes criteria vastgesteld in stap 1 en een 'test set') en (6) Implementatie (inclusief monitoring/maintenance activiteiten en terugkoppeling over performance in gebruik). [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

1.4 Waar in het proces zijn volgens u de oorzaken van ongewenste discriminatie te situeren? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e
Problemen: Vooringeno menheden van de experts worden meegenome n in de probleemstel ling	Problemen: Het onevenwicht ig selecteren van data.	Problemen:	Problemen:	Problemen: Er wordt niet gecontroleer d op discriminati e.	Problemen: De beperkingen van het model worden onvoldoende meegegeven aan de gebruikers.
Voorbeelden : Er wordt veronderstel d dat er geen discriminati e is en er wordt dus ook geen aandacht aan besteed	Voorbeelden : Een supervised model met als 'target' de set van gevonden overtreders, terwijl niet iedereen (goed) is gecontroleer d.	Voorbeelden :	Voorbeelden :	Voorbeelden : Een combinatie van (1) en (2): Het gebruik van een niet- representatie ve set voor de controle.	Voorbeelden : Vrijwel elk model dat last heeft van (1), (2) en (5)
Oorzaken: Ik ben geen psycholoog, maar doe een gokje: De modelontwik kelaars en hun omgeving worden niet of zelden gediscrimine erd	Oorzaken: Beperkte controlecapa citeit en vooraf bestaande discriminatie .	Oorzaken:	Oorzaken:	Oorzaken: Ook weer een combinatie van (1) en (2)	Oorzaken: De beperkingen worden onvoldoende gezien door de ontwikkelaar s, laat staan doorgegeven aan de gebruikers.

1.5 Transparantie van AI-besluiten is een *best practice*; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo

specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzame ling	(3) datavorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e
Problemen: <i>Er wordt geen rekening gehouden met transparanti e.</i>	Problemen:	Problemen:	Problemen: <i>Sommige problemen zijn niet goed met transparante modellen op te lossen.</i>	Problemen:	Problemen: <i>Het sluit niet aan bij de intuïtie.</i>
Voorbeelden : <i>In de probleemstel ling wordt alleen aandacht besteed aan efficiëntie en opbrengsten</i>	Voorbeelden :	Voorbeelden :	Voorbeelden : <i>Gezichtsherk enning.</i>	Voorbeelden :	Voorbeelden : <i>Een beslisboom is (mits niet te groot) goed te overzien, zeker voor één bepaald geval. De keuzes in de beslisboom zijn echter niet noodzakelijk allemaal een intuïtief.</i>
Oorzaken: <i>Er wordt niet gestuurd op transparanti e. Soms is transparanti e ook niet wenselijk, bijvoorbeeld bij het bestrijden van vindingrijke criminelen. Dit is</i>	Oorzaken:	Oorzaken:	Oorzaken: <i>Het aantal zaken waar een computer rekening mee moet houden is niet uit te leggen; het probleem is te complex.</i>	Oorzaken:	Oorzaken: <i>Soms vereist een probleem meer complexiteit dan een gebruiker aankan.</i>

overigens het belangrijkste obstakel, de volgende stappen zijn grotendeels transparant te maken.					

1.6 Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe?

Zeker! Het toetsen van vooroordelen is een belangrijke stap in het verminderen van discriminatie. Daarnaast zijn beslissingen die mensen maken van nog beperktere transparantie. Code kun je iig nog nalezen.

2. Mogelijke aanpakken

Verschillende organisaties en bedrijven hebben principes en standaarden ontwikkeld om discriminatie door AI systemen te minimaliseren. Data scientists hebben het hierbij vaak over het reduceren van bias en het zorgdragen voor *fairness*.

2.1 Neemt u of uw organisatie zorgen omtrent discriminatie momenteel mee in keuzes en beslissingen aangaande het gebruik of de inrichting van AI? Zo ja, ten aanzien van welke fase(s)? Leg aub uit.

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e
<i>Meenemen dat er op getoetst moet worden.</i>	<i>Een aselechte steekproef doen.</i>		<i>Indien nodig kiezen voor een uitlegbaar model en experts betrekken.</i>	<i>Toetsen op de aselechte steekproef.</i>	<i>Uitleg geven aan de gebruikers over mogelijke tekortkoming en van het model.</i>

2.2 Zo ja, maakt u of uw organisatie gebruik van specifieke standaarden/principes/richtlijnen (voeg aub bron toe)?

Momenteel is mijn organisatie bezig deze te formaliseren. Op dit moment gebruiken we richtlijnen zoals het gebruik van aselecte steekproeven.

2.3 Welke activiteiten/technieken gebruikt u of uw organisatie om mogelijke pijnpunten in het kader van ongelijke behandeling/discriminatie **te identificeren**?

Ja Nee

- x ☐ Formele risicoanalyse
Leg uit: Een toets op AVG, discriminatie, kwaliteit van modelleren en een op wettelijke basis
- x ☐ Checklists
Leg uit: Er zijn checklists die bepalen of de uitgebreide versie van bovengenoemde toetsen nodig is.
- ☐ x Toolkits
Leg uit: Deze worden momenteel ontwikkeld
- x ☐ In gesprek met gebruikers
Leg uit: Dit wordt meegenomen in gesprekken bij probleemstelling, modellering, evaluatie en implementatie
- x ☐ In gesprek met andere betrokkenen
Leg uit: Er bestaat een ethische commissie die het probleem scant op mogelijke ethische dilemma's.
- ☐ ☐ Anders, namelijk...
Leg uit:

2.4 Welke activiteiten/technieken gebruikt u of uw organisatie om discriminatie/ongelijke behandeling **te meten**?

Ja Nee

- x ☐ **Data op biases** checken aan de hand van fairness criteria. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit: (Hoge correlaties met) bijzondere persoonskenmerken. De criteria voor hoge correlaties zijn verschillend per probleem.
- x ☐ Model evalueren aan de hand van formele fairness criteria bij **model selectie**. Zo ja, welke criteria gebruikt u hiervoor?

Leg uit: Als het model op de onafhankelijke steekproef duidelijk minder presteert. Dit is een duidelijk voorbeeld van selection bias. De criteria zijn verschillend per probleem. Een voorbeeld is een val meer dan 0,04 van de AUROC.

- x ☐ Model evalueren aan de hand van formele fairness criteria bij **model in productie**. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit: Regelmatige check op modellen. De criteria zijn hier verschillend, afhankelijk van de tijd. Ook hier is de aselecte steekproef een belangrijk onderdeel.
- x ☐ Actieve evaluatie ten behoeve van discriminatie/ongelijke behandeling met Gebruikers; Zo ja, *leg uit: Ze worden gevraagd om ook eens naar anderen dan de usual suspects te kijken om zo blinde vlekken te voorkomen.*
Domeinexpert; Zo ja, *leg uit: Ze worden gevraagd naar mogelijke discriminatie.*
Andere betrokkenen; Zo ja, *leg uit: De commissies/personen die de toetsen uit de vorige vraag uitvoeren.*
- ☐ ☐ Anders, namelijk...
Leg uit:

2.5 Welke activiteiten/technieken worden in uw team gebruikt om discriminatie/ongelijke behandeling **te reduceren**?

Ja Nee

- x ☐ Meer data verzamelen
Leg uit: Aselecte steekproef doen.
- x ☐ Analyse van selectie variabelen
Leg uit: Checken op (hoge correlaties met) bijzondere persoonskenmerken.
- x ☐ Data aanpassen/filteren (pre-processing)
Leg uit: Oversamplen van data die minder vaak gecontroleerd is.
- ☐ x Speciale technieken om unfairness te reduceren tijdens het leren (in-processing)
Leg uit:
- ☐ x Speciale technieken om bias te mitigeren van predictions (post processing)
Leg uit:
- x ☐ Anders, namelijk overleg met gebruikers/experts
Leg uit: Het is soms niet inzichtelijk waar eventuele discriminatie zit; het helpt dan om vakmensen te spreken.

3. Juridisch raamwerk

3.1 Bent u op de hoogte van bestaande wetgeving, richtsnoeren of Codes of Conducts ten aanzien van AI? Zo ja, welke en welke spelen specifiek een rol bij de ontwikkeling of het gebruik van AI systemen binnen uw organisatie als het gaat om discriminatie?

Artikel 1 van de grondwet, AVG.

3.2 Op welke moment komt deze wetgeving aan bod in de ontwikkeling of het gebruik van AI systemen binnen uw organisatie en op welke manier?

Toets op AVG, toets op wettelijke basis. Deze komen op meerdere fasen tvan de ontwikkeling en tijdens productie aan de orde.

3.3 Zijn er in het verleden knelpunten geweest bij het operationaliseren van de wetgeving in de ontwikkeling, implementatie en gebruik van AI systemen binnen uw organisatie en hoe heeft uw organisatie (of is zij voornemens) deze knelpunten te adresseren?

De wet is geschreven door juristen, modellen worden gebouwd door beta's. Het vertalen van de wet naar technische voorschriften is lastig. Mijn organisatie is hier mee bezig.

3.4 Wat zou u idealiter van een wet of code of conduct verwachten ten aanzien van het bestrijden van discriminatie en ongelijke behandeling in AI-systemen?

Een duidelijk omschreven doel. De regels kunnen dan samen met data-analisten uitgewerkt worden.

4. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analisten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analisten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- x ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: De juridische stukken zijn in de verkeerde taal en de academische stukken hebben geen rechtsgeldigheid.
- x ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analisten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: Dit geldt voor de juridische stukken
- x x De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit: Deels: voor een deel van de data-analisten zal dit altijd een ver-van-mijn-bed-show blijven, maar er zijn voldoende mensen die iets kunnen met deze richtlijnen.

- x x De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: Hetzelfde antwoord als hierboven.

- x ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Dit geldt voor de academische stukken; het zijn er veel en ze hebben geen rechtsgeldigheid.

- x ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Kijk naar het onderzoek dat jullie doen; dat gaat ook bijdragen aan een verbetering hierin. En het feit dat jullie het onderzoek doen geeft aan dat het dus nu nog onvoldoende is.

- x ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: De toezichthouders hebben last van dezelfde problemen als de data-analisten. Daarnaast is het besef onder bestuurders nog groeiende.

- x ☐ Anders, namelijk: *Er wordt onvoldoende op gestuurd. Dit is het belangrijkste. Data-analisten doen (net als andere mensen) grotendeels waarvoor ze betaald krijgen. En als dat is voor niet-discriminerende, transparante modellen, dan gaan ze die maken, hoe ingewikkeld ook.*

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Goede samenwerking tussen universiteiten en de wetgever; dat maakt beide relevanter.

7.4.2 Respondent 2

Name:

Organization:

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Any human-like behavior a machine can exhibit.

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Data Engineer

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

Online Retail

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Data Preparation.

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

Customer Service

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒ ☐ An **important** problem

Explain: In some cases we should avoid discrimination in order to uncover real pattern that would help us better understand the AI model.

☐ ☐ An **urgent** problem

Explain:

☐ ☐ A **common** problem

Explain:

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☒ ☐ An **important** problem

Explain: Choose not to disclose

☐ ☐ An **urgent** problem

Explain:

☐ ☐ A **common** problem

Explain:

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes ☐ No ☒ Healthcare
Explain: Illness can be race/sex specific. But this is beneficial and not unwanted

Yes ☒ No ☐ Housing
Explain: Country of origin plays an important role in housing sector. This might be exhibited by the AI model as well.

☐ ☒ Wellness
Explain: Same as healthcare

☐ ☒ Retail
Explain:

☒ ☐ Policing/criminal law
Explain:

☒ ☐ Labourmarket/employment
Explain:

☒ ☐ Education
Explain:

☐ ☐ Otherwise, namely
Explain:

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> <i>Improper scaling of data</i>	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...
<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> <i>samples belong to similar traits can be under/over represented</i>	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...

<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: Imbalanced data.</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: Understandi ng decisions of complex model is not so easy</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: A non- fraudulent customer can be classified as a fraud without proper reasoning by the model.</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: Black-box nature of AI models</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

Chose not to disclose.

(1) problem definition	(2) <i>data</i> collection	(3) <i>data</i> preparation	(4) modelling	(5) evaluation	(6) implementati on

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

Chose not to disclose.

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Yes No

☒ ☐ Formal Risk Assessment
Explain:

☐ ☐ Checklists
Explain:

☐ ☐ Toolkits
Explain:

☒ ☐ Conversations with users
Explain:

☒ ☐ Conversations with other parties involved
Explain:

☐ ☐ Otherwise, namely...
Explain:

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Yes No

☒ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?
Explain: Imbalance in dataset

☐ ☐ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?
Explain:

☐ ☐ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?
Explain:

☐ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain....
- Other people involved; If so, explain....

☐ ☐ Otherwise, namely...
Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

Yes No

☒ ☐ Collect more data
Explain:

☒ ☐ Analysis of selection variables
Explain:

☐ ☐ Adjusting / filtering data (pre-processing)
Explain:

☒ ☐ Special techniques to reduce unfairness during learning (in-processing)
Explain:

- ☐ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

No.

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

Releasing the model to PRO environment. Not sure about the way.

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

I am not aware about it.

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

Set a standard procedure which would evaluates the model and give a "biasness score". This biasness score can be used to decide if the model should be considered discriminatory or not.

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☐☐

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain:

☐☐

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain:

☒☐

The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain:

- ☐ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain:

- ☐ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain:

.....

- ☐ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain:

.....

- ☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles

Explain:

.....

- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

7.4.3 Respondent 3

Naam: Maarten Stol

Organisatie: BrainCreators

0. Algemeen

Hoe zou u zelf in 1 zin Artificial Intelligence (AI) definiëren? *De huidige, moderne toegepaste AI is het verplaatsen van een probleem: van het domein van menselijk probleem oplossend vermogen, naar het domein van geautomatiseerde algoritmische, logische, en/of statistische methoden, met de bedoeling dat de nadelen van deze verplaatsing gecompenseerd worden door de eventuele voordelen.*

Wat is uw functie/rol (b.v. data scientist, product owner, project manager, data-analist, data engineer, machine learning engineer, manager, onderzoeker, leidinggevende)? *Leidinggevende (van het research team)*

In welke industrie bent u werkzaam (b.v. gezondheidszorg, financiën, energie, online retail, wetenschap)? *IT innovatie, toepassing geautomatiseerd asset management in industrie, voornamelijk gebaseerd op visuele kunstmatige intelligentie.*

Binnen welke AI-domein(en) vinden uw werkzaamheden met name plaats (b.v. computer vision, natural language processing, recommender systems, (geautomatiseerde) besluitvorming)? *Geautomatiseerde besluitvorming, voornamelijk op basis van computer vision en machine learning.*

Bij de ontwikkeling of het gebruik van welk soort AI-applicaties bent u betrokken (b.v. fraudedetectie, vraagvoorspelling, customer service)? *Visual object recognition (voornamelijk prototyping vooraf aan de productiefase) toegepast in de asset management in infrastructuur, real-estate, en manufacturing.*

1.3 In welke AI toepassingen (sectoren/domeinen) verwacht u dat ongewenste discriminatie/ongelijke behandeling zich met name zal voordoen?

NB: Moeilijk te beantwoorden, helaas weet ik te weinig van elk van deze sectoren afzonderlijk. In het algemeen verwacht ik dat op elke sector vormen van discriminatie kunnen en zullen gaan voorkomen. Maar (zie mijn Ai definitie) de vraag is telkens: zijn de nadelen de voordelen waard? En, hoe gaan we het beste om met de nieuwe nadelen? Voorbeeld: in de gezondheidszorg en geneeskunde is bekend dat sommig onderzoek te gender-eenzijdig geweest is (denk: medicijnen die alleen goed op mannen getest zijn). Deze trend kan zich in het begin voortzetten en zelfs versterken bij de invoering van verdere automatische besluitvorming. Aan de andere kant maken algoritmische methoden wellicht op de lange termijn mogelijk om een meer individuele gezondheidszorg te gaan aanbieden zoals dat eerder nog niet mogelijk was (denk aan behandeling ahv gen-testing en uitgebreide individuele data verkregen via smart-devices) Zijn de nieuwe nadelen het waard? Waarschijnlijk wel? Hebben we een keuze? Ja, maar we begrijpen nog niet goed welke.

☐ ☐ Huisvesting
Leg uit:

☐ ☐ Retail
Leg uit:

☐ ☐ Arbeidsmarkt
Leg uit:

☐ ☐ Anders, namelijk...
Leg uit:

Ja Nee

☐ ☐ Gezondheidszorg
Leg uit:

Ja Nee

☐ ☐ Welzijn
Leg uit:

☐ ☐ Politie en justitie
Leg uit:

☐ ☐ Onderwijs
Leg uit:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) Probleemdefinitie (identificeren bedrijfsprobleem en vertaling naar ML taak), (2) Initiële dataverzameling en analyse, (3) Datavoorbereiding, (4) Modelleren (verschillende modellen uitproberen en een model selecteren), (5) Evaluatie (geselecteerde model evalueren aan de hand van succes criteria vastgesteld in stap 1 en een ‘test set’) en (6) Implementatie (inclusief monitoring/maintenance activiteiten en terugkoppeling over performance in gebruik). *[Als u een andere fasering voor ogen heeft, geef dat gerust aan]*

1.4 Waar in het proces zijn volgens u de oorzaken van ongewenste discriminatie te situeren? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e
Problemen:	Problemen:	Problemen:	Problemen:	Problemen:	Problemen:
Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :
Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:

Probleemdefinitie: Dekt de definitie de klant vraag eigenlijk wel goed? Zijn er onbewust compromissen gesloten bij het definiëren? Zijn er verborgen consequenties, eventueel zelfs buiten het oorspronkelijke toepassingsgebied?

Dataverzameling: In theorie gaat machine learning er van uit dat de deployment data uit dezelfde distributie komt als de training data. In de praktijk is dit meestal niet zo. Soms heeft dit geen belangrijke consequenties. Soms zijn de consequenties niet duidelijk. Soms zijn de consequenties niet acceptabel. De vraag is: wie beoordeelt dit? Als de mogelijke consequenties van maatschappelijke aard kunnen zijn dan is de machine learning engineer wellicht niet de juiste persoon voor deze beoordeling. Dit geldt voor punt 2 t/m 5.

Implementatie: Hieronder versta ik de deployment van de live applicatie. Een Ai applicatie heeft een levensloop, gedurende welke onvermijdelijk veranderingen optreden. Veranderingen kunnen zich bevinden in: de aannames waaronder de applicatie gebruikt wordt door mensen en andere applicaties, het doel waarvoor de applicatie gebruikt wordt (creatief hergebruik komt vaak voor in software), de doelgroep van personen die door de applicatie beïnvloed worden, etc. Een nieuwe uitdaging rond de deployment van Ai applicaties is dat ze een nieuwe, continue verantwoordelijkheid opwerpen om al deze punten te blijven monitoren. Dit kan soms zeer kostbaar zijn, en hierin schuilt het risico. Er zal vraag komen naar een nieuw soort management: algoritme management, die al deze aspecten behelst en controleert.

1.5 Transparantie van AI-besluiten is een *best practice*; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van

transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e
Problemen:	Problemen:	Problemen:	Problemen:	Problemen:	Problemen:
Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :
Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:

Transparantie en verklaarbaarheid: ik heb dit in een lezing ooit gekstcherend “het eind van de verlichting” genoemd: de mens is sinds de verlichting in staat de wereld om hen heen te begrijpen. Ook relevant is dat de mens zich deze verklaring kan veroorloven in termen van energie, tijd, moeite, kosten, voeding, grondstoffen, etc. (Industriële revolutie, het ontstaan van de middenklasse, en fossiele brandstoffen waren hiervoor essentieel). Het is niet ondenkbaar dat de mensheid zo afhankelijk wordt van intransparante Ai en onverklaarbare controlemechanismen, dat er een horizon (of zelfs een eind) komt aan het begrijpen van de wereld. Dit is minder gek dan het klinkt. Als we complexe beslissingen moeten maken in een fractie van een seconde, op zeer grote schaal, voor zeer veel mensen, en dat bovendien moeten doen voor strakke deadlines en binnen het budget van de investeerder, en dat de enige algoritmen die dit kunnen intransparant zijn zoals in de deep learning, dan zou het zomaar kunnen dat verklaringen een luxe worden. Of zelfs iets antieks. Alleen weggelegd voor de zeer rijken, als überhaupt al mogelijk (het energietekort en de groeiende wealth-gap zijn ook hier weer relevante signalen).

1.6 Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe?

Ik verwacht van wel. Ai is gereedschap dat we kunnen verfijnen zoals we zelf willen. De uitdaging zit in begrijpen hoe, waar, wanneer, etc. Een belangrijke stap zou zijn dat we voor elke Ai applicatie niet alleen correlaties modelleren tussen de input en gewenste output, maar dat we ook een causaal model maken van het data generatie proces. Dat stelt ons wellicht in staat beter te controleren onder welke omstandigheden het model van gedrag verandert, of moet veranderen. Discriminatie zou dan gezien worden als een generalisatie probleem binnen de machine learning dat betrekking heeft op verschillen tussen training en deployment datadistributie, en op die manier gekwantificeerd en gecontroleerd kunnen worden. Deze aanpak alleen zal waarschijnlijk niet alle vormen van discriminatie tegen kunnen gaan, omdat discriminatie ook kan optreden door fouten op het systeemniveau (denk aan integratie van verschillende applicaties), maar het zou een goed begin kunnen zijn.

2. Mogelijke aanpakken

Verschillende organisaties en bedrijven hebben principes en standaarden ontwikkeld om discriminatie door AI systemen te minimaliseren. Data scientists hebben het hierbij vaak over het reduceren van bias en het zorgdragen voor *fairness*.

2.1 Neemt u of uw organisatie zorgen omtrent discriminatie momenteel mee in keuzes en beslissingen aangaande het gebruik of de inrichting van AI? Zo ja, ten aanzien van welke fase(s)? Leg aub uit.

Nee, niet of nauwelijks. Hoogstens op informeel niveau, en zonder systematische aanpak.

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e

2.2 Zo ja, maakt u of uw organisatie gebruik van specifieke standaarden/principes/richtlijnen (voeg aub bron toe)?

2.3 Welke activiteiten/technieken gebruikt u of uw organisatie om mogelijke pijnpunten in het kader van ongelijke behandeling/discriminatie **te identificeren**?

Ja Ne
 e

☐ X Formele risicoanalyse
Leg uit:

☐ X Checklists
Leg uit:

☐ X Toolkits
Leg uit:

X ☐ In gesprek met gebruikers
Leg uit: Voor de kennisoverdracht van menselijk domeinexpert naar het Ai platform spreken wij vaak informeel maar lang met betrokken experts. Deze gesprekken zijn de enige plek waar pijnpunten geïdentificeerd zouden kunnen worden in onze huidige werkwijze.

☐ X In gesprek met andere betrokkenen
Leg uit:

☐ ☐ Anders, namelijk...
Leg uit:

2.4 Welke activiteiten/technieken gebruikt u of uw organisatie om discriminatie/ongelijke behandeling **te meten**?

Toelichting: wij meten en werken vaak aan data bias in de onbeladen zin van het woord. Machine learning classificatie kent altijd bias die gecontroleerd moet worden. Daar zijn allerlei methoden voor, van het balanceren van trainingsdata, het aanpassen van model evaluatie, tot het meten van model performance in zogenaamde data-slices (subsets van data die niet overeenkomen met een classificatie categorie, maar wel betekenisvol zijn aangaande de deployment). Omdat onze toepassingen over het algemeen niet over personen gaan, zien wij deze meetmethoden niet in het licht van maatschappelijke fairness. Toch staan de methoden klaar om gebruikt te worden zo gauw we ons wel op dat gebied zouden begeven.

Ja Nee

☐ X **Data op biases** checken aan de hand van fairness criteria. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:

☐ X Model evalueren aan de hand van formele fairness criteria bij **model selectie**. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:

☐ X Model evalueren aan de hand van formele fairness criteria bij **model in productie**. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:

☐ X Actieve evaluatie ten behoeve van discriminatie/ongelijke behandeling met Gebruikers; Zo ja, *leg uit....*
Domeinexpert; Zo ja, *leg uit....*
Andere betrokkenen; Zo ja, *leg uit*

☐ X Anders, namelijk...
Leg uit:

2.5 Welke activiteiten/technieken worden in uw team gebruikt om discriminatie/ongelijke behandeling **te reduceren**?

Zie toelichting bij 2.4

Ja Nee

- ☐ ☐ Meer data verzamelen
Leg uit:
- ☐ ☐ Analyse van selectie variabelen
Leg uit:
- ☐ ☐ Data aanpassen/filteren (pre-processing)
Leg uit:
- ☐ ☐ Speciale technieken om unfairness te reduceren tijdens het leren (in-processing)
Leg uit:
- ☐ ☐ Speciale technieken om bias te mitigeren van predictions (post processing)
Leg uit:
- ☐ ☐ Anders, namelijk ...
Leg uit:

3. Juridisch raamwerk

3.1 Bent u op de hoogte van bestaande wetgeving, richtsnoeren of Codes of Conducts ten aanzien van AI? Zo ja, welke en welke spelen specifiek een rol bij de ontwikkeling of het gebruik van AI systemen binnen uw organisatie als het gaat om discriminatie?

Wij nemen uiteraard bestaande wetgeving mee in ons applicatie design. Tot nu toe heeft dit echter meestal betrekking op GDPR, software licensing, en vragen rond data eigendommen. Discriminatie en andere maatschappelijke thema's met betrekking tot fairness komen hier op dit moment niet of nauwelijks in aan de orde.

3.2 Op welke moment komt deze wetgeving aan bod in de ontwikkeling of het gebruik van AI systemen binnen uw organisatie en op welke manier?

Zie toelichting bij 3.1

3.3 Zijn er in het verleden knelpunten geweest bij het operationaliseren van de wetgeving in de ontwikkeling, implementatie en gebruik van AI systemen binnen uw organisatie en hoe heeft uw organisatie (of is zij voornemens) deze knelpunten te adresseren?

Nee.

3.4 Wat zou u idealiter van een wet of code of conduct verwachten ten aanzien van het bestrijden van discriminatie en ongelijke behandeling in AI-systemen?

Toetsingsmethoden in plaats van doelen alleen. Het is prima om te zeggen dat we bepaald gedrag van de applicatie niet willen accepteren. Het is echter een heel andere vraag hoe we een applicatie hieraan systematisch kunnen toetsen. Zolang we alleen spreken in termen van doelen, blijft ethische Ai niet veel meer dan een wishlist. Wat we nodig hebben zijn strak geformuleerde methoden die als maatstaf kunnen dienen tijdens het design, en voor/na de deployment van de applicatie. Dit hoeft niet te resulteren in een enkele eendimensionale score die boven een drempelwaarde moet uitkomen. Het is al goed genoeg als er een scala aan scores is die als uitkomst kunnen dienen met betrekking tot verschillende criteria. Als er maar een eenduidig uitvoerbare test geformuleerd is die iedereen op dezelfde manier kan toepassen.

4. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analisten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analisten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- X ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken
Leg uit: Vaak is er een beperkt budget vanuit een klant, en dient de toepassing een dubbel doel: het probleem zelf oplossen, en tegelijk een testcase tonen die binnen de organisatie de potentie van Ai demonstreert. Veel organisaties zijn in het beginstadium van hun Ai adoptie niet klaar om de grondige aanpak te rechtvaardigen die nodig is om de hun engineers ook nog eens deze richtlijnen op te pakken.
- X ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analisten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt
Leg uit: ...of ze in elk geval niet in staat stelt de overbrugging te maken van hun eigen technisch overwegingen naar de maatschappelijke, ethische doelen.
- X ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid

tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk
Leg uit: Zie toelichting bij 3.4. Het probleem zit niet in de algemeenheid, maar in het verschil tussen doel en methode. Transparantie is een doel. Wat een engineer nodig heeft is een eenduidig uit te voeren toets die hen in staat stelt om hun werk te evalueren in termen van fairness, discriminatie, en maatschappelijke waarden. Doelen zijn prima, maar niet genoeg voor hen die de applicaties ook echt moeten maken. Nog een voorbeeld: software engineers (dus niet Ai) zijn gewend om, voordat ze hun algoritmes schrijven, eerst zogenaamde unit-tests te schrijven, ie, een toets waar het gedrag van de software aan moet voldoen. Wat men wellicht nodig heeft zijn (een soort van) unit-tests voor Ai die toetsen op maatschappelijke waarden.

- ☐ X De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis, interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit: Dit hoeft niet het probleem te zijn. Er is geen reden waarom universele principes niet voldoende zouden zijn om specifiek toepassingen aan af te meten. Opnieuw, het verschil is tussen doelen en toetsingsmethoden. Zie 3.4
- ☐ X Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit: Ook dit hoeft geen probleem te zijn. Ik neem aan dat er veel overlap is tussen verschillende standaarden. Dat kenmerkt de vroege fase van Ai waar we ons nu in bevinden. Dit zal op den duur meer unificeren.
- X ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit: Misschien. Voldoen aan standaarden kost tijd en geld. Het risico en de consequenties om er niet aan te voldoen zijn wellicht nog niet goed duidelijk. Net als met de GDPR zal er waarschijnlijk jurisprudentie moeten ontstaan voordat men heldere voorbeelden heeft om deze keuzes te navigeren.
- X ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit: Ja. Zie vorig punt.
- ☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design* voor AI Systemen?

Ik kan niet genoeg het belang benadrukken van eenduidig uitvoerbare toetsingsmethoden. Er zijn genoeg doelen geformuleerd. Als we in de voedingsindustrie zeggen dat een product niet meer dan X% van een familie aan giftige bijproducten mag bevatten, dan wordt er ook een scheikundige toets ontwikkeld om het product mee te keuren. Ook het productieproces zelf kan op deze manier eisen aan toetsen verbinden, bijvoorbeeld niet meer dan zoveel liter zoetwatergebruik per kilo product. Dit is meetbaar, en toetsbaar. Zonder toets heeft het geen enkele zin om opnieuw een lijst met doelen te maken, dan blijft het bij een wishlist.

Een eerste start voor Ai is inderdaad werken met bucket lists zoals in deze enquête: probleemdefinitie, datavoorbereiding, etc. Hier wordt het designproces langs een tijdslijn gelegd. Er zouden voor elke bucket apart toetsings methoden ontwikkeld kunnen worden om zo het designproces te sturen. Er zijn echter ook andere indelingen die wellicht bruikbaar zouden kunnen zijn.

Bijvoorbeeld, bij BrainCreators zien we het designproces ook langs lijnen van domein-expertise kennis transfer. Op welke manieren kan expertise en kennis van een domein expert zijn weg vinden naar de automatisering op het platform? Open deuren zijn: gelabelde data, en rule based decision making. Er zijn echter nog andere wegen, zoals bijvoorbeeld regularisatie die geïnformeerd wordt door expliciete domein kennis tijdens de training (regularisatie in de machine learning sense van het woord: het introduceren van gewenste bias gedurende de training fase van het model), het causaal modelleren van het data generatie proces om zo beter te begrijpen hoe het model zich zal gedragen onder interventies die resulteren in een afwijkende data distributie (en dus afwijkend model gedrag), en het dirigeren van het samenwerken van meerdere (of alle) modellen die een organisatie in gebruik heeft dmv workflows die eenduidig te checken zijn op (on)gewenste eigenschappen.

Als u hier verdere vragen over heeft, aarzel dan niet om contact op te nemen. Ons bedrijf is zeker bereid mee te denken aan dit belangrijke onderwerp en de oplossingen die hiervoor nodig zullen zijn.

Vriendelijke groet,
Maarten Stol, BrainCreators
maarten.stol@braincreators.com

7.4.4 Respondent 4

Name: Machine learning researcher

Organization: Multi-national - digital product vendor

0. General

How would you define Artificial Intelligence (AI) in one sentence?

AI in the practice of detecting generalizable patterns out of large collections of data and making machines react on them.

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Machine learning researcher

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

Digital product/service vendor

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Computer vision

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

B2B digital product

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒☐

An

important

problem

Explain: AI is becoming a part of the everyday life more and more each year. The bias and discrimination in AI algorithms can easily be hidden from the users and even experts.

☒☐

An

urgent

problem

Explain: AI algorithms are starting to be used in critical aspects of life. While a bias in a resume screening program can cause many frustrations, discrimination in an algorithm that is designed to help a judge can lead to a horrible mistake.

☐☒

A

common

problem

Explain: I think a good deal of such problems are discovered and tackled by the scientists in the algorithm development phase, before getting to products and touching people's lives.

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☐☒

An

important

problem

Explain: As like many applications of AI, my work domain is to help improve quality of life. While bias in my domain is unwanted too, honest algorithm mistakes may lead to poor quality in service received rather than a costly harm.

☐☒

An

urgent

problem

Explain: Again, as above.

☐☒

A

common

problem

Explain: I cannot imagine many aspects in my work that the algorithms may lead to discrimination.

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes No

☒☐

Healthcare
Explain: Unlike the other cases, the bias in healthcare and wellness algorithms may less likely cause problems of ethical nature. However, bias towards a specific set of symptoms in a diagnosis algorithm can push doctors to make wrong decisions.

☒☐

Wellness
Explain: The health monitoring smart-wear may provide biased advice given a bias in the training set.

☒☐

Policing/criminal law
Explain: Today there are machine learning algorithm to help judges decide on complicated criminal cases.

☐☒

Education
Explain: I am not aware of any case.

Yes No

☒☐

Housing
Explain: The banks may use blackbox algorithms to decide whether to trust a person with mortgage.

☐☒

Retail
Explain: I am not aware of an example.

☒☐

Labourmarket/employment
Explain: Resume reviewing algorithms can be biased towards a specific characteristic that is fed to it as training data.

☐☐

Otherwise, namely
Explain:

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance

in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: collecting data from a limited set which is not inclusive of all possible varieties in the population</i>	<i>Problems: treating all data samples the same when preparing the data.</i>	<i>Problems: ...</i>	<i>Problems: It is common practice to evaluate algorithms on another set of data. If that data is not selected properly from the whole prepared dataset, the algorithm will have biases.</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: collecting face images of only white- colored people for a face detection algorithm</i>	<i>Examples: Cropping the input image to deep learning algorithms to a fixed size is common practice. It is wrong to crop all images with different aspect ratios and zoom levels to the same crop size.</i>	<i>Examples: ...</i>	<i>Examples: When members of several classes do not exist in the test dataset of a classificatio n algorithm.</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: The algorithm will be</i>	<i>Causes: the algorithm will be</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

	<i>unlikely to detect black-colored faces.</i>	<i>biased towards a specific aspect ratio in input images.</i>			

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: the bias in algorithms happen when the data of a group is not collected. People today are mostly concerned about how their data is collected rather than why their data is not collected.</i>	<i>Problems: ...</i>	<i>Problems: modelling AI algorithms is a complicated task that requires specific skillset. It is not a trivial task to make them understandable to all citizens</i>	<i>Problems: ...</i>	<i>Problems: AI implementations are usually considered proprietary to companies. Making them transparent to all citizens can cause competition and loss of value.</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

Yes, an AI with properly collected and prepared representative data can produce results that is against intuitions that we have and may subconsciously involve in our decisions.

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
			<i>It is expected from the engineers to design models that are efficient for a large variety of target costumers for the company. This means no bias toward a group but cover as many groups as possible</i>		

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

Unfortunately, I am not aware of any resources that my company uses in this regard.

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Yes No

- ☐ ☐ Formal Risk Assessment
Explain:
- ☐ ☐ Checklists
Explain:
- ☐ ☐ Toolkits
Explain:
- ☒ ☐ Conversations with users
Explain: For parts of our products, we get feedback from end users of software. The company tries to address these feedbacks whether they involve discrimination or not.
- ☒ ☐ Conversations with other parties involved
Explain: As my company is a B2B enterprise, we get constant feedback on services from the businesses, however concerns about discrimination falls into the costumer business's choice.
- ☐ ☐ Otherwise, namely...
Explain:

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Yes No

- ☒ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?
Explain: The data collection is a costly procedure for my company and we have limited resources for that. Thus, the targets for data collection are selected beforehand with as much diversity coverage as possible.
- ☐ ☒ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?
Explain: I am not aware of any
- ☐ ☒ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?
Explain: I am not aware of any
- ☐ ☒ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain....
- Other people involved; If so, explain....
- ☐ ☐ Otherwise, namely...
Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

Yes No

☒	☐	Collect more data <i>Explain: Though our team is not responsible for collecting data, the researchers always emphasis on the effect of larger and diverse datasets on AI performance and convey this to responsible teams.</i>
☐	☒	Analysis of selection variables <i>Explain:</i>
☒	☐	Adjusting / filtering data (pre-processing) <i>Explain: long and diverse procedures of analysing the raw data, from eyeballing to statistics. Trying to find the reason for odd or missing values.</i>
☒	☐	Special techniques to reduce unfairness during learning (in-processing) <i>Explain: The researchers always keep an eye on the AI literature. However, in my field of work, unfairness in models are not a recurring topic of research.</i>
☐	☐	Special techniques to mitigate bias of predictions (post processing) <i>Explain:</i>
☐	☐	Otherwise, namely... <i>Explain:</i>

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

The only ones that I am aware of are related to GDPR. This means that specific sorts of data cannot be used as input to AI or will not be kept in the company for later access of the models.

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

The legal team is involved in collection, processing and storage of data in our company.

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

It should prevent basing decisions on critical fields absolutely on algorithm outputs (for the current status of algorithms). It should also guarantee that the people using these AI systems are experienced enough. That is to prevent people pressing buttons to run AI algorithms and basing their decisions on that without knowing the procedures of producing input and steps towards output.

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☒ ☐ The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: The companies should be forced to create training for employees to make such principles known to them.

☒ ☐ The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: As these documents are intended for developers and engineers, they should take rather a scientific language rather than a legal one.

☒ ☐ The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain: division of requirements based on sectors and even algorithms can be helpful for the whole society

- ☐ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI
Explain:
- ☐ ☒ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)
Explain: I think there is a lack of communication channel to introduce the right standards to the right group of analysts. The number of guidelines is not the issue here.
- ☐ ☒ The standards have only recently been developed; it takes time for these to be adopted in practice
Explain: Depending on the investments and efforts the governments make on forcing standards, the acceptance time could be rather short
- ☒ ☐ There is a lack of supervision and enforcement of the existing standards and principles
Explain:
- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems* (which will also appear in English)?

While it is useful to gather opinions of people who are in the AI field on this matter, I think people with little idea on AI should also be explained about the role of AI in their life and then be asked if they have any experience with discrimination against them.

7.4.5 Respondent 5

Name: Sako Arts

Organization: Wolfpack

0. General

How would you define Artificial Intelligence (AI) in one sentence?

A manmade agent that shows intelligence

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

AI practice lead

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

Technology – Software Engineering

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Hyper Personalization, decision-making, recommender systems, anomaly detection, trend forecasting, AI in deployment

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

Risk detection, behaviour analysis, credit and dunning automation, sales forecasting, intrusion detection

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒ ☐ An **important** problem

Explain: While discrimination is the whole goal of AI in many cases you should be very aware which features are of influence on the prediction and if that makes sense/ is desirable.

☒ ☐ An **urgent** problem

Explain: Too often discrimination is not noticed since not enough analysis is done on the produced models

☒ ☐ A **common** problem

Explain: Discrimination on features that make no sense is very common, discrimination on ethical features like race, gender or ethnicity is less common. Often data scientist are aware of these features and exclude them.

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☒ ☐ An **important** problem

Explain:

☒ ☐ An **urgent** problem

Explain:

☒ ☐ A **common** problem

Explain:

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes No

☐ ☐ Healthcare
Explain:

Yes No

☒ ☐ Housing
Explain:

☐ ☐ Wellness
Explain:

☐ ☐ Retail
Explain:

☒ ☐ Policing/criminal
law
Explain:

☒ ☐ Labourmarket/employment
Explain:

☐ ☐ Education
Explain:

☐ ☐ Otherwise, namely
Explain:

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

The part where I think is most common is woven within 1, 2 and 3. Basically data research. This starts with the problem definition but the most important part is the feature definition. One should always explicitly research each available feature and see what it actually encodes in context of the domain and the problem definition. By doing this you will always find some features that are poorly understood by the business and could contribute to overfitting, information leakage or unwanted discrimination. A simple example could be a timestamp that is always empty in accordance with one of your target categories. A more advanced one is a categorical feature that encodes a specific file number, say D-07, which unknowingly for the data scientist is only used for alien residents. So unknowingly the data specifically encodes for foreigners.

Secondary in the validation stage it should be thoroughly research which features (heavily) influence the prediction. This could indicate overfitting or information leakage but also unexpected features that unknowingly encode for discrimination like mentioned above.

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on

<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...
<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

Well that would be in 5 and 6, however, in many situations you are working with black box models. You should always create some insight in them using model achnotic methods like LIME or SHAP, but these offer more apprcimations which makes it impossible to be transparent about

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...
<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

Well yes, since ML is slave of your data it can reveal discriminating features that have been unproportionally correlated in your training data with your target question.
Furthermore anomaly detection could play a role here.

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

(1) <i>problem definition</i>	(2) <i>data collection</i>	(3) <i>data preparation</i>	(4) <i>modelling</i>	(5) <i>evaluation</i>	(6) <i>implementati on</i>
x	x	x		x	

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

None specific

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Yes No

☐ ☐ Formal Risk Assessment
Explain:

☐ ☐ Checklists
Explain:

☐ ☐ Toolkits
Explain:

☒ ☐ Conversations with users
Explain:

☒ ☐ Conversations with other parties involved
Explain:

☒ ☐ Otherwise, namely...
Explain: Data research, model evaluation and analysis

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Yes No

- ☐ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?
Explain:
- ☐ ☐ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?
Explain:
- ☐ ☐ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?
Explain:
- ☒ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain....
- Other people involved; If so, explain....
- ☒ ☐ Otherwise, namely...
Explain: Simply analysing the results and see if unexpected results appear in model performance or feature importance

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

Yes No

- ☐ ☐ Collect more data
Explain:
- ☒ ☐ Analysis of selection variables
Explain:
- ☒ ☐ Adjusting / filtering data (pre-processing)
Explain:
- ☐ ☐ Special techniques to reduce unfairness during learning (in-processing)
Explain:
- ☐ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☐☐

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain:

☐☐

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain:

☐☐

The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain:

- ☐ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI
Explain:
- ☒ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)
Explain:
- ☐ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice
Explain:
- ☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles
Explain:
- ☒ ☐ Otherwise, namely, they are not presented in an intuitive and simple way. They should be more presented like the many cheat sheets and infographics you can find in AI blog posts.

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

7.4.6 Respondent 6

Name: Sagar Gugwad

Organization: Prefer not to disclose

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Controlled cognitive abilities by human defined system

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Data engineer/ machine learning engineer

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

insurance, online retail and mining

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Recommender systems, Natural language processing

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

Customer service, mainly intent detection for customer care services

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒ ☐ An **important** problem
Explain: AI systems malformed cognitive abilities might lead to discrimination.

☐ ☐ An **urgent** problem
Explain:

☐ ☐ A **common** problem
*Explain: Discrimination cannot occur because the way how AI systems are trained.
Discrimination*

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☒ ☐ An **important** problem
Explain: It might affect people sentiments and feelings leading to distressful situations.

☐ ☐ An **urgent** problem
Explain:

☐ ☐ A **common** problem
Explain:

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes No
☐ ☐ Healthcare
Explain:

Yes No
☐ ☐ Housing
Explain:

☐ ☐ Wellness
Explain:

☐ ☐ Retail
Explain:

☐ ☐ Policing/criminal
law
Explain:

☒ ☐ Labourmarket/employment
*Explain: This sector
generally experiences lots
of discrimination
irrespective of AI.*

☐ ☐ Education
Explain:

☐ ☐ Otherwise, namely
Explain:

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: Data domain might have have biased</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: Consider for example of a bank where all the employees are male. The collected</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>

	<i>data will be biased in terms of gender.</i>				
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: It would be difficult to convince user on the decision made. If the model is trained for particular context.</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

Not sure, AI will help to reduce but certainly it might introduce few more bias.

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain. No not got chance to consider this problem.

(1) <i>problem definition</i>	(2) <i>data collection</i>	(3) <i>data preparation</i>	(4) <i>modelling</i>	(5) <i>evaluation</i>	(6) <i>implementati on</i>

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Yes No

- ☒ ☐ Formal Risk Assessment
Explain: Risk Assessment carried out in general might have some pointers regarding discrimination. There is no dedicated evaluation.
- ☐ ☐ Checklists
Explain:
- ☐ ☐ Toolkits
Explain:
- ☐ ☐ Conversations with users
Explain:
- ☐ ☐ Conversations with other parties involved
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Yes No

- ☐ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?
Explain:
- ☐ ☐ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?
Explain:
- ☐ ☐ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?
Explain:
- ☒ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; The domain expert will be helpful to evaluate unequal treatment.
- Other people involved; If so, explain....
- ☐ ☐ Otherwise, namely...
Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

Yes No

- ☐ ☐ Collect more data
Explain:
- ☐ ☐ Analysis of selection variables
Explain:
- ☐ ☐ Adjusting / filtering data (pre-processing)
Explain:
- ☐ ☐ Special techniques to reduce unfairness during learning (in-processing)
Explain:
- ☐ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

Not aware

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

N/A

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

N/A

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

N/A

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☐☒

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: Most of the big organizations have defined guidelines for developing AI systems.

☐☒

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: Most of the guidelines are explained from the user experience point of view in simple words.

☒☐

The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain Yes there is scope to be more specific,

- ☐ ☒ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI
Explain:
- ☒ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)
Explain: Yes, least restrictive will be mostly choosen.
- ☒ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice
Explain: yes, most of the standards are still under development and might take sometime for standardization and adoption.
- ☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles
Explain:
- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

7.4.7 Respondent 7

Name: Juan Manuel Gonzalez Huesca

Organization:

0. General

How would you define Artificial Intelligence (AI) in one sentence?

It's the field which studies how machines (e.g. computers) can learn from data (as humans learn from experience) to perform human-like tasks.

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Data scientist

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

Manufacturing industry

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Supervised and unsupervised learning

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

Quality prediction in manufacturing processes

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒ ☐ An **important** problem

Explain: Nowadays, AI is widely used in our lives (e.g. fraud detection, loans approval, autonomous cars, supporting interview processes) that it has become very important to ensure impartiality and fairness in the outcome of its decisions, otherwise it can broaden current biases and discriminative attitudes impacting negatively many people.

☒ ☐ An **urgent** problem

Explain: Due to the fast paced adoption of AI across several industries, it's urgent to ensure mechanisms to avoid this to happen. This should be done to all systems that are already in the field and also before new ones are deployed.

☒ ☐ A **common** problem

Explain: Through recent years this topic has been raising awareness among AI practitioners and regulatory bodies which has helped by raising awareness in the community. Despite these efforts, AI is applied in so many sectors that it's very difficult to ensure this has been taken care of in all the cases and in a proper way.

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☒ ☐ An **important** problem

Explain: The use of AI is growing in the industry and reaching more sectors within the industry so eventually there should be formal mechanisms to ensure fairness in the AI systems.

☐ ☒ An **urgent** problem

Explain: Not at this point but it may change as most of the AI applications are focused on understanding manufacturing processes which impacts mainly operational procedures.

☐ ☒ A **common** problem

Explain: Due to the nature of the manufacturing industry where we mainly deal with production measurements, discrimination that may impact humans is not common but this may changes as new sectors within the industry are adopting the usage of AI.

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes No

☐ ☒ Healthcare
Explain: I expect the features and dataset to be well chosen in a way that a model doesn't learn biases.

☐ ☒ Wellness
Explain: As above, I expect the feature and dataset to be independent of biases and focused on the AI task.

☒ ☐ Policing/criminal law
Explain: Due to the nature of the data to be collected and analysed, mainly unsupervised, there is a higher chance to have biased results.

☒ ☐ Education
Explain: Due to the nature of the sector and the data to learn from, it has higher chance to be biased.

Yes No

☒ ☐ Housing
Explain: The nature of the data itself also would increase the chance to be biased.

☐ ☒ Retail
Explain: Data tends to be more structure plus retail campaigns target specific people groups, which would ensure by default the right features and datasets are used.

☒ ☐ Labourmarket/employment
Explain: It may be impacted also wit the data nature to have higher chance to be biased.

☐ ☐ Otherwise, namely
Explain:

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance

in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: A poor problem definition may lead to believe there is no bias problem</i>	<i>Problems: The features and the dataset itself doesn't represent all the groups equally</i>	<i>Problems: Those features who may induce biased are not properly handled</i>	<i>Problems: The right model is not used and modelling analysis is not in place</i>	<i>Problems: Model fairness is not evaluated</i>	<i>Problems: There is no track of fairness in production</i>
<i>Examples: The classification of pictures may not considered all relevant information to cover all groups of people equally</i>	<i>Examples: In a picture classification task, some groups may be under-represented</i>	<i>Examples: Features like gender or ethnicity should not be used in a CV assessment</i>	<i>Examples: Extra steps should be considered for under-represented groups during a classification task (e.g. class weight)</i>	<i>Examples: In a picture classification task, when only samples from the training group are used to evaluate the model</i>	<i>Examples: A picture classification task may become unfair with time if the training data is not appropriate</i>
<i>Causes: Not enough time is invested in this part</i>	<i>Causes: People is not aware or their main target market is limited</i>	<i>Causes: Unawareness of the situation</i>	<i>Causes: Unawareness of the situation or in how to handle them in AI</i>	<i>Causes: Unawareness of the situation or main target market is limited</i>	<i>Causes: Lack of plan to ensure fairness in production</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on

Problems: <i>The main problem I see is the unawareness of the existence of the problem, which is then propagated to all the following steps. Also, sometimes, there is limited time to bring a new AI product and fairness is overlooked.</i>	Problems: ...	Problems: ...	Problems: ...	Problems: ...	Problems: ...
Examples: <i>When a project has limited resources and time, fairness stays as a bonus which most of the time is never analysed</i>	Examples: ...	Examples: ...	Examples: ...	Examples: ...	Examples: ...
Causes: <i>Practitioners are unaware or doesn't understand the implications of fairness</i>	Causes: ...	Causes: ...	Causes: ...	Causes: ...	Causes: ...

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

Yes, greatly. By analyzing all the available data (mainly unstructured) which is generated and fed back into our society, we could see all the biased around us which will increase the awareness on them and enable us to work in mechanism to avoid them

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

There are no specific actions at this point in my current positions as we are mainly dealing with manufacturing tools and processes which is not impacted by fairness per se, but we deal with clustered data and we ensure a proper treatment. In the following sections, groups may be used to refer to different classes or data categories.

(1) <i>problem definition</i>	(2) <i>data collection</i>	(3) <i>data preparation</i>	(4) <i>modelling</i>	(5) <i>evaluation</i>	(6) <i>implementati on</i>
<i>Make sure all possible sources of unfairness are captured in the definition</i>	<i>All groups should be well represented</i>	<i>Mechanisms to handle different groups should be in place</i>	<i>Models are chosen to deal with potential groups differences</i>	<i>Every group is inspected and evaluated separately</i>	<i>Track and monitor performance for all groups</i>

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

No specific ones.

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Yes No

☐ ☒ Formal Risk Assessment
Explain:

☐ ☒ Checklists
Explain:

- ☐ ☒ Toolkits
Explain:
- ☐ ☒ Conversations with users
Explain:
- ☐ ☒ Conversations with other parties involved
Explain: Based on domain knowledge, we can know the existence of different groups in the data.
- ☒ ☐ Otherwise, namely:
Explain: data visualization and clustering

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Yes No

- ☐ ☒ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?
Explain:
- ☐ ☒ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?
Explain:
- ☐ ☒ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?
Explain:
- ☒ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain: data description mainly comes from domain experts in the manufacturing industry. They also confirm the existence when using visualization tools.
- Other people involved; If so, explain....
- ☐ ☐ Otherwise, namely...
Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

Yes No

- ☒ ☐ Collect more data
Explain: Try to balance the data with as many samples for all the groups as possible

- ☒ ☐ Analysis of selection variables
Explain: Variables are selected based on the AI task.
- ☒ ☐ Adjusting / filtering data (pre-processing)
Explain: Depending on the AI task, different pre-processing steps may be used to deal with groups difference (e.g. oversampling)
- ☐ ☒ Special techniques to reduce unfairness during learning (in-processing)
Explain:
- ☐ ☒ Special techniques to mitigate bias of predictions (post processing)
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

I am aware of general information from GDPR

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

There is no role as in my field this is not an issue

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

N. A.

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

Clear, simple and standard guidelines and procedures

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☒ ☐ The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: I am not fully aware of these documents and principles

☒ ☐ The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: Some of the documents are not easy to follow and implement

☒ ☐ The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain: They should target different audiences and for practitioners it should have specific examples across multiple fields

- ☒ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI
Explain: Landing the concepts into specific AI examples will greatly help to improve the understand and the acceptance of them
- ☒ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)
Explain: It could be beneficial for everybody to adhere to one framework instead of having several
- ☒ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice
Explain: This is true, but something expected when new guidelines and practices are introduced
- ☒ ☐ There is a lack of supervision and enforcement of the existing standards and principles
Explain: There should be a body which ensures this, even within organizations
- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

Be simple and pragmatic, whenever possible use real examples across multiple sectors and maybe create some templates to fill and ensure fairness across the development of an AI product.

7.4.8 Respondent 8

Naam:

Organisatie:

0. Algemeen

Hoe zou u zelf in 1 zin Artificiële Intelligence (AI) definiëren?

Systems that learn from the data and make decisions, eventually human in the loop.

Wat is uw functie/rol (b.v. data scientist, product owner, project manager, data-analist, data engineer, machine learning engineer, manager, onderzoeker, leidinggevende)?

Principal Data Scientist

In welke industrie bent u werkzaam (b.v. gezondheidszorg, financiën, energie, online retail, wetenschap)?

Banking

Binnen welke AI-domein(en) vinden uw werkzaamheden met name plaats (b.v. computer vision, natural language processing, recommender systems, (geautomatiseerde) besluitvorming)

Ik werk en coördineer het werk van 12 data science teams. De werkzaamheden van deze 12 teams bevatten alle bovengenoemde AI domeinen

Bij de ontwikkeling of het gebruik van welk soort AI-applicaties bent u betrokken (b.v. fraudedetectie, vraagvoorspelling, customer service)?

Fraudedetectie, Customer engagement, Detecting financial crime, crisis propagation in networks of companies, Anti-money Laundering, Speech analytics, etc.

1. Probleemanalyse

Bij het gebruik van AI kan (onbedoelde) discriminatie optreden.

Is het optreden van discriminatie door het gebruik van AI **in het algemeen** volgens u:

Ja Nee

Ja Een **belangrijk** probleem
☐ *Leg uit: iedereen heeft een maatschappelijk verantwoordelijkheid tegenover alle burgers met hun verschillende achtergronden. Wij willen graag de voordelen die aanwezig zijn in de maatschappij niet door AI laten bevestigen/herhalen. Wij willen juist door middel van AI deze vooroordelen fixen.*

☐ ☐ Een **urgent** probleem
Leg uit:

☐ ☐ Een **veelvoorkomend** probleem
Leg uit:

1.2. Is het optreden van discriminatie door het gebruik van AI in **uw werkdomein** volgens u:

Ja
☐ Een **belangrijk** probleem
Leg uit: Leg uit: Als bank, wij hebben een maatschappelijk verantwoordelijkheid tegenover alle burgers met hun verschillende achtergronden. Wij willen graag de voordelen die aanwezig zijn in de maatschappij niet door AI laten bevestigen/herhalen. Wij willen juist door middel van AI deze vooroordelen fixen.

☐ ☐ Een **urgent** probleem
Leg uit:

☐ ☐ Een **veelvoorkomend** probleem
Leg uit:

1.3 In welke AI toepassingen (sectoren/domeinen) verwacht u dat ongewenste discriminatie/ongelijke behandeling zich met name zal voordoen?

Ja

☐

Gezondheidszorg
Leg uit: Meeste ontwikkelingen in de Gezondheid zorg zijn geleid door een bepaalde type mensen (Witte mannen). Ben bang dat deze type mensen niet rekening zullen houden met andere groepen,

☐☐

Welzijn
Leg uit:

☐☐

Politie en justitie
Leg uit: Kleine criminaliteit (Diefstal, Fraude, etc) komt vooral uit achterstand wijken (immigranten, lage opgeleiden, etc). Ben bang dat AI zal deze groep voor altijd als criminelen labelen.

☐☐

Onderwijs
Leg uit: Kinderen in achterstand wijken zullen ook vanwege de data historie ook door AI gelabeld worden als niet goed lerende kinderen

Ja Nee

☐☐

Huisvesting
Leg uit:

☐☐

Retail
Leg uit:

☐☐

Arbeidsmarkt
Leg uit:

☐☐

Anders, namelijk...
Leg uit:

AI-systemen verlopen vaak via verschillende fasen/stappen zoals: (1) Probleemdefinitie (identificeren bedrijfsprobleem en vertaling naar ML taak), (2) Initiële dataverzameling en analyse, (3) Datavoorbereiding, (4) Modelleren (verschillende modellen uitproberen en een model selecteren), (5) Evaluatie (geselecteerde model evalueren aan de hand van succes criteria vastgesteld in stap 1 en een 'test set') en (6) Implementatie (inclusief monitoring/maintenance

activiteiten en terugkoppeling over performance in gebruik). [Als u een andere fasering voor ogen heeft, geef dat gerust aan]

1.4 Waar in het proces zijn volgens u de oorzaken van ongewenste discriminatie te situeren? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) Modelleren	(5) evaluatie	(6) implementati e
Problemen: Onbewust groep burgers in vakje willen classificeren	Problemen: AI trainen gebaseerd op de eerste beste beschikbaar data set. Trainen on the wrong sample. Ook “survivor bias” of the “power of the unseen genoemd”	Problemen: Niet goed omgaan met imbalanced data en outliers.	Problemen: Niet goed omgaan met imbalanced data en outliers.	Problemen: Model quality is not sufficient on all important data slices	Problemen: Automation of evaluation in production mode
Voorbeelden : Welke wijken zijn problemen wijken en welke mensen wonnen (achtergrond, opleiding, alleenstaande ouders, ect.) in deze wijken?	Voorbeelden : Abraham Wald remark during the WWII. The damaged portions of returning planes show locations where they can sustain damage and still return home; those hit in other places do not survive. Missing the data of those	Voorbeelden : Analisten/Data scientists gebruiken de data zonder toepassing van technieken die die de imbalance (brom van bias/discriminatie) in de data kan gedeeltelijk fixen. Of nog erger Analisten/Data scientists doe de outliers	Voorbeelden : Analisten/Data scientists gebruiken de data zonder toepassing van technieken die die de imbalance (brom van bias/discriminatie) in de data kan gedeeltelijk fixen. Of nog erger Analisten/Data scientists doe de outliers	Voorbeelden : Global accuracy of a model drop by 25% for one group in the sample.	Voorbeelden :

	who didn't survive	"gewoon" uit de samples laten verdwijnen	"gewoon" uit de samples laten verdwijnen		
Oorzaken: Niet bewust zijn van een discrimineerde probleemstelling	Oorzaken: Missing diep kennis van statistiek	Oorzaken: Missing diep kennis van statistiek. Missing duidelijk guidelines	Oorzaken: Missing diep kennis van statistiek. Missing duidelijk guidelines	Oorzaken: Missing de juiste modelling technieken	Oorzaken:

1.5 Transparantie van AI-besluiten is een *best practice*; ook hebben burgers het recht om geautomatiseerde besluiten die hen raken te begrijpen. Welke obstakels zijn er ten aanzien van transparantie van AI-processen en daarop gebaseerde beleids- en besluitvorming? Geef aub zo specifiek mogelijk aan waar in het proces deze problemen te situeren zijn (u hoeft uiteraard niet overal iets in te vullen).

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e
Problemen:	Problemen:	Problemen:	Problemen:	Problemen:	Problemen:
Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :	Voorbeelden :
Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:	Oorzaken:

1.6 Zou AI juist ook kunnen helpen bij het verminderen/verhelpen van bestaande (indirecte) discriminatie, bias, unfairness of een gebrek aan transparantie in huidige, niet-geautomatiseerde beleid- en besluitvormingsprocessen? Zo ja, hoe?

2. Mogelijke aanpakken

Verschillende organisaties en bedrijven hebben principes en standaarden ontwikkeld om discriminatie door AI systemen te minimaliseren. Data scientists hebben het hierbij vaak over het reduceren van bias en het zorgdragen voor *fairness*.

2.1 Neemt u of uw organisatie zorgen omtrent discriminatie momenteel mee in keuzes en beslissingen aangaande het gebruik of de inrichting van AI? Zo ja, ten aanzien van welke fase(s)? Leg aub uit.

(1) probleem definitie	(2) dataverzame ling	(3) datavoorber eiding	(4) modelleren	(5) evaluatie	(6) implementati e
Wij gebruiken ideeën vanuit de paper” NEAR TERM AI Hanna Gunn and Cathy O'Neil”	Wij gebruiken ideeën vanuit de paper” NEAR TERM AI Hanna Gunn and Cathy O'Neil”	Wij gebruiken ideeën vanuit de paper” NEAR TERM AI Hanna Gunn and Cathy O'Neil”	Wij gebruiken ideeën vanuit de paper” NEAR TERM AI Hanna Gunn and Cathy O'Neil”	Wij gebruiken ideeën vanuit de paper” NEAR TERM AI Hanna Gunn and Cathy O'Neil”	Wij gebruiken ideeën vanuit de paper” NEAR TERM AI Hanna Gunn and Cathy O'Neil”

2.2 Zo ja, maakt u of uw organisatie gebruik van specifieke standaarden/principes/richtlijnen (voeg aub bron toe)?

No Yet

2.3 Welke activiteiten/technieken gebruikt u of uw organisatie om mogelijke pijnpunten in het kader van ongelijke behandeling/discriminatie **te identificeren**?

Ja Nee

- ☐ ☐ Formele risicoanalyse
Leg uit: The Ethical Matrix from NEAR TERM AI Hanna Gunn and Cathy O'Neil”
- ☐ ☐ Checklists
Leg uit:
- ☐ ☐ Toolkits
Leg uit:

- ☐ ☐ In gesprek met gebruikers
Leg uit:
- ☐ ☐ In gesprek met andere betrokkenen
Leg uit:
- ☐ ☐ Anders, namelijk...
Leg uit:

2.4 Welke activiteiten/technieken gebruikt u of uw organisatie om discriminatie/ongelijke behandeling **te meten**?

Ja Nee

- ☐ ☐ **Data op biases** checken aan de hand van fairness criteria. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:
- ☐ ☐ Model evalueren aan de hand van formele fairness criteria bij **model selectie**. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:
- ☐ ☐ Model evalueren aan de hand van formele fairness criteria bij **model in productie**. Zo ja, welke criteria gebruikt u hiervoor?
Leg uit:
- ☐ ☐ Actieve evaluatie ten behoeve van discriminatie/ongelijke behandeling met Gebruikers; Zo ja, *leg uit*....
Domeinexpert; Zo ja, *leg uit*....
Andere betrokkenen; Zo ja, *leg uit*
- ☐ ☐ Anders, namelijk...
Leg uit:

2.5 Welke activiteiten/technieken worden in uw team gebruikt om discriminatie/ongelijke behandeling **te reduceren**?

Ja Nee

- ☐ ☐ Meer data verzamelen
Leg uit:
- ☐ ☐ Analyse van selectie variabelen
Leg uit:
- ☐ ☐ Data aanpassen/filteren (pre-processing)
Leg uit:

- ☐ ☐ Speciale technieken om unfairness te reduceren tijdens het leren (in-processing)
Leg uit:
- ☐ ☐ Speciale technieken om bias te mitigeren van predictions (post processing)
Leg uit:
- ☐ ☐ Anders, namelijk ...
Leg uit:

3. Juridisch raamwerk

3.1 Bent u op de hoogte van bestaande wetgeving, richtsnoeren of Codes of Conducts ten aanzien van AI? Zo ja, welke en welke spelen specifiek een rol bij de ontwikkeling of het gebruik van AI systemen binnen uw organisatie als het gaat om discriminatie?

3.2 Op welke moment komt deze wetgeving aan bod in de ontwikkeling of het gebruik van AI systemen binnen uw organisatie en op welke manier?

3.3 Zijn er in het verleden knelpunten geweest bij het operationaliseren van de wetgeving in de ontwikkeling, implementatie en gebruik van AI systemen binnen uw organisatie en hoe heeft uw organisatie (of is zij voornemens) deze knelpunten te adresseren?

3.4 Wat zou u idealiter van een wet of code of conduct verwachten ten aanzien van het bestrijden van discriminatie en ongelijke behandeling in AI-systemen?

4. Handreiking Non-Discrimination by Design voor AI

Wij zijn gevraagd door het Ministerie van Binnenlandse Zaken om een handreiking te maken voor het ontwikkelen en gebruik van AI systemen, in het bijzonder met het oog op het voorkomen en tegengaan van discriminatie en ongelijke behandeling. Deze handreiking zal met name gericht zijn op systeembouwers en data-analisten die AI-systemen bouwen en ontwikkelen en op organisaties die deze systemen gebruiken en toepassen in de praktijk. Kernvraag daarbij is: Hoe bouw je een AI-systeem en pas je die in de praktijk toe op een wijze dat gevaren voor discriminatie en ongelijke behandeling worden geminimaliseerd?

Uiteraard bestaat er reeds een keur aan standaarden, codes of conduct en beleidsdocumenten ten aanzien van AI, is de uitgebreide jurisprudentie met betrekking tot discriminatie en gelijke behandeling onverkort van toepassing op besluit- en beleidsvorming door middel van AI en zijn er in diverse academische studies additionele standaarden voor AI systemen uitgewerkt.

De assumptie die ten grondslag ligt aan de ontwikkeling van deze handreiking is dat de bestaande juridische en ethische documenten, principes en standaarden nog niet of slechts in beperkte mate worden gevolgd door data-analisten en systeemontwikkelaars die AI-systemen bouwen en organisaties die AI systemen gebruiken voor beleids- en besluitvorming.

Deelt u deze assumptie? Zo ja, wat is denkt u de oorzaak? (meerdere antwoorden mogelijk)

Ja Nee

- ☐ ☐ De bestaande documenten (richtsnoeren, jurisprudentie, literatuur, etc.) en de daarin vervatte principes zijn onvoldoende bekend bij mensen/organisaties die AI-systemen ontwikkelen/gebruiken

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn geschreven in juridisch/beleidsjargon dat data-analisten, systeemontwikkelaars, mensen/organisaties die AI in de praktijk gebruiken weinig zegt

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn te algemeen; voor mensen/organisaties die AI-systemen ontwikkelen/gebruiken zijn principes zoals ‘wees transparant’, ‘vermijd het maken van ongerechtvaardigd onderscheid tussen groepen’ en ‘vermijd bias in de dataset’ te algemeen – de principes zouden veel specifiekere moeten worden uitgewerkt om bruikbaar te zijn in de praktijk

Leg uit:

- ☐ ☐ De bestaande documenten en de daarin vervatte principes zijn op een universeel, abstract en context onafhankelijk niveau geformuleerd, terwijl de betekenis,

interpretatie en implementatie per context en per toepassingsgebied anders is
Leg uit:

☐ ☐ Er zijn simpelweg te veel richtsnoeren en standaarden; data-analysten, systeemontwikkelaars en mensen/organisaties die AI systemen in de praktijk toepassen weten niet aan welke standaard zij zich moeten houden (of kiezen de voor hen meest gunstige)
Leg uit:

☐ ☐ De standaarden zijn pas recentelijk ontwikkeld; het heeft tijd nodig voordat deze in de praktijk worden opgepakt
Leg uit:

☐ ☐ Het ontbreekt aan toezicht op en handhaving van de bestaande standaarden en principes
Leg uit:

☐ ☐ Anders, namelijk:

Heeft u nog overige opmerkingen, suggesties of tips voor ons bij de ontwikkeling van een handreiking *Non-Discrimination by design voor AI Systemen*?

7.4.9 Respondent 9

Name:

Organization:

0. General

How would you define Artificial Intelligence (AI) in one sentence?

A piece of software whose inner workings (rules) are learned from example data.

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Data Scientist & ML Engineer

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

Consulting, Finance

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Computer vision, forecasting, decision-making, recommender systems

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

Customer lifecycle, forecasting, risk

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒ ☐ An **important** problem
Explain: the effects of discrimination have a big impact on individual lives and on society

☒ ☐ An **urgent** problem
Explain: application of AI is growing exponentially in all aspects of industry and government

☒ ☐ A **common** problem
Explain: there are many examples of AI gone wrong in the recent past, even from the biggest tech companies

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☒ ☐ An **important** problem
Explain: financial decision are among the most impactful decisions that can be made by AI systems

☒ ☐ An **urgent** problem
Explain: there is a lot of pressure on financial institutions from regulatory and economic perspective, which means financial institution are increasingly forced to grab at supposedly efficient solutions

☐ ☐ A **common** problem
Explain: cannot say at this point how common or uncommon discrimination in the financial domain by AI systems is; I do not have enough insight/data yet

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes No
☐ ☐ Healthcare
Explain:

Yes No
☐ ☐ Housing
Explain:

☐ ☐ Wellness
Explain:

☐ ☐ Retail
Explain:

☒ ☐ Policing/criminal law
Explain: worldwide, police forces suffer from lack of resources, so automation is tempting. However, in policing there are more potential and complex biases and systematic effects than in many other domains.

☒ ☐ Labourmarket/employment
Explain: tasks in HR are labour intensive, and often subjective, while HR is a 'soft' profession. This makes it a prime target for selling pseudo-scientific AI system that pretend that they can get the 'truth' about people.

☐ ☐ Education
Explain:

☒ ☐ Otherwise, namely Finance
Explain: (see previous question)

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

Everywhere! Sorry, it's too much to write about. The causes are everywhere, and also roughly equally distributed between the phases (although that can depend on the specific case).

<i>(1) problem definition</i>	<i>(2) data collection</i>	<i>(3) data preparation</i>	<i>(4) modelling</i>	<i>(5) evaluation</i>	<i>(6) implementati on</i>
<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...

<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems:</i> ...	<i>Problems:</i> unsuitable data collection can compromise the problem deifinition	<i>Problems:</i> ...	<i>Problems:</i> black box models	<i>Problems:</i> the chosen optimization target can matter a lot	<i>Problems:</i> how are the results actually used?
<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

No, not by itself. AI just propagates all biases that already exists in the data and the system. You could create a system or a project that reduces bias and has an AI component, but it will not be the AI component that will do the heavy lifting of the bias reduction; that would be done by redesign of the data gathering process, data preparation process, optimization targets (plural!) and governance.

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain. **Yes, but I do not have the details or the extent of it.**

(1) <i>problem definition</i>	(2) <i>data collection</i>	(3) <i>data preparation</i>	(4) <i>modelling</i>	(5) <i>evaluation</i>	(6) <i>implementati on</i>

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

We are currently working on a concerted effort to determine tools and processes for all phases as part of organization wide protocol.

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Centralization effort in progress, I don't know what individual teams do.

Yes No

☒ ☐ Formal Risk Assessment
Explain: as a financial institution, there are existing regulatory processes in this regard

☐ ☐ Checklists
Explain:

☐ ☐ Toolkits
Explain:

☐ ☐ Conversations with users
Explain:

☐ ☐ Conversations with other parties involved

Explain:

☐ ☐ Otherwise, namely...

Explain:

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Centralization effort in progress, I don't know what individual teams do.

Yes No

☐ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?

Explain:

☐ ☐ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?

Explain:

☐ ☐ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?

Explain:

☐ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain....
- Other people involved; If so, explain....

☐ ☐ Otherwise, namely...

Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

My current usecase has no risk factors.

Yes No

☐ ☐ Collect more data

Explain:

☐ ☐ Analysis of selection variables

Explain:

☐ ☐ Adjusting / filtering data (pre-processing)

Explain:

- ☐ ☐ Special techniques to reduce unfairness during learning (in-processing)
Explain:
- ☐ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

I am aware that over 100 exist, and have saved many of them. We have not yet designated a single 'official' guideline.

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

Before exploration, and then again after exploration but before production, there are risk and privacy assessments.

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

What do you mean 'operationalization of legislation'? As a financial institution, there is existing legislation that is being followed, but this is not AI-specific. I have not yet noticed this causing a bottleneck.

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

This is a very big question, that I cannot hope to answer exhaustively. The following are only some speculations in the moment, and I may change my mind in the future when I think about it more:

Requirements on

- governance structures,*
 - transparency towards users/clients,*
 - monitoring (incl bias metrics),*
 - external audits,*
 - accessible redress channels,*
 - operational integrity (tracability, versioning, lineage, logging, etc)*
-

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☒☐

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: yes, this is the case, but not the problem. Of course people know *of* the documents, but there are too many of them, how to choose? People do not and will not memorize the contents. The problem is *applying* them.

☐☐

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: don't know, depends on the document

☒☐

The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain: yes, but see below. I am not sure to what extent it is possible or desirable due to the complexity, to try to write elaborately detailed documents.

- ☒ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain: very true, everyone is talking about it and thinking about it and writing about it, but hardly anyone is actually doing it. The principles are rather clear already in a general sense. In order to make further progress, practical application to real cases is needed, so that the learnings can be documented and the existing documents and principles be enhanced with practical examples, tools, suggestions, etc

- ☒ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain: this is true. Additional documents will not solve this. An organization could just task a few people to do a little research and pick one to make official. Or the government could pick one.

- ☒ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain:

- ☒ ☐ There is a lack of supervision and enforcement of the existing standards and principles

Explain: for sure this would give more incentive (see below). Assuming enforcement from the government is meant. Supervision/enforcement from the organization is also rare to non-existent, and they don't have an incentive for it either.

- ☒ ☐ Otherwise, namely the most important one! There is no incentive, no support and no commitment to apply the principles from middle-management/stakeholders. Developers are not the ones you have to convince to follow the principles; they definitely want to. But they are under too much pressure to deliver (products, features, KPI's, etc) and simply do not have the time to even think about applying them.

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems* (which will also appear in English)?

The inclusion of Responsible Data Science principles in the design, development and maintenance of AI systems cannot be seen as a separate component that can simply be put in place next to or on top of whatever it is that organizations are doing at the moment. The only way to make it work is if the organizational processes are designed to integrate with the principles in the first place. Specifically, without proper Machine Learning Operations (MLOps) practices and infrastructure and without appropriate Way of Working, it is nearly hopeless to try to implement responsible data science principles robustly.

I strongly support that the government is trying to get involved in guidelines and regulations for ethical AI. I also wish you luck in your work, because it is unclear to me what the positioning and scope of your guide should be to make it more useful than the plethora of works that are already out there.

7.4.10 Respondent 10

Name: Violeta Misheva

Organization: ABN Amro bank

0. General

How would you define Artificial Intelligence (AI) in one sentence?

AI is the design and development of algorithms to solve problems which were before solved only by machines.

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Data scientist

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

Finance(banking)

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Mixture of those, plus more classical ML: forecasting, classification, regression

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

Also varies: from fraud detection, to internal processes and operations..

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒☐

An

important

problem

Explain: It is definitely important problem and should be prevented as much as possible.

☒☐

An

urgent

problem

Explain: It is definitely an urgent problem that needs to be addressed. Currently, many organizations are not spending enough efforts on this (as much as we are aware) and they train and use algorithms which can result in perpetuating existing societal biases.

☐☐

A

common

problem

Explain: Not clear since we only know about certain cases where very clear and obvious discrimination occurs, less so about the average use cases that applies ML or AI. However, given that algorithms use historical data to train, and data is the result of human decision-making, and we humans are inherently biased, perhaps it is a much more common problem than we think..

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☒☐

An

important

problem

Explain: Finance and banking is a domain which can have a large impact on an individual's life; therefore it's very important we ensure algorithms we train in the financial domain are as fair as possible.

☒☐

An

urgent

problem

Explain: I do believe it's an urgent problem. We do not want to wait for an obvious bias to be revealed before we start thinking about unbiased and discrimination-free algorithms which decide who gets a mortgage and who not, for example.

☐☐

A

common

problem

Explain: Not clear for me how common it is since not enough investigation has been done across the field (that I am aware of.)

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes ☒ No ☐ Healthcare
Explain:

Yes ☒ No ☐ Housing
Explain:

☐ ☐ Wellness
Explain:

☐ ☐ Retail
Explain:

☒ ☐ Policing/criminal law
Explain:

☒ ☐ Labourmarket/employment
Explain:

☐ ☐ Education
Explain:

☒ ☐ Otherwise, namely :
 finance
Explain:

I expect issues of unwanted discrimination in these fields first of all because these fields are notorious in displaying a historical bias towards a certain group of the population to a different degree across countries and contexts. I also think such fields could have a large impact on important decisions regarding an individual's life and would thus consider them more important to start with.

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

Note: I am not certain what the 'Causes' box refers to. Is it related to what causes the problem, or what consequences does that problem have (especially if not addressed)??

(1) problem definition	(2) <i>data</i> collection	(3) <i>data</i> preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: historical data could contain</i>	<i>Problems: when creating feature, many biases</i>	<i>Problems: ...</i>	<i>Problems: Not choosing an evaluation metric which</i>	<i>Problems: ...</i>

	<i>many types of biases</i>	<i>could be introduced as well</i>		<i>can ensure certain chosen fairness</i>	
<i>Examples: ...</i>	<i>Examples: models to detect fraudulent activity can be biased/focused on certain subgroups of the population</i>	<i>Examples: creating features that include sensitive attributes of their proxies (related to religion, ethnic background, age, sex, etc.)</i>	<i>Examples: ...</i>	<i>Examples: optimising for accuracy when in fact we may care about precision in certain subgroup of the population, or ensure equal parity, etc.</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: conscious and unconscious biases in human decision-making</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementation
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: The development of non-transparent features in the model</i>	<i>Problems: development of 'black box' algorithms</i>	<i>Problems: classical transparent methods tend to perform slightly worse than more black box methods</i>	<i>Problems: Big potential problem here is not have clear monitoring checks in place for any explainability approaches in place</i>

<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: When working with NLP data, and employing some complex embedding approach</i>	<i>Examples: deep learning but also more classical ML (ensemble of models)</i>	<i>Examples a random forest will outperform a decision tree in most cases</i>	<i>Examples: It's easy to validate the performance of a model once it's deployed, but how does one validate the explanations, if those are also put in production?</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: As a result, the modeller may choose for more complex approaches just to gain some boost in performance</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

Algorithms are not biased, humans are. Since we use historical data to build our algorithms, they will directly inherit the biases present in the historical data. Analysing the decisions of an algorithm for the presence of bias can help illuminate historical biases present in the data and /or bias in the decision-making of the algorithm.

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

We split our process in 3 stages: exploration, development and deployment. We have concerns with regards to the first 2 stages, i.e. everything except the deployment/implementation.

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

We are currently developing them. We have developed a process for the development and deployment of algorithms and are currently in the process of embedding approaches for fairness and explainability.

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Yes No

☒☐

Formal Risk Assessment

Explain: every model goes through a change risk assessment

☒☐

Checklists

Explain: There is a privacy checklist, even for models which remain in exploration.

☒☐

Toolkits

Explain: We are currently testing different fairness metrics (individual vs group fairness) and mitigation strategies. For that, we will use developed packages and open-source libraries/toolkits, not develop our own (but is a work in progress).

- ☒ ☐ Conversations with users
Explain: as a first step to identify potential problematic outcomes of a model, we are testing the application of the ethical matrix by Cathy O'Neil and others.
- ☐ ☐ Conversations with other parties involved
Explain:
- ☒ ☐ Otherwise, namely...
Explain: We also have a few other procedures in place, such as a DPIA (data protection impact assessment).

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Yes No

- ☒ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?
Explain: We do not have a selected criteria yet (because they will also depend on the context) but rather a list of available metric/criteria to test for.
- ☐ ☐ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?
Explain: Same answer as question above.
- ☒ ☐ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?
Explain:
- ☒ ☐ Active evaluation for discrimination / unequal treatment with
 - Users: This is the first step in our process, with the previously mentioned ethical matrix
 - Domain expert: parties representing legal, compliance, privacy, as well as colleagues who represent various user groups are also part of this process.
 - Other people involved; If so, explain....
- ☐ ☐ Otherwise, namely...
Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

Yes No

- ☐ ☐ Collect more data



Explain:

☐ ☐ Analysis of selection variables

Explain:

☐ ☐ Adjusting / filtering data (pre-processing)

Explain:

☒ ☐ Special techniques to reduce unfairness during learning (in-processing)
Explain: Not one approach yet, but we are experimenting with various approaches (such as re-weighting for certain groups)

☒ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:

☐ ☐ Otherwise, namely...

Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

I am aware of the GDPR and the way it impact what data we can use in the model, how we can access it and how long/where we can store it. Depending on the domain, there are other requirements for models in terms of transparency (especially models in risk) but I don't know to what extent this is embedded in law or not.

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

As far as I am aware, in the initial stages: saying something about what data we should use, what features we can include in the data, what proxies; and eventually, what type of a model we may need to train (inherently transparent one vs not)

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

I am not aware and that familiar with this.

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

Again, my knowledge in the legal domain is superficial, but I'd expect some guidelines with regards to discrimination and detecting and mitigating bias. At the same time, I hope any code of law would be flexible enough not to make the model development process super stringent.

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

- ☐ ☐ The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain: I think it depends on the organization and the employee themselves. It's be a job of the organization/institution to educate their employees on the existing regulations by providing short workshops and trainings.

- ☐ ☐ The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain: Same answer as above.

- ☒ ☐ The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain: I think the existing guidelines are too general indeed and should be more concrete.

☐ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI
Explain: I agree overall but then again I think that the differences are not as wide as one might expect. I suspect to see discrimination happening along similar lines in different contexts.

☐ ☒ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)
Explain: It's up to the organizations/companies to simplify these rules and regulations for their employees. This will in fact be necessary to ensure a unified process across the organization.

☒ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice
Explain:

☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles
Explain: I am not aware of that..

☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

7.4.11 Respondent 11

Name:

Organization:

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Algorithms inspired by human cognition (also called agents) that could perform as good as the brain when it comes to tasks such as computation, learning, etc.

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Data team lead

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

Digital marketing

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

NLP and recommender system

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

Demand prediction

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒ ☐ An **important** problem
Explain:

☐ ☒ An **urgent** problem
Explain: it is not urgent yet because all sectors have not adopted the use of AI yet. But in the future, it could become urgent.

☒ ☐ A **common** problem
Explain:

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☐ ☒ An **important** problem
Explain:

☐ ☒ An **urgent** problem
Explain:

☒ ☐ A **common** problem
Explain: Depending on the type of input that we feed into the algorithm, you could get biased recommendation. It is a common issue in every domain. It is not urgent or important for us because it does not impact anyone's life or jeopardize our business in any form.

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes No

☒ ☐ Healthcare
Explain: It could happen when algorithms wrongly suggest specific treatment without considering individual differences.

☐ ☒ Wellness
Explain:

☒ ☐ Policing/criminal law
Explain: when the training data is biased towards specific minorities which due to social and political reasons in their time, they lack proper integration or education and accordingly statistically committed more crimes comparing to other ethnicities. Social and political situations could easily bias the dataset and as a result bias the algorithm recommendation.

☒ ☐ Education
Explain: When algorithm recommends something in individual level. Every student learns differently,

Yes No

☐ ☒ Housing
Explain:

☐ ☒ Retail
Explain:

☒ ☐ Labourmarket/employment
Explain: Depending on the training dataset, automatic hiring process could be heavily biased.

☐ ☐ Otherwise, namely
Explain:

and it could be risky
to treat a specific
student because
algorithm trained
based on a pool of
other students
recommend so.

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

<u>(1) problem definition</u>	<u>(2) data collection</u>	<u>(3) data preparation</u>	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...
<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...
<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...

<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

Maybe due to unintentional biases toward specific race, some people historically get treated worse than others. An AI with balanced dataset could help with reducing this sort of discriminations.

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

(1) problem definition	(2) <u>data</u> <u>collection</u>	(3) <i>data</i> preparation	(4) <u>modelling</u>	(5) <u>evaluation</u>	(6) implementati on

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

We have training cycles meaning as we collect more data over time, we re-train the existing models and correct some of the conclusions if needed. We also always discuss the model selection procedure carefully.

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

Yes No

- ☐ ☐ Formal Risk Assessment
Explain:
- ☒ ☐ Checklists
Explain:
- ☐ ☐ Toolkits
Explain:
- ☒ ☐ Conversations with users
Explain: I recommend it strongly
- ☒ ☐ Conversations with other parties involved

Explain: this really helps to define the problem better

☐ ☐ Otherwise, namely...

Explain:

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

Yes No

☒ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?

Explain: full statistical analyses to make sure that data represents the real population correctly.

☒ ☐ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?

Explain:

☐ ☐ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?

Explain:

☒ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain....
- Other people involved; If so, explain....

☐ ☐ Otherwise, namely...

Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

Yes No

☒ ☐ Collect more data

Explain: Always

☒ ☐ Analysis of selection variables

Explain:

☒ ☐ Adjusting / filtering data (pre-processing)

Explain:

☐ ☐ Special techniques to reduce unfairness during learning (in-processing)

Explain:

- ☒ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:
- ☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

No

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

Regarding GDPR, we used legal advises to make sure that we treat our data with enough care.

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

-

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

Questioning the process and offering a clear guideline that checks if the AI process (including the 6 stages described above (problem definition, etc)) is implemented correctly. Detailed analyses/checks should be done before any AI system gets used in production.

Similar concept as audit. i.e. before an AI system is in use in scale, a checklist should be followed, and the results of the analysis requested by the checklist should be stored in case of future investigations by the authorities.

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

I think there should be only one guide and it should be enforced by the authorities.

Yes No

☒ ☐ The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems
Explain:

☒ ☐ The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice
Explain:

☒ ☐ The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice
Explain:

- ☒ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain:

- ☒ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain:

.....

- ☒ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain:

.....

- ☒ ☐ There is a lack of supervision and enforcement of the existing standards and principles

Explain:

.....

- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

The guideline should be understandable not only for data analyst but also for top level management.

7.4.12 Respondent 12

Name: Carmen Bratosin
Organization: ESI (TNO)

0. General

How would you define Artificial Intelligence (AI) in one sentence?

Ability of computers to perform tasks in an automatic manner, being able to react and adapt to stimuli

0.2 What is your function / role (e.g. data scientist, product owner, project manager, data analyst, data engineer, machine learning engineer, manager, researcher, supervisor)?

Researcher

0.3 In which industry do you work (e.g. healthcare, finance, energy, online retail, science)?

science

0.4 Within which AI domain(s) do your activities mainly take place (e.g. computer vision, natural language processing, recommender systems, (automated) decision-making)

Nlp, process mining

0.5 In the development or use of what kind of AI application(s) are you involved (e.g. fraud detection, demand prediction, customer service)?

high tech systems diagnostics

1. Problem analysis

When using AI, (unintentional) discrimination can occur.

In your opinion, is the occurrence of discrimination through the use of AI **in general**:

Yes No

☒ ☐ An **important** problem
Explain: the use of AI was promoted to reduce bias however when bias is in the training set, bias will be transferred to the AI

☐ ☐ An **urgent** problem
Explain:

☐ ☐ A **common** problem
Explain:

1.2. The occurrence of discrimination through the use of AI in **your work domain** is:

Yes No

☐ ☐ An **important** problem
Explain:

☐ ☐ An **urgent** problem
Explain:

☐ ☐ A **common** problem
Explain:

Not relevant since we are mostly work with machine data

1.3 With regard to which AI-applications (sectors / domains) do you expect unwanted discrimination / unequal treatment to occur in particular?

Yes ☐ No ☒ Healthcare
Explain:

Yes ☐ No ☒ Housing
Explain:

☐ ☒ Wellness
Explain:

☐ ☒ Retail
Explain:

☒ ☐ Policing/criminal law
Explain:

☒ ☐ Labourmarket/employment
Explain: as explained above if you train on applications with bias towards certain group this will be translated into the learned behavior

☐ ☒ Education
Explain:

☐ ☐ Otherwise, namely
Explain:

AI systems often run through different stages / steps such as: (1) Problem definition (identifying business problem and translation to ML task), (2) Initial data collection and analysis, (3) Data preparation, (4) Modelling (testing and selecting different models), (5) Evaluation (evaluating selected model based on success criteria established in step 1 and a 'test set') and (6) Implementation (including monitoring / maintenance activities and feedback on performance in use). [If you envisage a different phasing, please indicate so]

1.4 Where in the process do you think the causes of unwanted discrimination can be situated? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) data collection	(3) data preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: data labelling</i>	<i>Problems: ...</i>	<i>Problems: ...</i>	<i>Problems: ...</i>
<i>Examples: ...</i>	<i>Examples: When collecting data where certain sex is preferred for specifc jobs</i>	<i>Examples: humans labelling the data they will introduce their bias</i>	<i>Examples: ...</i>	<i>Examples: ...</i>	<i>Examples: ...</i>
<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>	<i>Causes: ...</i>

1.5 Transparency of AI decisions is a best practice; citizens also have the right to understand automated decisions that affect them. What obstacles are there to transparency of AI processes and policy- and decision-making based on them? Please indicate as specifically as possible where in the process these problems can be situated (of course, you do not have to write down something in every box).

(1) problem definition	(2) <i>data</i> collection	(3) <i>data</i> preparation	(4) modelling	(5) evaluation	(6) implementati on
<i>Problems:</i> ...	<i>Problems:</i> <i>bias in the</i> <i>data</i>	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...	<i>Problems:</i> ...
<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...	<i>Examples:</i> ...
<i>Causes:</i> ...	<i>Causes:</i> ...	<i>Causes:</i> ...	<i>Causes:</i> ...	<i>Causes:</i> ...	<i>Causes:</i> ...

1.6 Could AI also help reduce/remedy existing (indirect) discrimination, bias, unfairness or a lack of transparency in current, non-automated policy and decision-making processes? If so, how?

I believe it is possible but that means that in the data preprocessing bias should be identified and corrected

2. Potential approaches

Several organizations and companies have developed principles and standards to minimize discrimination by AI systems. Data scientists often talk about reducing bias and ensuring fairness.

2.1. Are you or your organization currently taking into account concerns over discrimination in choices and decisions regarding the use or design of AI? If so, with regard to which phase(s)? Please explain.

<i>(1) problem definition</i>	<i>(2) data collection</i>	<i>(3) data preparation</i>	<i>(4) modelling</i>	<i>(5) evaluation</i>	<i>(6) implementati on</i>

I am not directly involved in activities where discrimination can play a role

2.2. If so, do you or your organization use specific standards / principles / guidelines (please add source)?

2.3 What activities / techniques do you or your organization deploy to **identify** potential issues of unequal treatment / discrimination?

I am not aware

Yes No

☐ ☐ Formal Risk Assessment
Explain:

☐ ☐ Checklists
Explain:

☐ ☐ Toolkits
Explain:

☐ ☐ Conversations with users
Explain:

☐ ☐ Conversations with other parties involved

Explain:

☐ ☐ Otherwise, namely...

Explain:

2.4 What activities / techniques do you or your organization deploy to **measure** discrimination / unequal treatment?

I am not aware

Yes No

☐ ☐ Checking for **biases in the data** based on fairness criteria. If so, which criteria do you use?

Explain:

☐ ☐ Evaluating the model based on formal fairness criteria for **model selection**. If so, which criteria do you use?

Explain:

☐ ☐ Evaluating the model based on formal fairness criteria for the **model in production**. If so, which criteria do you use?

Explain:

☐ ☐ Active evaluation for discrimination / unequal treatment with
- Users; If so, explain....
- Domain expert; If so, explain....
- Other people involved; If so, explain....

☐ ☐ Otherwise, namely...

Explain:

2.5 What activities / techniques are used in your team to **reduce** discrimination / unequal treatment?

I am not aware

Yes No

☐ ☐ Collect more data

Explain:

☐ ☐ Analysis of selection variables

Explain:

☐ ☐ Adjusting / filtering data (pre-processing)

Explain:

☐ ☐ Special techniques to reduce unfairness during learning (in-processing)

Explain:

☐ ☐ Special techniques to mitigate bias of predictions (post processing)
Explain:

☐ ☐ Otherwise, namely...
Explain:

3. Legal Framework

3.1 Are you aware of existing legislation, guidelines or Codes of Conducts regarding AI? If so, which one(s)? Which one(s) play a role in the development or use of AI systems within your organization when it comes to discrimination?

No

3.2 At which point in the process does the legal framework play a role in the development or use of AI systems within your organization and in what way?

I am not aware

3.3 Have there been any bottlenecks in the operationalization of legislation with regard to the development, implementation and use of AI systems within your organization in the past and how does your organization (or does it intend to) address these bottlenecks?

I am not aware

3.4 What would you ideally expect from a law or code of conduct with regard to preventing/mitigating/combating discrimination and unequal treatment through AI systems?

I am not aware

4. Guide on Non-Discrimination by Design for AI-systems

We have been asked by the Dutch Ministry of the Interior to develop a Guide for the development and use of AI systems, particularly with a view to preventing and combating discrimination and unequal treatment. This guide will be aimed primarily at system builders and data analysts who design and develop AI systems and at organizations that use and apply these systems in practice. The main question is: How do you build an AI system and apply it in practice in such a way that risks of discrimination and unequal treatment are minimized?

Naturally, there already is an abundance of standards, codes of conduct and policy documents with regard to AI, the extensive case law on discrimination and equal treatment applies in full to AI-based policy- and decision-making and there is a wide variety of academic literature in which additional standards for AI systems have been developed.

The assumption for developing a new guide is that existing legal, ethical and social principles are not, or only to a limited extent, followed by data analysts and system builders that develop AI-systems and organizations that use these systems for policy and decision-making.

Do you agree with that assumption? If so, what are the causes? (multiple answers possible)

Yes No

☐☐

The existing documents (guidelines, case law, literature, etc.) and the principles contained therein are insufficiently known to people / organizations that develop / use AI systems

Explain:

☐☐

The existing documents and the principles contained therein are written in legal / policy jargon that mean little to data analysts, system developers and people / organizations using AI in practice

Explain:

☐☐

The existing documents and the principles contained therein are too general; for people / organizations developing / using AI systems, principles such as "be transparent", "avoid making unjustified distinctions between groups", "avoid bias in the dataset" are too general - they should be elaborated in much more detail to be useful in practice

Explain:

- ☐ ☐ The existing documents and the principles contained therein have been formulated on a universal, abstract and context independent level, while the meaning, interpretation and implementation of each and every principle is different per context and per type of application of AI

Explain:

- ☐ ☐ There are simply too many guidelines and standards; data analysts, system developers and people / organizations that apply and use AI systems in practice do not know which standard they should adhere to (or choose the most favourable/least restrictive)

Explain:

.....

- ☐ ☐ The standards have only recently been developed; it takes time for these to be adopted in practice

Explain:

.....

- ☐ ☐ There is a lack of supervision and enforcement of the existing standards and principles

Explain:

.....

- ☐ ☐ Otherwise, namely....

Do you have any other comments, suggestions or tips for us when developing a guide *Non-Discrimination by design for AI Systems (which will also appear in English)*?

7.5 Verslagen van testgroepen

7.5.1 Testgroep 1

30 september 2020 16.00-17.30

Deelnemers:

- Janneke Gerards: Universiteit Utrecht
- Jan-Peter Loof: College voor de Rechten van de Mens
- Marlies van Eck: Hooghiemstra en partners
- Frederik Zuiderveen Borgesius: Radboud Universiteit

Team:

- Bart van der Sloot
- Alexander Hoogenboom
- Yvette Wagenveld
- Bram Visser

Bart van der Sloot: Heet welkom, zet de contouren van het project uit een en stelt alle deelnemers aan elkaar voor. Startvraag: Stel je bent een super ethisch iemand en je wil een AI-systeem bouwen zonder dat die discrimineert, hoe doe je dat? Kunnen we een handleiding voor zo iemand maken en zo ja, welke juridische principes zouden daar in moeten staan en hoe zouden die moeten worden geformuleerd?

Als inspiratiebron dient het ‘Little blue book’ van Jaap-Henk Hoepman, die een handleiding voor Privacy by Design heeft gemaakt. Wij zijn gevraagd door BZK om een handleiding voor Non-Discrimination by Design te maken.

Er bestaan namelijk al best wat standaarden voor discriminatie, AI en biases, maar deze lijken nog niet in de praktijk te worden gevolgd. Dat kan liggen aan onbekendheid, de veelheid van de richtsnoeren en met name ook met de vaagheid van de principes die daar in zijn vervat. ‘Zorg dat er geen bias in je dataset zit’; ‘Wees transparant’; etc. ‘Maar hoe dan en op welke punten’, zal dan wellicht de vraag zijn van de welwillende systeembouwer.

Eigenlijk zou je dus veel specifieker willen zijn met een handleiding. Maar dat is lastig. Veel juridische principes bevinden zich ofwel op een hoog over level – wees transparant, indirect onderscheid mag alleen als daar een redelijk belang voor is, etc. – ofwel hangen samen met de omstandigheden van het geval, zoals neergelegd in jurisprudentie en casuïstiek. Beide niveaus zijn dus ook lastig te vangen in een handleiding. Idealiter zou je tussen die twee niveaus in willen zitten.

Eerste vraag is kort: wat vinden jullie van het idee om een handleiding te maken? Is dat zinvol of niet; heeft het kans van slagen of niet en zo ja, wat vermag zo’n handleiding; en hoe zou die er uit kunnen zien?

Janneke Gerards: in 2002 heb ik een proefschrift geschreven dat ging over een toetsingsmodel voor rechters om gelijke-behandelingsvraagstukken te testen. Dit was een heel brede toets. Je redt het dan bijna alleen op een hoog abstractie model, wat nog steeds een heel moeilijk model is. Het helpt wel bij o.a. de besluitvorming. Vraag was ook of je op basis van zo’n model een proces zou kunnen digitaliseren. Conclusie was dat dat eigenlijk vrijwel onmogelijk was; er

zijn zo veel afwegingen, keuzes en zachte elementen in het model. Het rechtsgebied en de ontwikkelingen van AI zijn daarnaast heel erg in beweging, onze opvattingen over bijvoorbeeld proportionaliteit verandert steeds. Je kan dus een algemeen toetsingsmodel maken, maar verder zal het lastig zijn.

Frederik Zuiderveen Borgesius: ik ben het eens met Janneke. Je ziet ook vaak bij grote bedrijven dat zij ook gekleurd zijn. Gezichtsherkenningsoftware wordt bijvoorbeeld niet altijd getest op mensen van verschillende etniciteit/met een verschillende huidskleur, wat fouten kan opleveren. Door middel van zo'n handreiking of een Data Impact Assessment kan je zulke situaties eruit filteren. Er glibberen wel dingen doorheen, maar grote fouten kan je voorkomen.

Bart van der Sloot: je kan discriminatie/bias dus niet volledig voorkomen door middel van een handreiking, maar je kan wel de grootste fouten er uit filteren?

Frederik Zuiderveen Borgesius: klopt.

Marlies van Eck: je wil ervoor zorgen dat mensen door middel van een handreiking bepaalde punten expliciet meenemen in hun keuzes en beslissingen. Data scientists willen vaak horen of iets wel of niet mag. Je kunt dan gewoon de verboden gronden uit artikel 1 GW of AVG opsommen. Dat brengt je gelijk op een basale vraag, die niet ieder (overheids)organisatie zich stelt: waarom hebben we eigenlijk deze categorie gegevens? Vele organisaties hebben bijvoorbeeld data over man-vrouw, maar dat is vaak niet direct relevant. Data scientists vinden dat natuurlijk wel heel interessant. Maar het echte probleem is: je kunt wel bepaalde categorieën verbieden, maar daar kunnen ze zo om heen werken. Verbied je land van herkomst als categorie, dan maken ze systemen met proxies. Net zoals bij Netflix: je geeft niet door of je man of vrouw bent, maar dat weet het bedrijf natuurlijk wel.

Bart van der Sloot: Daar op aanhakend: zijn de klassieke discriminatiegronden dan wel relevant voor AI-systemen? Frederik, jij zei dat het die categorieën heel intuïtief en breed geaccepteerd zijn, maar is dat wel zo? Zeggen die categorieën een systeembouwer iets?

Frederik Zuiderveen Borgesius: er is al een bepaalde consensus over deze categorieën. Natuurlijk is het probleem dat AI discriminatie op basis van nieuwe gronden mogelijk maakt. Groepen kunnen worden samengesteld op basis van duizenden kleine kenmerken samen waaruit blijkt dat bepaalde mensen bijvoorbeeld meer ongelukken maken en dus een hogere verzekeringspremie moeten betalen. Er is daarover nog geen consensus dus moeten we dat niet in AI-systemen stoppen.

Bart van der Sloot: maar als we dat weglaten uit een handreiking, houden we dan nog iets over?

Janneke Gerards: proxies bestaat al heel lang. Het is ook al lang zo dat ongelijke behandeling plaatsvindt op grond van gronden die we niet verdacht vinden. Veel discriminatiekwesies gaan juist over dingen waarover morele consensus is en waarvan we vinden dat het niet aanvaardbaar is, bijvoorbeeld discriminatie op grond van kleur of geslacht. Problematisch onderscheid ligt vaak tegen deze gronden aan; wel is een goede dataset nodig om dat te controleren.

Bart van der Sloot: Dus een basis vinden binnen de gronden en de nieuwe vormen beter afwachten?

Janneke Gerards: een benchmark uitvoeren. Over de verboden gronden zijn we het eens, bij nieuwe situaties hebben we iets om naar te kijken. Dan kan je meer je rechtvaardigingsmodel inrichten en bijvoorbeeld hogere eisen stellen aan data. Dus niet helemaal vanaf zien.

Jan-Peter Loof: ik ben het eens met Janneke. Je ziet de consensus, die is er al decennialang op deze gronden. De consensus is niet dat directe discriminatie in alle gevallen verboden is, er kan altijd een objectieve rechtvaardiging zijn. Die rechtvaardiging is ingewikkelder in een AI-systeem vorm te geven, want het is een vloeiende schaal die afhankelijk is van omstandigheden en hoe noodzakelijk het is.

Bart van der Sloot: techneuten hebben gezegd dat de we handreiking maken voor 20 jaar geleden, want AI werkt zo niet meer. Zij zeggen dat de categorieën hun niks zegt. Je kan het wel uitsluiten, maar je wint er niks mee, want ze kunnen het eenvoudig omzeilen.

Janneke Gerards: je kan wel naar resultaten kijken. Als je een algoritmisch systeem hebt en je laat een standaard los, dan zie je dat dit veel negatiever is voor een bepaalde groep.

Bart van der Sloot: zit het in de disproportionaliteit?

Janneke Gerards: ja, klopt. Kan je wat we onredelijk vinden in cijfers uitdrukken? Daar zit een grens aan. Er is een bandbreedte waar je niks mee kunt. Je kan wel zeggen dat er rode vlaggetjes gaan branden op het moment dat je boven een grenswaarde komt. Dan is het belangrijk dat dat in ieder geval gecheckt wordt, zodat je niet dat resultaat krijgt waarbij een bepaalde groep door algoritme eruit wordt geplukt.

Frederik Zuiderveen Borgesius: het perfecte is niet te halen in een handreiking, maar er kan wel veel worden gewonnen. Een techneut die de handreiking leest is welwillend en als hem via een handreiking kan worden uitgelegd hoe het moet, is dat een goed begin. Ik denk dat er geen moreel probleem is.

Marlies van Eck: ik denk wel dat er een moreel probleem is. Amnesty heeft een rapport uitgebracht. Dat ging erover dat wanneer een auto met een Pools kenteken in Roermond wordt gezien, er wordt gedacht dat deze auto niet voor de outlet komt. Cijfers wezen uit dat er effectieve controle kon komen. Toen dat gepresenteerd werd dacht ik dat dit onrechtmatig was, naast mij dacht één iemand er ook zo over, maar de rest was er vooral van overtuigd dat de Polen een groot probleem waren. Ik zei dat dit ook niet van de rechter mocht, maar dat maakte niet uit. Je mag niet profileren op grond van land van herkomst maar dit kan wel als iedereen geprofileerd wordt. Zij vonden dat irreëel, want ze gaven aan dat iedereen weet dat bijvoorbeeld mensen uit Marokko een tweede woning hebben. Daar kun je dus naar kijken. Moreel gezien hebben we daar dus nog veel te winnen. We zijn mensen aan het categoriseren op basis van kenmerken terwijl dat niet mag. Een data scientist discrimineert ook, in neutrale zin van het woord. Het kan volgens mij geen kwaad om moraliteit naar voren te brengen.

Frederik Zuiderveen Borgesius: mee eens.

Bart van der Sloot: de gemene deler is dus eigenlijk: begin met de basis, want daar is veel te winnen. Gewoon uitleggen wat de basale dingen zijn – direct onderscheid op basis van ras mag bijvoorbeeld niet. Dat creëert bewustwording. Je kan dan ook meegeven goed te kijken naar

indirecte discriminatie. Denk je als techneut dat je wel de data ten aanzien van verboden gronden (direct of indirect) nodig hebt? Leg dan uit waarom. Snap je het niet of twijfel je? Vraag een jurist. Dat soort dingen.

Frederik Zuiderveen Borgesius: Ja, goede samenvatting. Daarnaast: je kan ook procedurele tips geven. Je moet niet-juristen kunnen uitleggen dat directe discriminatie nooit mag. Voor indirecte discriminatie heb je meer pagina's nodig, maar ook dat krijg je uitgelegd. Je moet proportionaliteit uitleggen: het is ingewikkeld maar zoek vroeg in het project hulp. Dat kan helpen, ze zullen wel gaan klagen dat je geen handige tips geeft, maar zo zit het eenmaal in elkaar.

Bart van der Sloot: daar denken we ook aan. Gewoon een stappenplan.

Alexander heeft voor ons jurisprudentie in kaart gebracht. Alexander, jij had de vraag of het überhaupt realistisch is om een handreiking te krijgen of moet dit per sector worden gedaan, omdat daar ook andere standaarden gelden.

Alexander Hoogenboom: EHRM en Awgb hebben andere gronden, maar overlappen soms wel. Van welke jurisprudentie gaan we uit? Daar worstelen we mee. Welk lijstje neem je mee?

Jan-Peter Loof: consensus zit dus eigenlijk op de beperkte lijst van de Awgb en het EU-recht en minder op het EVRM, daar zit namelijk niet het harde verschil tussen directe en indirecte discriminatie.

Alexander Hoogenboom: als je kijkt naar wat het EHRM bij onderwijs doet, het is onduidelijk wie waar bevoegd voor is.

Janneke Gerards: mij is het gelukt om een algemeen toetsingsmodel te maken op alle situaties, zoals zowel horizontale als verticale verhouding. Je moet het abstraheren, maar veel eisen komen in veel contexten terug. Het gaat met name over de punten waarop consensus. Ik ken weinig EHRM uitspraken waarin verdachte gronden en rechtvaardigingen aanwezig zijn. Vaak is er sprake van onrechtvaardig onderscheid.

Jan-Peter Loof: er zijn een paar gronden waar weinig jurisprudentie over is.

Janneke Gerards: gronden zijn in sommige context wel verdacht en in sommige context niet, zowel bij het HvJ als het College. Geslacht hoeft niet altijd een verdachte grond te zijn, dit hangt van de context af. Overall bestaan uitzonderingen op; alles is afhankelijk van de context; ieder rechtcollege interpreteert zaken net weer wat anders. Dat kun je allemaal niet in een model stoppen. Voor het doel van deze tool is het belangrijker om die morele vraag voorop te stellen. Die kan je goed na toetsen met algemene criteria, dit hoeft niet precies op de Europese normen.

Alexander Hoogenboom: het idee van rode vlaggetjes is goed. Als je dat hebt dan kom je best ver bij techneuten.

Frederik Zuiderveen Borgesius: de Awgb-gronden moeten wel genoemd worden, toch?

Alexander Hoogenboom: er zijn gronden waarover we niet zeker weten of een grond verdacht is. Wordt het voor Nederland of in het algemeen geschreven?

Janneke Gerards: wat betreft zaken over leeftijd, op EHRM-niveau is dat discutabel omdat er nog maar één zaak bestaat waarin daarover is besloten. Ze vinden het misschien wel verdacht. Er zijn wel veel criteria genoteerd of iets verdacht is. Aansluiten bij lijstjes vind ik riskant. We willen de lijst al heel lang uitbreiden, maar je hebt wel criteria nodig die te maken hebben met kwetsbare groepen. Als het in de Grondwet en onder het EU-recht valt, kan je redelijk zeker zijn dat het een grond is.

Frederik Zuiderveen Borgesius: zou het niet goed zijn om toch wel bijvoorbeeld het lijstje uit Awgb te noemen, zodat je het merkt als je bijvoorbeeld per ongeluk homoseksuelen bent vergeten toe te voegen aan het systeem. Ontwerp het systeem zo dat als iemand een relatie heeft, er ook de mogelijkheid bestaat dat dit met hetzelfde geslacht is en niet alleen met het andere geslacht.

Janneke Gerards: ik zou een ruim lijstje opnemen. Minimaal de gronden uit Awgb en uit de EU. Ook gronden toevoegen waaraan EHRM heeft aangenomen dat ze verdacht zijn. Daarnaast ook gronden opnemen die gelet op factoren kunnen worden toegevoegd.

Marlies van Eck: ik vraag me af hoe data scientists op de vragen hebben gereageerd. Hoe moeilijk is het voor hen om iets toe te voegen? Zij zijn wel bewust dat data iets representeert in tegenstelling tot iemand die software ontwikkelt. Hoe gingen de gesprekken?

Bart van der Sloot: data scientist zijn zich vaak al bewust van bias in de data en hoe keuzes in het model/systeem de resultaten beïnvloeden. Wel hebben data scientists een eigen jargon; bijvoorbeeld al die gronden die voor juristen duidelijk zijn, dat zegt het niet direct iets. Waarom is het per sé prolematischer om onderscheid te maken op basis van nationaliteit dan op basis van de kleur van je auto? Gaat het niet om het effect die dat onderscheid heeft? Bovendien, als je zowel de direct identifiers als de indirecte identifiers er uit wil halen, dan hou je eigenlijk geen bruikbaar systeem meer over. Alles verwijst indirect naar alles, bv postcode is voor 87% van de gevallen een goede voorspeller voor etniciteit; muzieksmaak voor 76%; kleur van de bank voor 65%; en type shampoo voor 54%. Wil je al die indirecte identifiers uit de dataset halen; verbied je data-analisten daar iets mee te doen? Dan hoe je steeds minder over.

Frederik Zuiderveen Borgesius: misschien verstandig om naar de uitkomsten te kijken?

Bart van der Sloot: dan zit je alleen ex post facto te kijken en dan is vaak ook niet helemaal duidelijk waarom een systeem zo is gekomen en/of dat dan is gekomen door een bias of niet. Voor beide valt iets te zeggen.

Frederik Zuiderveen Borgesius: ja, beide doen. Stel, binnen een systeem worden mensen uit een bepaalde postcode altijd op een bepaalde manier behandeld, als de ontwikkelaars dan naar het boeken kijken, zullen ze het misschien wel anders doen.

Marlies van Eck: Zuurmond zei dat als je iets neerlegt bij iemand uit de praktijk en die kijkt naar bijvoorbeeld een opgerold briefje. De ene ziet niets, de ander denkt: cokesnuiver, want die rollen briefjes op. Je hebt dus altijd meerdere mensen nodig met diverse achtergrond, ook bij het bouwen van het systeem en het inrichten van de data en voor het testen van het systeem. De

ene persoon kan zien dat een dataset of een systeem in de praktijk discrimineert, de andere mogelijk niet.

Bart van der Sloot: we hebben een document gemaakt over standaarden en fases en hoe kan je dat inrichten per fase. Werkt dit of niet? Hierop gebaseerd maken we de handreiking. Ik wil het nu voorleggen, roep het als iets in je opkomt.

Marlies van Eck: ik vind het een mooi document. Ergens staat: het recht op discriminatie? We zeggen in bestuursrecht ook wel het gelijkheidsbeginsel, maar je hebt ook het recht op een afwijkende behandeling als de situatie afwijkend is. Je hebt het recht niet over een kam te worden geschoren, in zekere zin dus een recht om “gediscrimineerd” te worden. Bij fairness zeg ik tegen de studenten dat de Amerikaanse literatuur en jurisprudentie het altijd op fairness gooit terwijl in de EU het op discriminatie wordt gegooit. Fairness is dus een lastige term.

Bart van der Sloot: zou jij de term fairness vermijden?

Marlies van Eck: ik zou het wel goed proberen te definiëren om in ieder geval aan te geven dat beide groepen verschillende dingen er onder verstaan en dan uitleggen wat juristen eronder verstaan. Ik ben voorstander van dat billijkheid in bestuursrecht een rol gaat spelen. Fairness wordt soms als fair vertaald en soms als billijk vertaald. Bijvoorbeeld, je hebt de Top600: een lijst met namen van jongens in Amsterdam met een strafblad. Journalisten deden onderzoek naar de Top400: een broertjes en zusjes lijst. Ze hebben een model gemaakt om te bepalen welke groep in de kring van Top600 zitten en high potential hebben om ook crimineel te zijn. Journalisten gaven aan dat ze het idee hadden dat iedereen op die lijst een kleur had. Je moet dus ergens wel bijhouden wat de etniciteit is van mensen om op dit soort vragen achteraf antwoord te kunnen geven.

Bart van der Sloot: om dat te controleren heb je data nodig. Als we willen dat een systeem niet discrimineert moeten we wel data verzamelen voor een check. Jij zegt dat moet je zo houden want als er dan een vraag komt, dan heb je in ieder geval de data om het te controleren.

Marlies van Eck: ja, want die journalisten konden het niet constateren. Hetzelfde was bij de toeslagenaffaire, de journalisten die aanwezig waren zagen dat zij de enige kaaskoppen waren.

Jan-Peter Loof: Ik vond alles wat ik hoorde langskomen in het document logisch. Waar je nu mee bezig bent is kijken naar de voorkant om in te richten of je bepaalde dingen kan voorkomen. Discriminatie kan onbedoeld zijn, maar dat zie je pas op het einde. Dat komt lang niet altijd door op slinkse wijze van proxies gebruik te maken; vaak is het met de beste bedoelingen gebeurd. De uitkomst moet in de gaten worden gehouden, om daarna opnieuw te kijken. Aan de voorkant voorkomen is belangrijk, maar het is ook van belang om te evalueren wat de uitkomsten zijn. Wat is er ingewikkeld aan? Wie gaat dat doen? Die uitkomsten worden ook in fysieke werkelijkheid pas duidelijk als een individu een klacht indient. Net als wat Marlies zei over het zaaltje bij de toeslagen. Wat gebeurt hier? Dat is het ingewikkelde eraan. Wat voor mechanisme heb je om discriminatie zichtbaar te maken en wie gaat dat doen? Ik heb hier geen antwoord voor maar het is wel relevant. Hierin zit de mogelijkheid tot verbetering.

Bart van der Sloot: is inderdaad een lastig punt. Moet wel benoemd worden.

Jan-Peter Loof: in zoverre het gelijke behandelingsrecht en handhavingsmechanisme is het heel klassiek. Een burger moet een instantie hebben om een klacht in te dienen, maar de vraag is in hoeverre die burger de effecten van het systeem kan doorzien. Die zie je soms pas op cumulatief geaggregeerd niveau. Burger zal niet snel in actie komen, want hij heeft het niet door. Dit is soms ook bij het sollicitatieproces te zien. Wat voor juridisch handhavingsmechanisme past bij deze digitale werkelijkheid?

Frederik Zuiderveen Borgesius: document is heel mooi, logisch en solide. Zou toevoegen dat in sommige sectoren nog meer eisen spelen die bijzonder zijn. Stel, een predicted policing systeem moet bijvoorbeeld ook rekening houden met de onschuldpresumptie; ook voorbeelden en contexten noemen te verduidelijking. Daarnaast zeggen dat als iemand denkt dat er wat speelt, dat ze dit moeten checken. Ik zou ze waarschuwen.

Bart van der Sloot: we willen de handreiking testen in drie sectoren om zo ook misschien nog een extra concretisering te kunnen maken voor die sectoren.

Frederik Zuiderveen Borgesius: dat maakt het beter. Je hebt dan de voorbeelden. Je kan dan die uitkomst checken en toevoegen aan de handreiking. Het is logisch dat je de AVG noemt, maar het wordt moeilijk af te bakenen als je alles gebruikt. Ik krijg veel vragen van data scientists. Veel niet-juristen gaan informatie Googelen en komen dan vaak op Amerikaanse publicaties, daar hebben ze dan vragen over. Daar werkt het toch een stuk anders.

Bart van der Sloot: wat vinden jullie van de link met de AVG?

Marlies van Eck: stel, je hebt in de medische sector scanningapparatuur die dingen moet herkennen, dan kunnen daar discriminatoire factoren inzitten want bepaalde dingen kunnen bij de ene persoon sneller zichtbaar zijn en daardoor sneller gescand worden dan bij de ander, maar dit zijn geen persoonsgegevens.

Bart van der Sloot: eens; maar de procedurele zaken uit de AVG zijn op zich handig. Een audit doen, documentatieplicht, transparantie, etc., kunnen denk ik nooit kwaad, ook al betreft het geen verwerking van persoonsgegevens.

Marlies van Eck: eens.

Janneke Gerards: ik vind het model indrukwekkend. Ik ben het eens met Marlies en Jan-Peter dat het goed is om ook de evaluatie en auditing verder uit te werken. Ander punt hangt samen met wat Frederik zei: het is heel breed en er staan veel elementen in die relevant zijn voor discriminatie maar het zijn algemene criteria voor hoe je een systeem bouwt. Ik vroeg me af hoe dit verhoudt tot alle bestaande richtlijnen en modellen. Wordt dit een integrale vervanging? Zo ja, dan moeten er elementen bij of focust dit op discriminatie en gelijk behandeling? En heb je dan wel al die elementen nodig? Veel lijkt te gaan over deugdelijkheid en betrouwbaarheid. Dat geldt ook een beetje voor het recht op privacy en onschuldpresumptie. Het gaat om ethische algoritmen en mensenrechten. Moet je dit allemaal in dit model willen schrijven? Of gaat dit model over detecteren van problemen en verboden gronden waarvoor rechtvaardiging bestaat? Hoe omvattend is dit model? Of is het te breed?

Bart van der Sloot: I feel your pain. Dat is precies het probleem bij aanvang van het project. Je wilt niet een handreiking maken waarvan data-analysten en systeembouwers denken, als ik

hier aan voldoe, doe ik het goed, terwijl het maar over een klein niche gaat en bovendien hangt vele met elkaar samen. Je kunt discriminatie niet in isolement zien, het hangt ook samen met redelijkheid, privacy, transparantie, etc. Aan de andere kant wil je ook geen Reviaanse handleiding maken die alle andere handleidingen vervangt, behalve de Bijbel en het telefoonboek. Er bestaat al zo veel en er zijn uiteindelijk zo veel regels waar je meer rekening mee moet houden dat het niet zinnig is om dat allemaal weer over te nemen in een nieuwe handreiking. Het wordt dus balanceren, waarbij het zoveel mogelijk moet gaan naar strikt discriminatie/gelijkebehandeling.

Heel veel dank voor jullie waardevolle feedback!

7.5.2 Testgroep 2

Expert group II AI system principles for non-discrimination

Expert participants:

- **Seda Gürses (SG)** (TU Delft): Technology policy and management, background in computer science, computational infrastructures
- **Harry van Geijn (HvG)** (Microsoft): Digital advisor, senior advisor to largest customers on how to implement technology and apply ethics to it. Long term interest in trustworthy technology
- **Violetta Misheva (VM)** (ABN Amro): Data scientist, PhD in econometrics, internal AI principles fairness, explainability. Works on these issues at ABN Amro
- **Yugesh Raghoenath (YR)** (Rotterdam): Data scientist, master in econometrics - Part time PhD student - probabilistic ML

Project team:

- **Esther Keymolen (EK)**: researcher TILT, philosophy of technology
- **Merel Noorman (MN)**: researcher TILT, STS and AI
- **Hilde Weerts (HW)**: research engineer, works on tools and applied research on ML, industrial engineer and computer science
- **Yvette Wagenveld (YW)**: student assistant
- **Bram Visser (BW)**: student assistant

Introduction:

MN kicks off the session with an introduction to the project and an explanation of the principles/standards:

- AI public values and human rights - Project commissioned by Ministry of Interior Affairs: can we apply those principles to AI systems? (fairness, no bias, non-discrimination)
- We want to translate legal formal norms and rules to AI system principles in order to come up with visual, easy to use booklet for policymakers and developers that can serve as a guide.

Problem analysis:

MN briefly highlight some problems that the research team has identified in the literature review:

- Bias, fairness, explainability
- Technical issues: bias in/out; concept drift; formalising fairness
- Societal issues: effectiveness and reliability, filter bubble and Mattheus effect, legitimacy, and trust

For this session, MN points out, that the research team is particularly interested in what the participants as experts think about the standards the team has identified and whether or not they are useful.

The team has tried to translate legal norms to concrete system principles (SP). They have used a process model for structuring development of data mining systems (CRISP DM) to map legal norms to different phases of the process. This mapping was based on the literature review of existing standards and principles.

The team has organised three expert sessions to find out what the experts think of the current overview of principles and standards? What is missing? What would the experts do differently?

Discussion principles and standards

EK leads the discussions about the documents that were sent to the participants ahead of the meeting (summary of the draft report and overview of the standards and principles). She starts off with the first question:

Q1: suppose you are an ethical system builder: what should a guide to imply non-discrimination look like? What are the most important things to have guidance on, what do you look for (key words, institutions, people):

HvG: Look for relevant stakeholders in the model: make up a **diverse team**. Not just look for technical guidance but also how to use model in a real situation/context.

SG: Agrees with **HvG**, what are the goals of the given industry in terms of population management? Is population management the right approach and will the system benefit some over others?

EK: Recap of SG's argument: SG wants to know how tech will be used? And what purposes and subpopulations it would serve?

SG: What is the change we want to introduce to the world with this system?

VM: Notes that she, as a data scientist in a big organization, is limited in what she can control: diversity is not always possible. Problems are mostly very concrete. What is the problem? Target audience? What solutions? She struggles with what fairness is in concrete situations. How to implement it? GDPR is a barrier to training systems in terms of sensitive data. She feels insecure about whether she's allowed to process data.

EK recap: "in order to combat discrimination, you feel insecure to process certain data?"

YR: agrees with **HvG**: most problems are organizational, not technical. Data scientists are asked to develop a model; the governance is not in place to ensure the right questions are asked. How to build a team? How to put in place tests? Important to ask the right questions from the start.

HvG: "I think the key in AI-systems is making sure to pose the ethical questions from the start and build in functionality for ethics from the start, instead of re-engineering it post factum."

Q2: About the standards, we have 6 steps. The first step is problem definition. What do you like about the first phase? What should we keep? What don't you like? Or what do you think isn't important?

VM: Likes: concrete and helpful questions in all three pillars

Doubts: "As our organization is diverse and big, it might not be applicable in every use case." Generic might work well, but in her organization they are already doing some of these things. Implementation should be embedded in existing processes which are scattered throughout the

organisation, because every part of an organization has a different structure, but it is a good starting point. It might be too generic

YR: Likes: In developing standards, usually a technical approach is chosen, while standardising the way in which we are compliant to legal standards is also a good approach. Very good starting point to avoid unnecessary technology. “For me the second question in the pillar [*is the necessary and proportional to collect and process more data for the project?*] should be the starting point.”

transparency: Rotterdam split up transparency (procedural transparency (record, justify choices and proactively performing) - technical transparency (might not always be possible))

These have different depths to get inside the black box. That gives you different ways to go at it.

HW: what do you mean with procedural transparency?

YR: it is both organizational and technical. The description of the process, the way in which we move from a business goal to a technical solution, is procedural. For example, documenting who is responsible for which decision is procedural. Technical transparency entails details with respect to the technical implementation of the algorithm, such as, which algorithm, which implementation (libraries/packages), with parameter initialisation.

SG: Likes: legal questions, documentation proposal. It is often iterative, so good to document thoroughly.

Dislikes/missing: a different way to organize it. Right now, it is focused on the model, however the model is built for an organizational reason (optimizing workflow eg). If you automate a process (and replace workers) how will the service affect the world and society? What kind of activities get metricized and open for central control, what happens to those parts that are not automated? You can ask more questions:

- Economic considerations: how much of the budget moves to the technical system? What kind of economic transformation is the organization going through?
- Governance: how much is going to be outsourced to service providers?
- Will the structure of the organization change as a result of the automation?

HvG: Good and comprehensive list. It is valuable. Agrees with Seda on missing things. Context in which tech is used is very important. Ethical thing about inclusivity: e.g. online mortgage application excludes people without internet access. Thus, the context is far more important than the tech itself.

Transparency: extremely tricky. What is transparency? To whom? To a technical expert? To a user? If the use of the model becomes subject to a lawsuit both parties must have equal access. Transparency is thus complex, deserves more elaboration in our standards. Ethical model of Microsoft emphasizes accountability. Which human being is responsible for each of the steps of the model? You can always track it back to a human. Human responsibility is key.

EK: Do we need to add a pre-step to our current 6 steps? Or imbed the suggestions about understanding the context in our existing standards?

HvG: It is an extension. You need to expand that. Add contextual questions: who are the stakeholders?

At Microsoft they have a list of use cases:

- is the model going to cause harm (physical or mental)?
- potentially infringe on liberties and human rights, does it exclude people?
- is the model denying people fundamental services? eg credit worthiness → go through ethical commission

SG: Important to put organization and institution at the center of the analysis to avoid making it seem like that there is only algorithm and people and the institutions disappear. This is not the case, there are also institutions that need to ensure protection of public values. As for “inclusion”, systems can be predatory, inclusion in such systems may hurt people. So inclusion is very important, but also very tricky (eg higher interest rate loans for people with higher risk scores may mean that the poor pay more for taking out loans, such a system will impact poor people unjustly)

HvG: agrees with Seda, but is an issue that should be handled way before the data scientist goes to work

Q3: Data collection and storage phase: what are the challenges for data scientist in terms of anti-discrimination:

YR: Systems are designed to be discriminatory: they categorize, identify patterns, etc. This is the goal of the project, because we want to optimize some process. We should develop discriminatory systems, that comply to our standards (moral standards, technical standards, standards imposed by legislation). Finding the right balance can be difficult. With regard to the technical aspects, it is difficult to reason about what is not there: representativeness, bias, missing values. E.g. using the value of homes to predict another target variable: you have a proxy for income/wealth. With these questions as they stand in the third column, the problem of proxies is not addressed. Maybe the second question could catch it, but this one is more about what is represented.

VM: With regard to predatory systems, she agrees with **SG** and **YR**: it requires ethical evaluation for a model. Data collection and storage is a very important step: many requirements for storage, access, duration, format etc.. She struggles with technical pillar. These are important questions but difficult to answer, because I’m not allowed to do some things. “How can I see if it’s discriminating if I’m not allowed to do this from a legal perspective?” Checking for proxy’s or protected categories: this example shows the disconnect between what is required and what is allowed from legal perspective.

HvG: problem is that the law describes a solution instead of addressing the problem, thus adding to the problem. Suggested addition: dependent on outside parties that provide data sets, have a clear view of where data comes from and what are their standards and objectives. New category of security risks, both data sets and models are attacked in different ways to manipulate. What is the goal and standards and processes of third-party partners? What is the quality of the sets? Important to consider.

YR: which organisation delivered the data, what is the date of the dataset... etc

SG: thinks there is fundamental tension in the order to do things right with respect to discrimination, doing ML right has a tendency to induce more data collection, as privacy

engineer we look at what data you REALLY need. ‘Data lake’ (*pooling a lot of data from different sources*) practices, for example, should be discouraged in organizations.

Comment on **HvG**: there is whole pipeline of data and models that data analysts depend on. We see a lot of issues there. An organization ends up subscribing to a pipeline and then it becomes difficult to have accountability. She gives the example of a study on Airbnb and gentrification. The study focused on neighborhoods that had Airbnb, leaving out the neighborhoods that are already treated unequally/underserved. It shows a tension between available data and how it can make certain populations and environments more or less visible.

If you look at the Google case (mislabelled discriminatory images): if a person complains about the model, Google will just remove that part or change the classification threshold.

Data collection is a continuous process, amorphous, all over the place, "not a static step".

HvG: they (*data lakes*) become circular references if you don't watch them!

SG: a lot of organizations lose control over data in this step

Q4: Data analysis and preparation: what do you like about it and what are doubts and worries:

HvG: Very hard subject, the data pipeline problem plays a role here as well (**SG**). Check that data is current and up to date. This opens up an entire box of problems, it is a very hard part. The problem is not listing the components, the devil is in the details. How to check? What to compare it to? What is correct? This is one of the core elements of data science.

EK: Is it too broad?

HvG: would not know how to make it more specific without referring to complete field of science. Could be useful to add a question like: have you done the work to ensure this?

VM: devil is in the details. hard to make in concrete. Likes the proposed questions: relevant expertise is useful, relevant data scientist and process engineers etc... how the process is lacking.

EK: who should we include?

VM: depends on context and problem: data scientist, product owners, business analysts, relation managers. various knowledge to build a complete picture. On the technical side: it is not so easy to use various software libraries. You cannot be fair to the entire population, it is not clear what kind of fairness we should employ, it is not that simple as using a package in a software library.

YR: Agrees with relevant expertise side of things: experts in the privacy field should have a proper legal background, organisations need data scientist that have statistical and machine learning backgrounds. The same goes for management. . Relevant expertise becomes more important as we go more into detail. Having people that know about the details helps to address these difficulties.

HW: What do you think about the neglected data subject (relevant stakeholders): at what phase should they be included?

Stakeholders = people working in org, civil society organisation, people impacted by the system. “The organisational view is 'my main theme' I am following.”

YR: as **SG** mentioned data collection phase is not linear. The entire development of the product is way more chaotic than the CRISP MD shows. There should be a way in which principles and checks described should be more cyclical, instead of a one-time thing. It should be implemented in a way that those checks are carried out throughout and not just at a certain phase because that is not how it works.

VM: Involve various stakeholders when defining the problem. Set up a process in which you should keep engaging the stakeholders. Colleagues from legal, compliance, ... “When we do not have access to stakeholders, we think about what kind of impact ML process can have on stakeholders.”

EK: Recap: Assess what stakeholders to involve both inside and outside the organisation

Q5: Modelling: Likes and dislikes: What do you think about the legal principles here? focus is on statistics, does that make sense from a data science perspective? Can you work with that?

VM: legal principles: some make sense, some are vague.

What does it mean to develop in a neutral manner? Make it more concrete to actionable actions for data scientist.

Big issue is the technical part: explain choices in terms of explainability and transparency: whole different process besides fairness, here is it mentioned as a matter of fact. What does it mean to explain a model to stakeholder? How to monitor and validate them? if we want to tackle explainability we should make it broader.

Participation of stakeholder: citizens and external partners: I would involve citizens in an earlier stage.

HvG: If a model becomes subject to a court case, explainability is important.

Technical side: understanding of the limitations of the system. Additional bias by limitations in tech: facial recognition can be harder on darker skin, not necessarily data dependent but technologically dependent

SG: just looking at the algorithm will not give you the whole answer but one needs to look at the entire picture (what data is the model trained on etc...)

1) we need more expertise, but it is about organisational expertise. Would love to see suggestions on how to bring stakeholders into the process. Diversity is important but needs to be complemented with expertise in the know of the consequences for different populations.

2) we can make sure parity measures (as in fairness metrics) are satisfied but that might not necessarily avoid discrimination. Parity measure is feasible but what is the impact on the world? Take relevant stakeholders into account!

HvG: Neutrality is hard to define. make sure different aspects of reliability etc are taken into account.

YR: Investigate strengths and weaknesses of models of pipeline. “In the different phases we start from different points of view.” Some department has business goals, but in the step in the overview we begin with statistical principles. Maybe begin with legal standpoint: what decision are we trying to make and what are the risk associated with it, instead of beginning with statistical approach.

HvG: You need to look at this in the broader context, front to back and this is important throughout, important not to be myopic. Assess use case carefully! As all these steps will differ according to use case, so keep broader context into perspective is the running thread.

EK: Indeed, important points made, guide should not give a false sense of security.

SG: E.g. bias mitigation: is this a term that data analysts feel comfortable with? Can you claim you mitigated all bias? She sees it more as 'impact'. Discriminatory impact or... If in the bigger context we identify that this system is bad for immigrant women, fairness frameworks are not going to help: it is hard to represent intersectionality. Fairness metrics do not allow you to identify the problem. So, what do you tell data scientists to do?

YR: Notes that the question is on point. It links to his earlier answer: Bias is important, but it is quite a narrow concept. There is much more to it: it is more about what do we do, what are we developing, what are strengths and weaknesses, negative consequences? Bias will always be part of the system because we are trying to develop discriminating systems. We should however only allow bias that complies with our standards and discriminate based on factors we agree with. Discriminating between accepting or denying a customer a mortgage might be okay based on wage, but discriminating based on race or religion is not. This choice should be aligned with the organisational and societal needs and values.

SG: How do you deal with errors? How do you prioritise?

VM: Prefers 'bias' over 'impact' as data scientist are not aware about problems that models may generate in the broader context. It is better to use bias to raise awareness amongst data scientists for broader social impact of systems.

YR: Much of the standards on the technical side are about making sure the model is of good quality. I would not necessarily restrict us to 'bias mitigation' 'fairness metric' etc.. It is not about the specific technique, but about the strengths and weaknesses of the system. Bias and fairness are only two of the candidates one could look at. Reliability of the model is a broader concept. Make the decision of when a model is reliable with the relevant stakeholders.

HvG: If a model provides answers, also mention confidence of the outcome. Tricky as well: when you say my model is only 60% accurate: even with low confidence level people will trust the outcome of a system more than they would trust another human's judgement

Q6: Evaluation and deployment:

VM: important to have deployment stage, we should treat this as a separate stage. She had more comments on the evaluation stage: where is the trade-off between fairness and bias mitigation and accuracy of the model. Fairness may reduce algorithmic performance. Bias can be dependent on evaluation model. So, document choice for evaluation model. Peer review can also be helpful, for checking for the code etc... Audit the whole modelling process end-to-end (in model training stage or evaluation stage)

HvG: A major concern is that damage is only done once the model is in use. So continuous monitoring of model in use is important as the model learns! (remember racist algorithm comments that Microsoft Tay).

Over time, the system changes so it may be fine at a certain point in time, but this does not guarantee that it will be fine 3 months later. Governance around use of the model is the most important part. The real work only starts once it is in place.

Key: we cannot solve everything with technology, organization has a broad understanding of use case and necessity of all the problems that come with it, not just a problem of data scientists

SG: Evaluate services that you are going to use. There needs to be clarity on what services are going to be used and how to evaluate before data analysts go to work.

“We cannot do privacy engineering in this ecosystem, it will also possibly be difficult to do it with bias.”

Non-discriminatory or fair system is not possible, there are always unequal outcomes. Important from a social justice perspective to not claim a system is fair.

HvG: the ethical framework in itself is biased (WEIRD countries - Western, Educated, Industrialized, Rich and Developed)

SG: Fairness is subjective, so how to advise people on how contestation is possible. if you claim system is non-discriminatory you close the opportunity for contestation. Very important to have a possibility for contestation. (Comment in chat: I think testing (A/B testing) and testing systems in the wild can have serious ethical implications, as they test versions of models which might have had discriminatory impact).

EK: overall framing of the guide is important

YR: good to split phase 5 and phase 6

Violetta mentioned good programming practices (peer reviews): get a list of libraries we want to work with because we know they are good, list of companies that have good practices etc...

HvG: important to address weaknesses and limitations! Being aware of the limitations of the product and data sets is crucial!

Q7: can you give your thoughts on our different visualisations:

(participants are shown a visualisation of a developing process made by Google for ML as an example of a more linear model)

Hvg: Not in favour, more in favour of Agile loop. Cycle for development and operational phases. Systems are constantly being maintained and developed. Dev ops.

YR: more in favour of Agile, building blocks instead of sequences.

SG: this is only an ML process, nothing about organizational workflow part of which is the development of the ML solution. so important to not focus on data analyst problem, but on organization. Give the bigger picture and not pile all the problems on the data scientist.

VM: more of a loop, introduce a loop between numbers.

7.5.3 Testgroep 3

Vrijdag 16 oktober van 14.00-16.00

Deelnemers:

- Francien Desechne : Universiteit Leiden
- Monique Steijns : WRR
- Sennay Ghebreab : Universiteit van Amsterdam
- Ronald Leenes : Universiteit Tilburg

Team:

- Bart van der Sloot
- Merel Noorman
- Yvette Wagenveld
- Bram Visser

Bart: We zijn gevraagd door Binnenlandse Zaken een handreiking voor techneuten te maken. Kun je wat Jaap-Henk Hoepman heeft gedaan ten aanzien van Privacy by design (Little Blue Book) ook doen voor non-discriminatie: non-discrimination by design. De juridische begrippen zijn vaak OF hoog over: redelijk, transparant, proportioneel OF heel specifiek in casuïstiek, mag dit wel of niet of moet bepaald criterium sus of zo uitgelegd. Wij proberen een tussenlaag te creëren van begrippen die concreet genoeg zijn voor een techneut om mee aan de slag te gaan, maar ook algemeen genoeg om een AI systeem mee te creëren dat niet discrimineert. Hoe zou je zo een handreiking kunnen maken? Hebben die ambities überhaupt kans op slagen? Heeft het zin?

Francien: verwachtingsmanagement is belangrijk want bij informatici is er een behoefte duidelijke definities te krijgen. Als de verwachting is ‘vermijd alle bias’ dan heeft het project geen zin. Het gaat erom hoe je er mee omgaat. Het zou wel duidelijk moeten maken wat de beperkingen van een bepaald design zijn. Denk aan de ‘21 definitions of fairness’: technische uiteenzetting die duidelijk maakt dat er niet één oplossing is en dat elke oplossing zijn beperkingen heeft. Als dat uit een handleiding zou kunnen spreken zou het goed zijn. Welke formalisering past op jouw dataset? Hoe is de wereld gemodeleerd in jouw situatie? Wat weet je over de info die je niet hebt etc... Dat soort hoger niveau richtlijnen zijn belangrijk voor bewustzijn van contextafhankelijkheid. Ben je bewust van de inherente beperkingen van de tools die je hebt. Bewust worden van de beperkingen van je systeem. Er zijn verschillende juridische noties van fairness; de keuze die je maakt voor formalisering creëert nieuwe bias die niet per se schadelijk is. Maar het is belangrijk om de impact van de keuzes die je maakt te kennen.

Senay: Eens met Francien. Om terug te komen op eerder punt van tussenlaag zoeken, tussen hoog over en concrete casuïstiek: weet niet of dat meest nuttige aanpak is want daarin kan je vaag blijven. Misschien toch goed om het systeembouwer perspectief te blijven houden en wijzen op spanningen en dynamieken en op die manier dingen aan te pakken. Wat bias betreft, helemaal mee eens, die zijn er maar je kan wel de harmful bias eruit halen. De handreiking zou duidelijk moeten maken dat het een iteratief proces is. Als je een relatie legt tussen de dingen die je doet kun je continu iteratief bias eruit halen. Dat is de laag die ik zie tussen ‘hoog over’ en ‘praktisch systeembouw denken’.

Bart: Spanningsveld tussen ‘hoog over’ en ‘praktisch’ bestaat, laat het spanningsveld zien tussen die twee. Het iteratieve proces medieert tussen die twee lagen. Correct?

Senay: ja, deze fluïditeit faciliteer je op deze manier.

Francien: Dat geldt ook voor de kolommen over legal, organisational en technical. Ook daartussen moeten iteratief proces tot stand komen. Veel vragen kun je niet vanuit een specialisme beantwoorden, daar heb je meerdere disciplines voor nodig. Handreiking zou denk ik meest effectief zijn als die goede processen kan geven om interactie en iteratie te laten plaatsvinden.

Bart: Eerder een soort stappenplan hoe interacties aan te gaan, dan stellen: dit zijn de principes en dat doe je zo.

Ronald: Aantal goede dingen gezegd. De opdracht gaat primair over discriminatie en hoe dat wettelijk verankerd is. Het idee om bias tegen te gaan en fairness te promoten is breder.

Bart: dat is een kernvraag: Moeten we proberen zo goed als kwaad non-discriminatie jurisprudentie vertalen naar technische realiteit OF moeten we vanuit technische kant en de biasliteratuur een document maken dat streeft naar zelfde waardes als non-discriminatie jurisprudentie.

Ronald: Als je als data scientist meer weet van recht of ethiek snap je beter wat er op het spel staat met bias, fairness en non-discriminatie. Daarvoor is training nodig. Als je een statische handreiking maakt ga je de mist in want dan gaat het ook zo behandeld en geïnterpreteerd worden, en zo sla je de plank mis. Aantal problemen: begrippen zijn lastig (bias, discriminatie, fairness), je moet ze operationaliseren maar er een checklist van maken vergt het volledig platslaan van ingewikkelde principes en dan ben je helemaal nergens. Door het stellen van vragen aan ontwikkelaars over karakteristieke dataset, over concrete risico's, over wat je aan het ontwikkelen bent, welke risico's kan je bedenken, etc, bereik je misschien meer. Welke schade kan je bedenken die kan ontstaan voor individu, maatschappij, groepen... op die manier discussie op gang krijgen om te weten welke risico's je wil beperken. Eens met Senay en Francien, het is een heen en weer gaan, multidisciplinair.

Bart: Dus dan niet zozeer een handreiking met principes misschien, maar meer met vragen voor een gesprek binnen de organisatie en met belanghebbenden.

Ronald: Ja. Kijk naar de AVG. Voor jou en mij zit daar al een duidelijk stappenplan in: (1) denk na over de noodzaak van het project, (2) welk doel heb je, (3) hoe sla je data op, (4) wie geef je toegang, etc. etc. Maar een techneut ziet/leest dat niet in de AVG. Dus klaarblijkelijk moet daar een verduidelijkings/concretiseringslag worden gemaakt. Ikea folder met tekeningetjes over kastje gaat het niet worden.

Francien: Als ik denk aan discriminatie in de juridische zin staat dat als informaticus ver van mij af. Geeft aan dat die afstand er enorm is alleen al in taal en begripsniveau. Een concept als *fairness* spreekt sterker aan aan de technische kant.

Monique: Om bij Ronald aan te haken. Eens dat het probleem breder ligt dan discriminatiegronden, in de scope van ongelijkheid. De vraag die bij mij speelt is: snapt men

de technologie niet? Of snapt men niet hoe inequity werkt? Voor je ze kan tonen hoe een systeem inequity in de hand werkt vereist eerst het begrijpen van inequity op zich. Hoe een bepaald besluit uit zo een systeem een groep ongelijk kan raken. Hoe ongelijkheid vaak al in de maatschappij zit en in hoe mensen kijken naar zaken/zelf gebiased zijn.

Bart: dat kunnen wij niet oplossen met een handreiking?

Monique: nee, maar besef van ongelijkheid in de samenleving is belangrijk, een voorwaarde om aan een systeem goede uitvoering te kunnen geven. Bewustwording over ongelijkheid is belangrijk. Interdisciplinaire samenwerking heeft zeker iets te bieden. Vergt brede en inclusieve blik, qua persoonlijke en professionele achtergrond, maar ook tussen privaat-publiek en over landsgrenzen heen. Misschien goed om soms ook advies te geven: “dit kan je nog in wetgeving vervatten?”

Bart: Op de eerste pagina van het document ‘Standaarden’ zien jullie een visualisatie van een systeembouw. Is dit goede visualisatie? Mist er iets? Is er iets te veel?

Francien: wat ik al aangaf, eerste plaatje (CRISP DM), daar zie je nog de feedback loops, maar de drie zuilen (legal, technical, organisational) zijn niet verbonden. Een heleboel problematiek speelt zich af op de interface/interactie tussen de kolommen/zuilen. Het zou goed zijn om visueel duidelijk te maken dat er interactie moet zijn. Viel me ook op in de beschrijving dat er geen pointers zijn over hoe de drie kolommen met elkaar communiceren.

Ronald: ik sluit me daar bij aan. In de praktijk proberen flow te creëren is goed, vanuit organisatieperspectief een plan hebben om juridische stappen te nemen alvorens technische uit te voeren, etc. Samengaan drie kolommen is lastig maar nodig, er zouden formele beslismomenten moeten zijn die bepalen pas naar volgende stap te kunnen gaan als bepaalde voorwaardelijke stappen zijn doorlopen. Het wordt gezien als gescheiden activiteiten maar dat is precies waar het misgaat, je moet weten wat er technisch gebeurt om te begrijpen dat er ongelijkheid is; techneuten moeten zich op hun beurt weer bewust zijn van wat juridisch nu ‘ongelijkheid’ is. Dialoog tussen techneuten en juristen is belangrijk. Belangrijk uitgangspunt: discussie over data en wat erin zit, juristen en ethici moeten dan begrip hebben van wat daar gebeurt.

Senay: Interacties zijn belangrijk maar het is ook belangrijk om toch te benadrukken dat inzichten ook ontstaan op alle drie de niveaus. Ook al gaat het over ongelijkheid, equity etc... Het kan ook zijn dat juist bij techneuten nieuwe inzichten ontstaan over ongelijkheid wat dan gecommuniceerd kan worden naar het juridische en het organisationele. Dus wisselwerking heel belangrijk.

Francien: Ik wil dit onderstrepen. Ik heb in 2017 met Linnet intersectionaliteit en discriminatie workshop georganiseerd dat bepaalde vormen van discriminatie niet wettelijk omschreven zijn. Met huidige tools zien we dat er blootstellingen zullen komen, nieuwe bestaande biases. Ook zijn techneuten in staat beperkingen en oplossingen bloot te stellen. Bijvoorbeeld de Compass-case; verschil in behandeling over het gemiddelde kun je afdwingen maar dan krijg je verschillen in de foutmarges. Zulke dingen kunnen duidelijk gemaakt worden door techneuten en vervolgens doorgegeven worden aan juristen.

Monique: Ik mis de domeinkennis; stel de overheid wil systeem bouwen over toekenning huurders. In hoeverre is de domeinkennis aanwezig bij 'legal' of 'organisational', of moeten we er niet een groep bijhalen met domeinkennis. Iets met woningen heb je woonbond nodig, enzovoort.

Francien: inderdaad, consultatie stakeholder 'throughout the process', is belangrijk voor het juiste begrip van de impact. Het doen van Impact assessment tijdens het proces met consultatie van belanghebbende, of zij die de impact ervan gaan ondervinden, om de impact te begrijpen.

Merel: we hebben naar verschillende modellen gekeken, Crisp dm op aanraden van technici. Kunnen we nog naar andere modellen kijken binnen AI? Idee van dit model was om vertaalslag te kunnen maken. Zijn er andere procesmodellen die ander perspectief bieden? Weten jullie nog een?

Ronald: semantiek: ik zie niet helemaal wat die buitenste cirkel doet... het wordt een meta evaluatie van... wat is het?

Francien: ik weet niet in hoeverre dit soort modellen actueel zijn, er zijn agile modellen die meer de chaos representeren maar ik durf er geen uitspraken over doen.

Ronald: ik ken er geen maar zijn er ongetwijfeld maar vraag me af of het zo belangrijk is omdat het in wezen gaat om de stappen.

Bart: Dan naar de inhoud van de principes; zij de principes die we hebben neergezet ten aanzien van organisatorisch, technisch en juridisch juist. Hebben we daarmee alle bestaande principes goed samengevat? Helder vormgegeven?

Monique: Blijft lastig qua plaatsing van principes in kolommen. Noodzakelijkheid is iets legal's, maar natuurlijk ook iets technisch en iets wat op organisatorisch niveau moet worden bekeken. Daarnaast spelen natuurlijk een aantal gerelateerde vragen: is het haalbaar, is het effectief, kan het zonder AI-systeem?

Ronald: Ik heb het gevoel dat het anders in elkaar zit. Je begint niet met juridische vragen maar met een probleem dat opgelost moet worden met AI systeem, dan pas ga je over andere zaken nadenken. Projectdefinitie fase start vanuit organisatie of business, afhankelijk van de context, maar daar moet probleemstelling komen: dit is het probleem en we willen het oplossen met dit systeem. Dan nadenken over hoe het te gaan vormgeven met systeemprincipes. Bijvoorbeeld huurhuistoekenning: denken we beter te doen met AI -> probleemstelling. Welke data hebben we daarvoor nodig? Welke technologie? Dan pas kan je de legal tak invliegen. Hoe komen we dan aan die data en wat mogen we ermee? Er zit meet detail workflow in, de chaos die ontstaat denk ik pas in later stadium ipv helemaal in het begin. Er zit workflow in. Stakeholder involvement: hoe gebeurt dat? Wanneer gebeurt dat? Het lijkt me ook zoals het nu staat zijn het drie bullet-lijsten, er zit geen stroom in. Maar er zit impliciet wel een volgorde in. Je kan niet benchmarken als je niet weet wat je aan het doen bent. Benchmark zit temporeel gezien aan het eind en niet bovenaan. Er zijn afhankelijkheden tussen kolommen. Interacties tussen de kolommen zouden moeten duidelijk worden.

Francien: ik heb ook de neiging om eerder het hele proces te beschrijven en labels te geven in welke dimensies dat dan plaatsvindt. Sommige zijn technisch, sommige technisch en juridisch.

Sorteren op basis van labels in plaats van in aparte folders te steken. Maar dat is ook wel uitdagend en misschien contraproductief voor de handreiking.

Bart: Wat zijn jullie gedachten over de 2^e fase? Zijn dat goede principes? Mist er iets?

Ronald: ‘Purpose’ and ‘purpose limitation’ moet in eerdere fase. Juridisch moet ‘purpose limitation’ plaatsvinden vóór data verzameling dus moet bij probleemdefinitie. Legitimatie tak moet ook voorafgaan aan data verzameling. ‘Data quality’ zou ik ook bij ‘initial data collection’ doen. Verdeling ‘data protection’-beginselen zou ik anders doen.

Bart: wat bij 2 staat naar 1 en wat bij 3 staat naar 2

Ronald: Ja, behalve ‘confidential’

Bart: wat betreft technische kant, zijn jullie bekend met andere principes?

Sennay: Nee

Ronald: begrijp niet helemaal mismatch tussen p1 en rest. Is probleemdefinitie zelfde als business understandig? Op plaatje 1 is data understanding. Wat zit er eigenlijk in die data? Dat is niet hetzelfde als ‘initial data collection and storage’. Dat levert dus verschillende trajecten op. Als je al data hebt... historische bias, selectie bias... zijn andere soorten vragen als je zelf in hand hebt als je zelf data verzamelt. Dus misschien 2 trajecten uit elkaar trekken. De vragen die je moet stellen verschillen van of je data al hebt of nog moet verzamelen.

Sennay: in praktijk zijn dat idd 2 aparte zaken, maar tegelijkertijd continue proces van bias eruit halen. Ik sluit me aan bij Ronald, als je data krijgt neem je dingen over die misschien niet eerlijk of compleet zijn. Als je het zelf doet en is continue proces ben je continue je data aan het verrijken en algoritmes aan het optimaliseren. Het verzamelen van data moet goed gedocumenteerd worden, dat is wat wij de studenten meegeven: alles wat je verzamelt documenteren; documenteer ook als er databronnen worden samengevoegd. Let ook op nieuwe trend: het algoritme naar de data toebrengen, bijvoorbeeld naar datakluisen bij burgers. Dan weet je dus niet precies hoe het zit met de data; dat geeft weer andere principes en werkwijze. Co-creëren begint ook op andere manier te leven. Tussen de drie kolommen zou ik pijltjes trekken; iteratieve mindset liever dan het lineaire. Ruimte creëren waar continue wordt nagedacht over data, keuzes en impact; inzetten op co-creatie.

Monique: Over de interactiviteit: als je kijkt naar technische kolom: staat ‘Historical bias in data? If so, which one?’ Moet dat ook niet bij de kant van legal? Welke labels ga je gebruiken? Moet dat ook niet bij legal? Of in beide kolommen? En ‘only collect data that is necessary’: ik zou ernaast zetten om aan te tonen hoe je dat hebt gedaan. Hoe heb je die noodzakelijke data verzameld?

Sennay: Er zijn thema’s die in alle drie de kolommen nodig zijn want dat zijn de discussies die interdisciplinariteit vereisen.

Bart: Dan de derde fase.

Ronald zei al ‘data quality’ beter naar stap 2. Gevoelige gegevens, heb je die nodig? Zoja, waarom? etc.... Domeinexpertise zou extra rol kunnen krijgen hier.

Francien: ik word meteen ge-alert als er objectiviteitcheck in staat. Is objectiviteit überhaupt mogelijk?

Bart: dit komt uit statistische wereld, die moeten objectieve dataset hebben. Je zou kunnen zeggen: er bestaat in de wereld niet zoiets als objectiviteit. Maar voor statistici geldt het wel als ideaal; omdat AI toch voor een groot deel statistiek is, hebben we dat ook als streefideaal opgenomen. Maar dat vind jij niet goed?

Ronald: Je moet duidelijk maken wat bedoeld wordt met objectiviteit. Dit gaat over keuze van modellen en analyse methoden, dus hoe doe je dat dan objectief? Daarnaast zit ik een beetje met: ik heb het idee dat uitwerking stappen losgezongen raakt van model eerste pagina. Voorbeeld: welke attributen zou ik willen hebben bij ontwikkelen systeem. Als je van nul begint ga je data verzamelen. Bias gaat pas komen als je de data binnenhaalt. Er zit in die eerste fase iets anders dan data schonen, er zit een iteratie tussen preparatie en ‘understanding’, en ook tussen ‘preparation’ en ‘modeling’ die je moet proberen vatten. Sommige dingen kan je niet doen als je nog geen data hebt. In die drie stappen zitten veel iteraties en verschillende vragen. Principes hoe ze er nu voor staan stellen vragen op verkeerde moment.

Sennay: Dynamiek en interactie proberen te vatten zonder dat het iets lineair moet zijn. Dat moet je dan inbedden in organisatie. Hoe pak je de niet-lineariteit aan die inherent is aan lerende systemen?

Ronald: Dat is de lastigheid van de handreiking; een circulaire realiteit platslaan tot iets tweedimensionaal. Er is een cyclus in de biases, die drie of vier cirkels bij elkaar brengen en lineair beschrijven; aangeven op welk punt inbreng of expertise nodig is. Het bij elkaar komen van disciplines op een beslismoment is essentieel.

Monique: ‘inform citizens right to provide additional data’: goed maar om welk system gaat het en welke data is er al? Moment waar burger kan inchecken en data kan verifiëren, is dat wat hier bedoeld wordt?

Sennay: je zou burgers/belanghebbenden al bij de eerste stap moeten betrekken en voortdurend moeten meenemen in de besluitvorming. Daar wordt het AI-systeem alleen maar beter en sterker van en zorgt uiteindelijk voor meer legitimiteit en draagvlak.

Monique: ik denk dat je heel duidelijk moet maken hoe je dat precies gaat doen in het proces. Duidelijk maken hoe het vorm krijgt in het proces. Dit is eigenlijk iets wat pro-actief zou moeten, waar in het proces ga je informeren en welke informatie geef je dan.

Bart: Over naar fase 4

Francien: ik zie ‘objectivity’ hier dan weer in legal terechtkomen. Dat brengt mij een beetje in de war. Is er ‘legal compliance’ nodig met statistische principes? Snap dat dit normatieve principes zijn, maar zijn ze juridisch?

Bart: beetje tussenfase, zowel in AVG als verschillende EU documenten staat dat statistische processen objectief moeten plaatsvinden. Daarnaast is er een soort Code voor statistische organisatorisch binnen de EU die dit soort principes nader uitwerkt.

Ronald: het voelt alsof er een juridische plicht ligt maar implementatie zit bij technisch of organisatorisch. In zekere zin is dat met AVG-beginselen ook zo, maar voelt alsof betekenis juridische box niet bij juridische kolom zit maar bij de organisatorische of technische.

Francien: Ik zou deze principes eerder in organisatorisch of technisch zien

Bart: wat vind je van de principes zelf?

Francien: daar ben ik het mee eens maar vraag me af waarom in juridische kolom? Omdat het vooral in het bouwen van een systeem zijn het niet de juristen die het gaan verwezenlijken zoals Ronald zei. Het is een check of deze dingen door andere gedaan zijn. Het is niet de taak van juristen 'to develop statistics'.

Ronald: voelt alsof je 'requirements' op abstracte manier formuleert, maar je kan ze in verschillende kolommen neerzetten. De organisatie of business moet bepaalde modellen en technieken kiezen en implementeren, dus de voorwaarde in 'statistical principles' kan je aangeven maar bron van de verplichting heeft juridische grondslag. Dat geldt eigenlijk voor alle dingen... Wie moet dat waarborgen en wie is er verantwoordelijk voor? Juristen moeten check of het gebeurt maar heeft verder geen betekenis voor juristen.

Bart: Stap 5

Francien: 'right to discriminate' moet zijn 'right to non-discriminatie'

Ronald: of recht op individuele behandeling in plaats van recht op non-discriminatie.

Francien: goed om te zien 'use different fairness metric'. Staat niet wat dan bepaalt of een test for fairness geslaagd is. Zou goed zijn als iedereen dat doet want je kan niet goed scoren op alles tegelijk. Maakt je dan bewust van punten waarop je niet goed scoort en kan dan kijken of het problematisch is. Uitleg over wat het betekent is hier wel wenselijk.

Ronald: ontbreekt iets in de richting van 'accuracy'. 'Evaluatie van model', hoort daar bij dat het model doet wat het beoogt te doen? Lijkt me wel een rol te spelen in dit stadium.

Francien: bron van bias in algoritme systemen zit vaak in hoe systeem geïnterpreteerd wordt. Zit vaak in 'attributed meaning' van de uitkomst van het systeem. 'Accuracy' of 'check dat systeem doet wat het moet doen' en kritische reflectie daarop is belangrijk in acht te nemen. Het begrijpen van de uitkomst van het proces valt bij mij ook onder 'evaluatie'.

Bart: De deployment fase

Sennay: interactie met degene waarop je het toepast. Als je de burger erbij haalt is dat hier de plek.

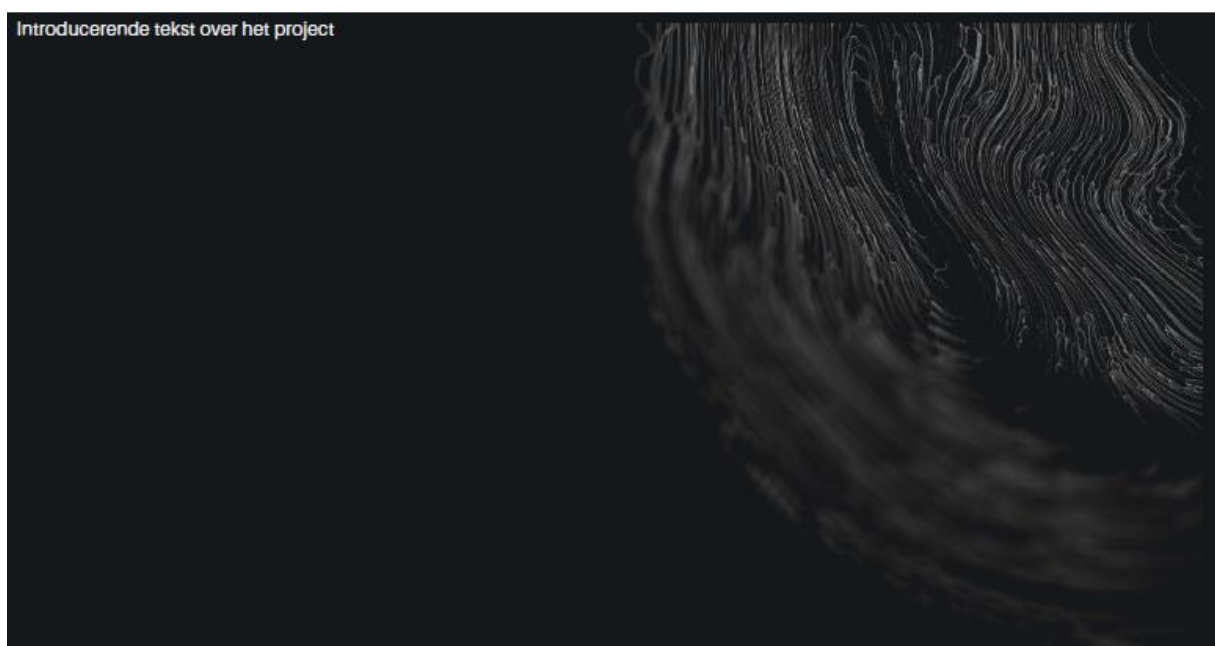
Ronald: omgekeerd, artikel 22 GDPR is wat laat in deze fase, dat weet je bij stap 1 ook al dus mag naar voren wat mij betreft.

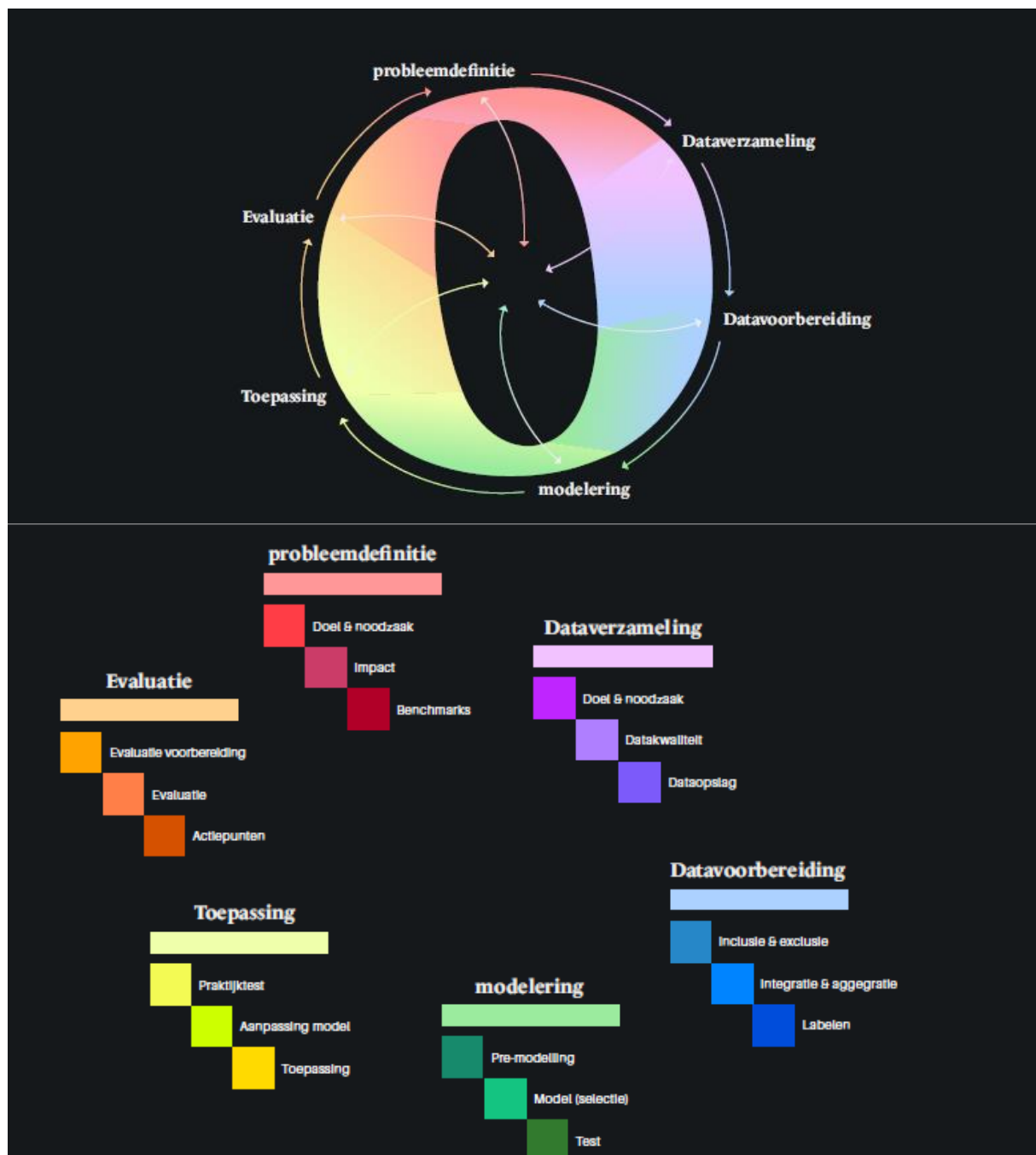
Monique: als het misgaat dat je zelf nog met systeem kan communiceren of onjuistheden kan corrigeren. Is dat ook in deze stap?

Ronald: lijkt te ontbreken maar hoort hier wel. ‘Right to correction’ enzovoort is hier nodig. Bepaalde ‘data subject rights’ kunnen pas uitgerold worden later. Correctierecht speelt nu pas, maar artikel 12 – 15 moet je uitsmeren over de cyclus. ‘Data Subjects rights’ komen hier pas aan bod. Ook bezwaar tegen beslissingen. Wordt een erg ingewikkeld model.

Bart: bedankt voor nuttige toevoegingen, zeer interessant. Blijkt nog complexer en iteratiever dan we dachten.

7.6 Eerste design Handreiking







Inhoudsopgave

3	CRISP DM systeem	25	2. Dataverzameling
9	Introductie	26	Doel en noodzaak
11	Voorbeelden toegestane besluiten		Data kwaliteit
	Voorbeelden verboden besluiten		Data opslag
12	Inleiding		Voorbeelden
		27	Arbeidsmarkt
13	Handreiking	28	Strafrechtketen
14	Diversiteit	29	Medisch
15	Praktijk		Principes
16	Controleerbaarheid	30	Juridisch
		31	Technisch
17	1. Probleemdefinitie	33	Organisatorisch
18	Doel en noodzaak		
	Impact	34	3. Datavoorbereiding
	Benchmarks	35	Inclusie en exclusie
	Voorbeelden		Integratie en aggregatie
19	Arbeidsmarkt		Labelen
20	Strafrechtketen		Voorbeelden
21	Medisch	36	Arbeidsmarkt
	Principes	37	Strafrechtketen
22	Juridisch	38	Medisch
23	Technisch		Principes
24	Organisatorisch	39	Juridisch
		40	Technisch
		41	Organisatorisch

42	4. Modelering	59	6. Evaluatie
43	Pre-modelering	60	Evaluatievoorbereiding
	Model (selectie)		Evaluatie
	Test		Actiepunten
	Voorbeelden		Voorbeelden
44	Arbeidsmarkt	61	Arbeidsmarkt
45	Strafrechtketen	62	Strafrechtketen
46	Medisch	63	Medisch
	Principes		Principes
47	Juridisch	64	Juridisch
48	Technisch		Technisch
51	Organisatorisch	65	Organisatorisch
52	5. Toepassing		
53	Praktijktest		
	Aanpassing model		
	Toepassing		
	Voorbeelden		
54	Arbeidsmarkt		
55	Strafrechtketen		
56	Medisch		
	Principes		
57	Juridisch		
	Technisch		
58	Organisatorisch		

fase - 0

Introductie

Artikel 1 van de Nederlandse Grondwet verbiedt discriminatie: 'Allen die zich in Nederland bevinden, worden in gelijke gevallen gelijk behandeld. Discriminatie wegens godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht of op welke grond dan ook, is niet toegestaan.' Non-discriminatie is dus de basis van ons rechtssysteem en onze samenleving.

Dit document is bedoeld voor systeembouwers, data analisten en AI-experts en legt uit welke vragen en principes leidend zijn bij het ontwikkelen en het implementeren van een AI-systeem met het oog op het discriminatieverbod. Stel je wilt een AI-systeem zo non-discriminatoir mogelijk maken, waar moet je dan aan denken?

Deze handreiking geeft uitgangspunten en vragen en geen ge- of verboden. Vaak is de hoop dat juristen 'gewoon zeggen' wat wel en wat niet mag, maar zo werkt het recht niet. Het recht geeft een standaard, een uitgangspunt of een principe, maar daarop bestaan altijd uitzonderingen. Bovendien zal een rechter altijd rekening houden met de context, of wat juristen noemen 'de omstandigheden van het geval'.

Dat geldt ook voor het non-discriminatierecht. Onderscheid maken op basis van geslacht is bijvoorbeeld verboden, tenzij dit een relevante factor is. Zoekt een castingbureau een actrice om de vrouwelijke hoofdrol op zich te nemen, dan mag het natuurlijk mannen uitsluiten. Sterker nog, positieve discriminatie kan in sommige gevallen geoorloofd zijn; als een organisatie bijvoorbeeld met name mannelijke werknemers heeft, dan mag die besluiten dat bij gelijke geschiktheid aan vrouwen de voorkeur wordt gegeven.

Ook de vraag wat nu precies 'onderscheid maken' is, is niet eenduidig te beantwoorden. Het uitgangspunt in het recht is dat gelijke gevallen gelijk behandeld moeten worden,

Introductie

fase - 0

en ongelijke gevallen ongelijk. Maar wat is precies gelijk en wat niet? Welke factoren zijn relevant voor het besluit en welke niet? Al dat soort vragen kunnen niet in het algemeen worden beantwoord in deze handreiking, omdat ze afhangen van de context en de vraag hoe een AI-systeem functioneert, waartoe het dient en welke waarborgen er zijn getroffen. Belangrijk is dat het recht niet alleen uitgaat van een beperkt aantal gronden waarop in principe geen besluiten mogen worden genomen, zoals ras, geslacht of geaardheid, maar dat het ook discriminatie 'op welke grond ook' verbiedt. Dit vergt van een systeembouwer dat die zich goed bewust is op basis van welke groepen het AI-systeem onderscheid maakt en of dat onderscheid te rechtvaardigen is. Is het onderscheid bedoeld of onbedoeld; is het relevant of niet?

Daarnaast is het van belang dat niet alleen 'directe discriminatie' verboden is, maar ook 'indirecte discriminatie'. Als de politie bijvoorbeeld in het kader van een project vooral personenauto's met een kenteken uit midden of Oost-Europa controleert, dan is dat verboden omdat de eigenaars van die auto's veelal personen uit midden of Oost-Europa zijn. Zij worden dus meer getroffen door deze maatregel dan personen met een andere nationaliteit. Zo ook kan bijvoorbeeld een postcodegebied een proxy zijn voor ras en etniciteit.

9

Introductie

fase - 0

Voorbeelden van toegestane besluiten

Een medische organisatie selecteert alleen mensen met een bepaalde etnische achtergrond voor het testen van een medicijn, omdat er aanwijzingen zijn dat er binnen die groep grotere en ernstigere bijwerkingen zijn.

10 De belastingdienst besluit Thaise-massage-salons aan een nadere check te onderwerpen, omdat er aanwijzingen zijn dat die worden gebruikt voor het witwassen van drugsgelden.

Voorbeelden van verboden besluiten

Een medische organisatie weigert orthodox gereformeerden te behandelen, omdat die zich niet vaccineren tegen ziekten en artsen geloven dat ze het dan aan zichzelf te wijten hebben als ze ziek worden.

De belastingdienst besluit een groep nader te controleren en doet dat niet op basis van etniciteit, maar op basis van de vraag of iemand een tweede huis heeft, waarbij 90% van die groep een immigratieachtergrond heeft.



Introductie

11



12

Omdat het discriminatierecht zelf weinig ex ante regels bevat, wordt in deze handreiking naast dit rechtsgebied ook gekeken naar het privacy- en gegevensbeschermingsrecht, statistische principes, organisatorische handvaten en technologische best practices. Samen bij elkaar vormt dat een pakket aan principes voor het bouwen van een non-discriminair AI-systeem. De relevante vragen en uitgangspunten worden in deze handreiking weergegeven aan de hand van fasen die zijn gebaseerd op het Cross-industrie standaardproces voor datamining (CRISP-DM). Per stap zullen steeds eerst de vragen worden geclusterd.

Introductie

13	fase - 0 daarna zullen ze ter illustratie kort worden uitgewerkt aan de hand van drie hypothetische casus, waarbij niet alle, maar een aantal vragen worden uitgewerkt, en tot slot zullen concrete juridische, organisatorische en technische handvatten worden gegeven. Drie uitgangspunten zijn leidend voor het hele proces en zullen in alle fasen in acht moeten worden genomen. 1. Diversiteit De meeste vragen die in dit document staan kunnen niet in isolement worden beantwoord. Het document probeert een brug te slaan tussen de juridische wereld en de technische wereld. Terwijl juristen gewend zijn aan hele algemene uitgangspunten zoals 'wees transparant' en 'vermijd discriminatie' moeten bij het bouwen van een AI-systeem concrete keuzes worden gemaakt. Welk deel van het proces moet bijvoorbeeld transparant zijn en voor wie moet het proces begrijpelijk zijn: de systeembouwer zelf, een rechter of voor een leek? Hetzelfde geldt voor discriminatie: de keuze om de ene bias te vermijden betekent vaak dat een andere bias de kop op steekt. Het is daarom belangrijk om het team dat aan een AI-project werkt zo divers mogelijk samen te stellen. Dat geldt voor de verschillende expertises en professionele achtergronden van mensen, maar ook voor de persoonlijke achtergrond binnen het team. Let daarbij op etniciteit, geslacht, geaardheid, culturele en religieuze achtergrond, leeftijd en andere aspecten die bij de concrete toepassingsfunctionaliteit van het AI-systeem relevant kunnen zijn. 2. Contextgevoeligheid AI-systemen hebben uiteindelijk invloed op de praktijk, vaak binnen een specifieke context. Daarom is het belangrijk om al bij het ontwerp van het systeem specifieke domeinkennis Introductie
14	fase - 0 binnen het team te hebben. Vaak worden AI-systemen gebouwd in isolatie, met een mismatch met de realiteit ten gevolg. Dat heeft niet alleen nadelige effecten op de effectiviteit, maar kan ook ongerechtvaardigde discriminatie in de hand werken, bijvoorbeeld omdat de bouwers van een AI-systeem niet hebben gezien dat een bepaald datapunt een proxy is voor een van de gronden die het anti-discriminatierecht een rol spelen. Daarnaast is het belangrijk om in een vroegtijdig stadium belanghebbenden bij het ontwerp te betrekken. Stel dat een AI-systeem met 80% accuraatheid een bepaalde diagnose kan stellen, terwijl artsen dat slechts met 60% kunnen, maar het nadeel van het systeem is wel dat het is getraind op data met betrekking tot mannen en dus een veel hogere foutmarge heeft ten aanzien van vrouwen. Dat betekent niet dat het per definitie verboden is om zo'n systeem in te voeren, maar wel dat het belangrijk is om al in een vroegtijdig stadium in gesprek te gaan met patiëntenorganisaties over de inrichting van het ontwikkel- en besluitvormingsproces en het mitigeren van potentiële problemen. Kunnen er alsnog data worden verzameld die het beeld compleet maken? Moet het AI-systeem alleen worden toegepast ten aanzien van diagnoses bij mannen? Kan een deel van de besparing in kosten en middelen worden ingezet voor extra capaciteit bij het diagnosticeren van vrouwen zonder een AI-systeem? Etc. 3. Controleerbaarheid Het proces en de stappen die worden gezet moeten inzichtelijk, systematisch en controleerbaar zijn. Daarom is het belangrijk dat alle keuzes goed worden gedocumenteerd en verantwoord. Dit zorgt er voor dat achteraf kan worden gecontroleerd of er fouten zijn gemaakt en het systeem later beter kan worden geüpdatet. Het uitgangspunt is daarnaast dat onderdelen in het proces zo concreet mogelijk worden gedocumenteerd, Introductie

fase - 0

zodat alle stappen herhaalbaar en verifieerbaar zijn.

Belangrijk is om te zorgen voor interne en externe controle op de procesonderdelen. Dat kan bijvoorbeeld door een second opinion te vragen van externe experts, door testen te doen op dezelfde data met een ander algoritme of door het proces twee keer te doorlopen met een andere fairness definitie. Belangrijk is daarbij dat het geheel een iteratief proces is: het AI-systeem is een lerend systeem en is voortdurend in ontwikkeling en dus moet er constant worden gemonitord of aan alle voorwaarden en non-discriminatie principes is voldaan. Het doorlopen van deze handreiking is dus geen garantie voor het volledig uitbannen van discriminatie bij AI-systemen. Ook is het nadrukkelijk niet bedoeld als een lijst met vragen die simpelweg afgevinkt kunnen worden. Het is opgesteld om op een gesystematiseerde en bewuste wijze keuzes te maken over hoe non-discriminatie principes in AI processen kunnen worden ingebed.

15

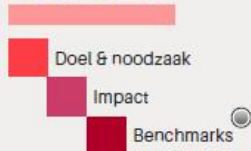
Leeswijzer

Tot slot. Deze handreiking is geschreven voor de projectleider van een AI-systeem en heeft ten doel de projectleider de juiste fasen te laten onderscheiden, de juiste mensen op juiste momenten bij elkaar te zetten en hen de juiste vragen te laten stellen. Uiteraard moet worden bedacht dat besluitvormingsprocessen die geen gebruik maken van AI nu ook vaak gebiased zijn, maar dat er weinig kennis en data over dit soort beoordeelbaarheid is. Het werken met AI is dan ook niet per se een risico in termen van discriminatie, het is juist een kans om processen neutraler en rechtvaardiger te maken en beter in beeld te krijgen welke groepen worden getroffen door bepaald handelen, beleid of besluiten.

Introductie

fase - 1

Probleemdefinitie



Belangrijke punten in deze fase

Wat zijn de financiële, computationele en organisatorische kosten voor dit systeem en welke kosten zouden er zijn als er voor een niet-AI gedreven oplossing zou worden gekozen?

Welke impact hebben fout-positieven en fout-negatieven op de burger en de maatschappij?

Wat is het doel van het project? Is het noodzakelijk om met AI te werken of kan het probleem ook zonder een AI-systeem worden geadresseerd?

Note: deze punten zijn nader te bepalen.

Doel en noodzaak

1. Wat is het probleem en hoe gaat Ai helpen het probleem op te lossen?
2. Wat is het doel van het project? Is het noodzakelijk om met Ai te werken of kan het probleem ook zonder een Ai-systeem worden geadresseerd?
3. Welke veronderstellingen over het onderscheid tussen verschillende groepen liggen ten grondslag aan de formulering van het doel van het systeem?
4. Bestaan er verschillende opvattingen over het doel van het systeem en zijn de verschillende belanghebbenden daarin gehoord?
5. Wat is het probleem en volgens wie en waarom moet daar iets aan gebeuren?
6. Welke groepen worden onderscheiden in de probleemdefinitie(s) en waarom?
7. Welke verandering en voor wie moet het systeem teweegbrengen en waarom?

17

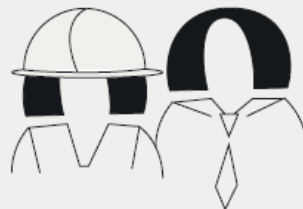
Impact

8. Moeten er voor dit project meer data worden verzameld en verwerkt dan reeds beschikbaar zijn binnen de organisatie en welke gevolgen heeft dat voor burgers?
9. Welke impact heeft het systeem op burgers en de maatschappij ten positieve en ten negatieve?
10. Wordt het systeem gebruikt om informatie te verkrijgen, om besluiten voor te bereiden of om zelfstandige besluiten te nemen en welke gevolgen heeft dat voor de mate waarin Ai bepalend zal zijn in de praktijk?
11. Welke impact hebben fout-positieven en fout-negatieven op de burger en de maatschappij?
12. Welke procedures zijn ingebouwd voor belanghebbenden om een beslissing (fout-negatieven/foutpositieven) aan te vechten?

Benchmarks

13. Wat zijn de financiële, computationele en organisatorische kosten voor dit systeem en welke kosten zouden er zijn als er voor een niet-Ai gedreven oplossing zou worden gekozen?
14. Wanneer is het Ai-systeem een succes, bijvoorbeeld bij welk percentage van effectiviteit, en wanneer moet deze benchmark zijn gehaald, bijvoorbeeld na 1 maand of 2 jaar?
15. Welk percentage in foutnegatieven en foutpositieven is acceptabel?
16. Wat betekenen de verschillende succes-criteria voor verschillende groepen?

Probleemdefinitie

voorbeelden Ai-inzet**Arbeidsmarkt**

fase - 1

Het probleem is dat het manueel beoordelen van brieven veel tijd kost en manuele selectie ge-biased is (human bias). Dit systeem dient om een voorselectie te maken van sollicitatiebrieven. Dit systeem heeft ten doel het proces efficiënter en objectiever te maken.

18

Voor de training van dit systeem wordt gebruik gemaakt van reeds binnengekomen sollicitatiebrieven. Het systeem maakt zelf beslissingen. Foutpositieven betekent een minder efficiënt systeem, foutnegatieven betekent geen baankans voor een sollicitant.

Succes bij 20% kostenreductie en 5% minder onterechte afwijzing t.o.v. het bestaande proces. Moet zich uiterlijk na 1 jaar hebben bewezen. Maximaal fout positief: 60% (t.o.v. de mensen die worden uitgenodigd voor een gesprek) en fout negatief: 2% (t.o.v. de mensen die zouden worden uitgenodigd voor een gesprek door manuele selectie)

Probleemdefinitie

voorbeelden Ai-inzet

fase - 1



Strafrechtketen

19

Het probleem is dat de politie nu vaak pas na de feiten ter plaatse is. Dit systeem dient om te voorspellen waar en wanneer een overtreding zal plaatsvinden en aldus preventief en gericht patrouilles in te zetten. Dit systeem heeft ten doel het proces kosten-efficiënter, nauwkeuriger en objectiever te maken.

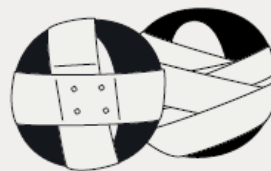
Gebruik van reeds bestaande databases, aangevuld met gescrepte social media data. Het systeem maakt voorspellingen en informeert de corpsleiding. Foutpositieven betekent efficiëntieverlies, foutnegatieven betekent dat overtreding zal plaatsvinden.

Succes bij 10% hogere effectiviteit van politie-surveillances, gemeten aan het aantal arrestaties en staande houdingen. Moet zich na 3 jaar hebben bewezen voor een gesprek door manuele selectie). Gedurende de 3 jaar worden surveillances uitgevoerd op basis van het bestaande proces en anderen met behulp van het AI-systeem. Die laatste mogen geen hogere foutpositieven of -negatieven opleveren.

Probleemdefinitie

voorbeelden Ai-inzet

fase - 1



Medische sector

20

Het probleem is dat het manueel beoordelen van brieven veel tijd kost en manuele selectie ge-biased is (human bias). Dit systeem dient om een voorselectie te maken van sollicitatiebrieven. Dit systeem heeft ten doel het proces efficiënter en objectiever te maken.

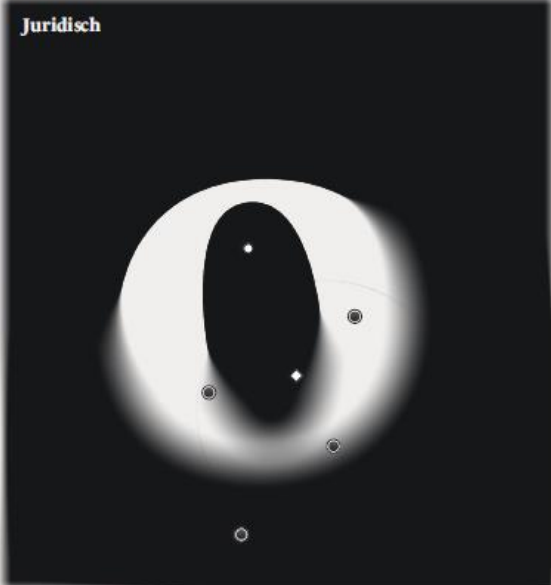
Data worden verzameld gedurende de trial-fase van dit systeem in verschillende instituten over de wereld. Het systeem maakt zelf beslissingen bij de duidelijke gevallen en geeft de keuze aan een arts bij een hoge onzekerheidsmarge. Foutpositieven leiden tot angst, foutnegatieven leiden in extremo tot de dood.

Succes bij 10% hogere precisie van de diagnoses, gemeten ten opzichte van de fout-negatieven die nu bestaan. Kosten mogen niet hoger zijn dan het nu bestaande proces. Er zit geen einddatum aan dit proces, maar bij de constatering van een hogere foutmarge t.a.v. foutnegatieven wordt het project onmiddellijk stopgezet. Artsen blijven diagnoses uitvoeren zoals nu, daarnaast doet een AI-systeem diagnoses. Een onafhankelijke arts vergelijkt en controleert de resultaten.

Probleemdefinitie

21

Juridisch



Probleemdefinitie

fase - 1

- ❶ Er dient vooraf een helder en concreet doel voor het systeem te worden gedefinieerd.
- ❷ Ga na of er persoonsgegevens worden verwerkt, dat wil zeggen gegevens die kunnen worden herleid tot concrete individuen. [Klik hier voor uitleg.](#)
- ❸ Er dient een Impact Assessment te worden uitgevoerd, waarin wordt gekeken naar de impact van het systeem op alle mensen-rechten en hoe die gemitigeerd worden. Als er hoge risico's blijven bestaan na aanvullende maatregelen dan moet voorafgaande toestemming van de Autoriteit Persoonsgegevens worden gevraagd. [Klik hier voor een model.](#)
- ❹ Juridisch gezien moeten systemen noodzakelijk, proportioneel en subsidiair zijn. Dat betekent dat de nadelen van zo'n systeem niet groter mogen zijn dan de voordelen en dat er geen andere, minder ingrijpende manieren zijn om hetzelfde doel te verwezenlijken. [Klik hier voor uitleg.](#)

22

Technisch

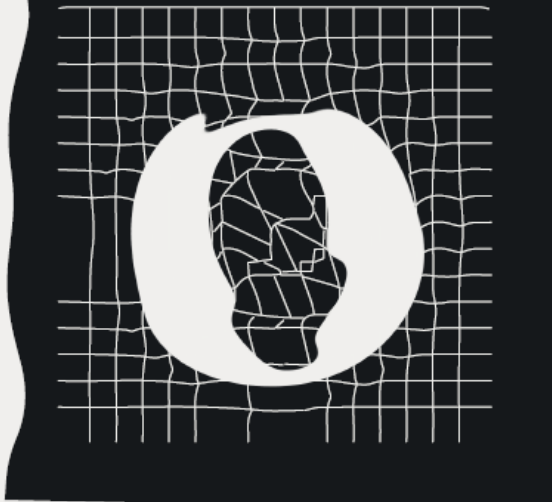


Probleemdefinitie

fase - 1

- ❶ Beargumenteer en documenteer de keuze voor AI in relatie tot context: b.v. keuze voor doelvariabele, classificatietaak, prestatiedoelen, etc.
- ❷ Formuleer en documenteer de logica en het waarom van het AI systeem.
- ❸ Ga na en documenteer wat de beperkingen zijn van de verschillende aanpakken met betrekking tot de representatie van het probleem en de belanghebbenden.
- ❹ Beargumenteer de keuze voor een type AI-systeem: is het een eenvoudige beslissboom, een zelflerend systeem of deep learning?
- ❺ Waartoe wordt het systeem ingezet: het verkrijgen van inzicht, het voorbereiden van besluiten of het nemen van zelfstandige besluiten? [Klik hier voor uitleg.](#)
- ❻ Bepaal de benchmarks ten aanzien van foutpercentages en -negatieven, de effectiviteit en de looptijd van het project.
- ❼ Voer een technische evaluatie uit. Een voorbeeld is de Ethics Canvas als ontwikkeld door het Open Data Institute. [Klik hier voor de Canvas.](#)

Organisatorisch



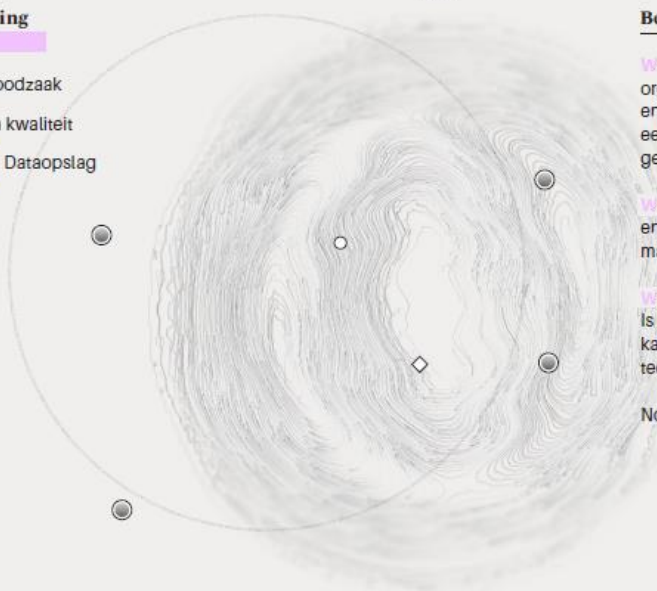
fase - 1

- 1 Loop de volgende vragen langs:
 - Heeft het team de toegang tot de benodigde middelen?
 - Heeft het team de benodigde bevoegdheden?
 - Is alle relevante expertise aanwezig in het team?
 - Is het team zo divers mogelijk?
 - Is een persoon met domeinkennis betrokken bij het project?
 - Wie is er voor welk onderdeel van het proces verantwoordelijk en hoe vindt de interne communicatie en besluitvorming plaats?
- 2 Breng in kaart binnen welk domein het systeem een rol zal gaan spelen en wat de maatschappelijke context is.
- 3 Breng in kaart wie de belanghebbenden zijn en bij wie de voordelen van het systeem en bij wie de nadelen komen te liggen.
- 4 Betrek de relevante belanghebbenden of vertegenwoordigende (burgerrechten) organisaties in een vroegtijdig stadium.
- 5 Documenteer alle stappen en keuzes in het proces en overleg die keuzes zowel binnen het team als met belanghebbenden.
- 6 Bij twijfel, contacteer een externe expert voor een second opinion.
- 7 Leg vast wanneer het proces zal worden stopgezet en wat de exit strategie is.

Probleemdefinitie

fase - 2

Dataverzameling



Belangrijke punten in deze fase

Wat zijn de financiële, computationele en organisatorische kosten voor dit systeem en welke kosten zouden er zijn als er voor een niet-AI gedreven oplossing zou worden gekozen?

Welke impact hebben fout-positieven en fout-negatieven op de burger en de maatschappij?

Wat is het doel van het project? Is het noodzakelijk om met AI te werken of kan het probleem ook zonder een AI-systeem worden geadresseerd?

Note: deze punten zijn nader te bepalen.

Doel en noodzaak 1. Welke data zijn nodig voor dit project en waarom? 2. In hoeverre zijn deze gegevens al binnen de organisatie beschikbaar en in hoeverre moeten ze van buiten worden gehaald? 3. Is het toegestaan om deze data voor dit project te verzamelen en te verwerken?	Datakwaliteit 4. Welke bias zit er in de data (van binnen, buiten of gecombineerd) en welke consequenties heeft dat? 5. Zijn de data representatief en zijn alle relevante groepen in gelijke mate vertegenwoordigd? 6. Als verschillende databronnen worden gebruikt, hoe wordt er gezorgd dat deze data compatibel en vergelijkbaar zijn?	Dataopslag 7. Hoe lang worden de gegevens bewaard en hoe? 8. Worden de gegevens veilig en vertrouwelijk behandeld; welke gevolgen heeft een datalek voor groepen of categorieën personen? 9. Worden data gedeeld met andere partijen en wat is het gevaar dat die misbruik maken van de data met negatieve gevolgen voor groepen of categorieën personen?
25 voorbeelden Ai-inzet	Arbeidsmarkt 	
26 Data van 30 random vroegere sollicitatieprocedures zullen worden beoordeeld op basis van daadwerkelijke uitkomst en de uitkomst als gesuggereerd door het AI-systeem. Data is afkomstig van de organisatie zelf. Het gaat hier om secundair gebruik van data dat is geoorloofd vanwege het publieke belang van het voorkomen van gebiasde sollicitatieprocedures.	Er zit een grote bias in de bestaande dataset, onder meer op basis van etniciteit en afkomst. De bias in bestaande data is de reden om dit systeem te introduceren. Succes van het AI-systeem zal dus niet worden beoordeeld op basis van de mate waarin de uitkomsten gelijk zijn aan vroegere processen. Een onafhankelijk team zal beoordelen welke van de selectiemethoden de meest rechtvaardige uitkomst oplevert.	Data worden slechts voor de duur van dit project bewaard en daarna verwijderd. Gedurende het project staan de data op een interne cloud Deze is slechts toegankelijk is voor medewerkers van de afdeling werving en selectie en voor de ontwikkelaars van het AI-systeem.

voorbeelden Ai-inzet

Basc - 2



Strafrechtketen

27

Data over effectiviteit van vroegere surveillance en criminaliteitscijfers (beschikbaar binnen de organisatie). Deze data worden gekoppeld aan open data afkomstig van sociale media. De data worden geanonimiseerd en geaggregeerd op groepsniveau. De verwerking vindt plaats op basis van statistische data en niet op persoonsgegevens.

Er zit een historische bias in de dataset van de politie, onder meer ten aanzien van bepaalde wijken, personen met een immigratieachtergrond, lage sociaaleconomische status en mannen. Er zit een bias in de social media die vooral door jongeren worden gebruikt. Deze twee databronnen worden onafhankelijk van elkaar beoordeeld en niet direct gekoppeld. Er zal worden gecorrigeerd op deze bias.

Data worden in geanonimiseerde vorm oneindig bewaard omdat ook historische patronen van grote waarde zijn voor toekomstige processen.

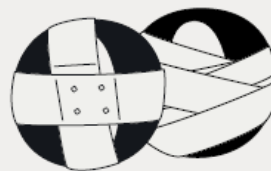
De data worden binnen een beveiligde ruimte opgeslagen.

Slechts leden van een speciale unit hebben toegang tot deze data.

Dataverzameling

voorbeelden Ai-inzet

Basc - 2



Medische sector

28

Er zijn zo veel mogelijk data nodig over diagnoses gesteld door artsen en second opinions door het AI-systeem. Hoe presteert het AI-systeem ten opzichte van een arts?

Er wordt slechts gewerkt met data die patiënten zelf afgeven.

Hiervoor geven zij geïnformeerde toestemming.

Er zal een bias zitten in deze patiëntpopulatie, namelijk vrouwen van 50+ met overgewicht. Het AI zal dus leren om met name ten aanzien van deze groepen accurate voorspellingen te doen.

Dat kan betekenen dat op termijn, het AI-systeem wordt ingezet voor diagnoses bij deze groep en artsen voor andere groepen, al naar gelang de accuratesse daartoe aanleiding geeft.

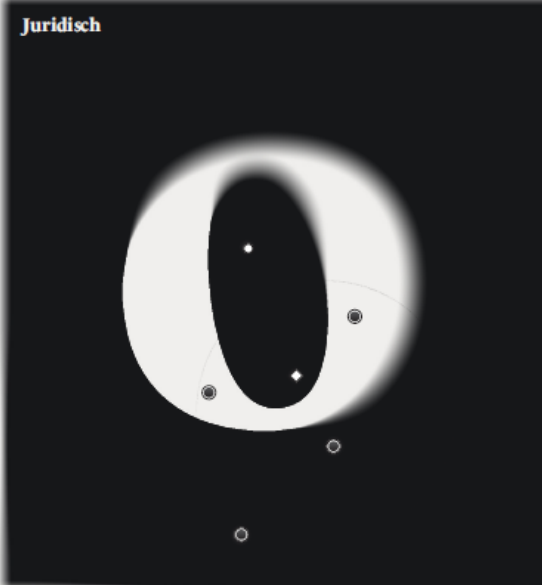
Data worden voor 20 jaar bewaard, om kwaliteitscontroles en assessments mogelijk te maken.

Alleen onderzoekers en data-analysten hebben toegang tot de data. De data van de verschillende instituten worden aan elkaar gekoppeld en worden gedeeld in een secured private cloud.

Dataverzameling

29

Juridisch



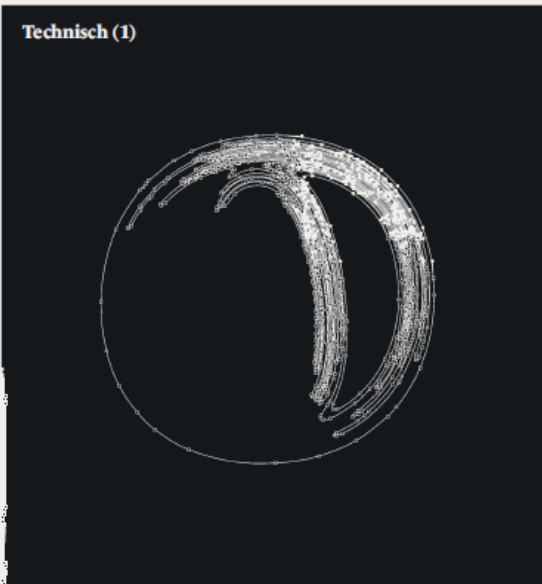
fase - 2

- ❶ Beoordeel of er een legitieme grondslag is voor het verwerken van persoonsgegevens, zoals een wettelijke plicht, informed consent of publiek belang. [Klik hier voor meer uitleg.](#)
- ❶ Als er gegevens over iemands ras, geaardheid, gezondheid, strafrechtelijk verleden of andere gevoelige gegevens worden verwerkt, dan is expliciete toestemming, een wettelijke grondslag of een zwaarwegend publiek belang nodig. [Klik hier voor uitleg.](#)
- ❶ Als er gegevens uit andere bron(nen) worden verkregen, dan moet worden geverifieerd of die data op een legitieme wijze zijn verkregen.
- ❶ Data moeten correct zijn, up to date worden gehouden en betrokkenen moet de mogelijkheid worden geboden om aanvullende gegevens aan te dragen. [Klik hier voor een voorbeeldsbrochure.](#)
- ❶ Als gegevens de EU-grens overgaan, en bijvoorbeeld naar de VS of China worden doorgevoerd, moet de organisatie waarmee de gegevens worden gedeeld zich aan de in de Europese Unie geldende gegevens-beschermingsregels houden. [Klik hier voor nadere uitleg.](#)

Dataverzameling

30

Technisch (1)

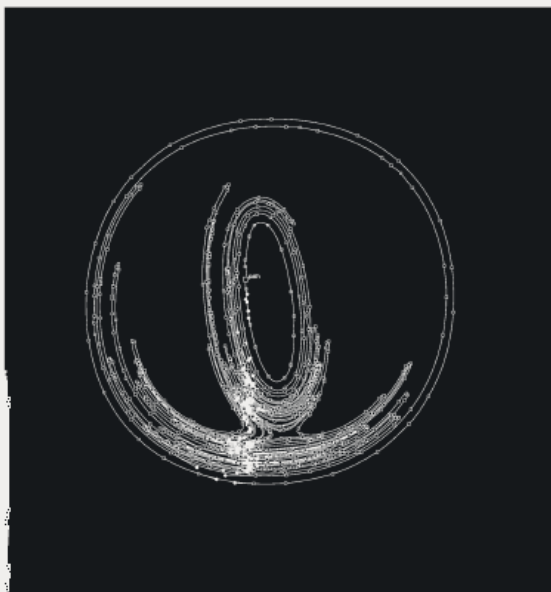


fase - 2

- ❶ Zorg dat data zo worden verzameld dat ze kunnen worden benut door het AI-systeem.
- ❶ Leg daarbij van te voren technische vereisten vast.
- ❶ Beoordeel data op:
 - Verdeling van attributen (bijvoorbeeld het doelattribuut van een voorspellingstaak)
 - Relaties tussen paren of een klein aantal attributen
 - Resultaten van eenvoudige aggregaties
 - Eigenschappen van significante subpopulaties
- ❶ Beoordeel data op mogelijk aanwezige bias aan de hand van o.a.:
 - Verdeling van de doelvariabele in verschillende groepen.
Voorbeeld: Wanneer de proportie 'positieven' voor vrouwen anders is dan voor mannen, kan dit een teken zijn van historische bias.
 - Relaties tussen paren of een klein aantal attributen (features):
 leidt dit tot proxies?
Voorbeeld: postcode kan een proxy zijn voor etniciteit.
 - Verdeling van attributen en representatie van relevante subpopulaties.
Voorbeeld: wanneer een training dataset van gezichtsherkenningsssoftware weinig (i.e., lage representatie) foto's bevat van donkergetinte gezichten (i.e., andere verdeling van attributen) bevat, bestaat het risico dat het systeem slechter zal werken voor deze groepen.

Dataverzameling

31



fase - 2

- ❶ **Controleer data nogmaals op bias:**
 - Historische bias: ongelijkheid in de samenleving is vaak ook terug te zien in op zich neutraal verzamelde data
 - Representatie bias: ontstaat wanneer delen van de inputruimte zijn onder- of overgerepresenteerd
 - Measurement bias: meten we wel wat we willen meten? Zijn verkoopcijfers bijvoorbeeld wel een goede proxy voor de verkoopkwaliteiten van werknemers?
- ❷ **Zorg voor de technische veiligheid en vertrouwelijkheid van de data.** Zorg dat de data geëncrypteerd en gecompartmentaliseerd zijn.

Dataverzameling

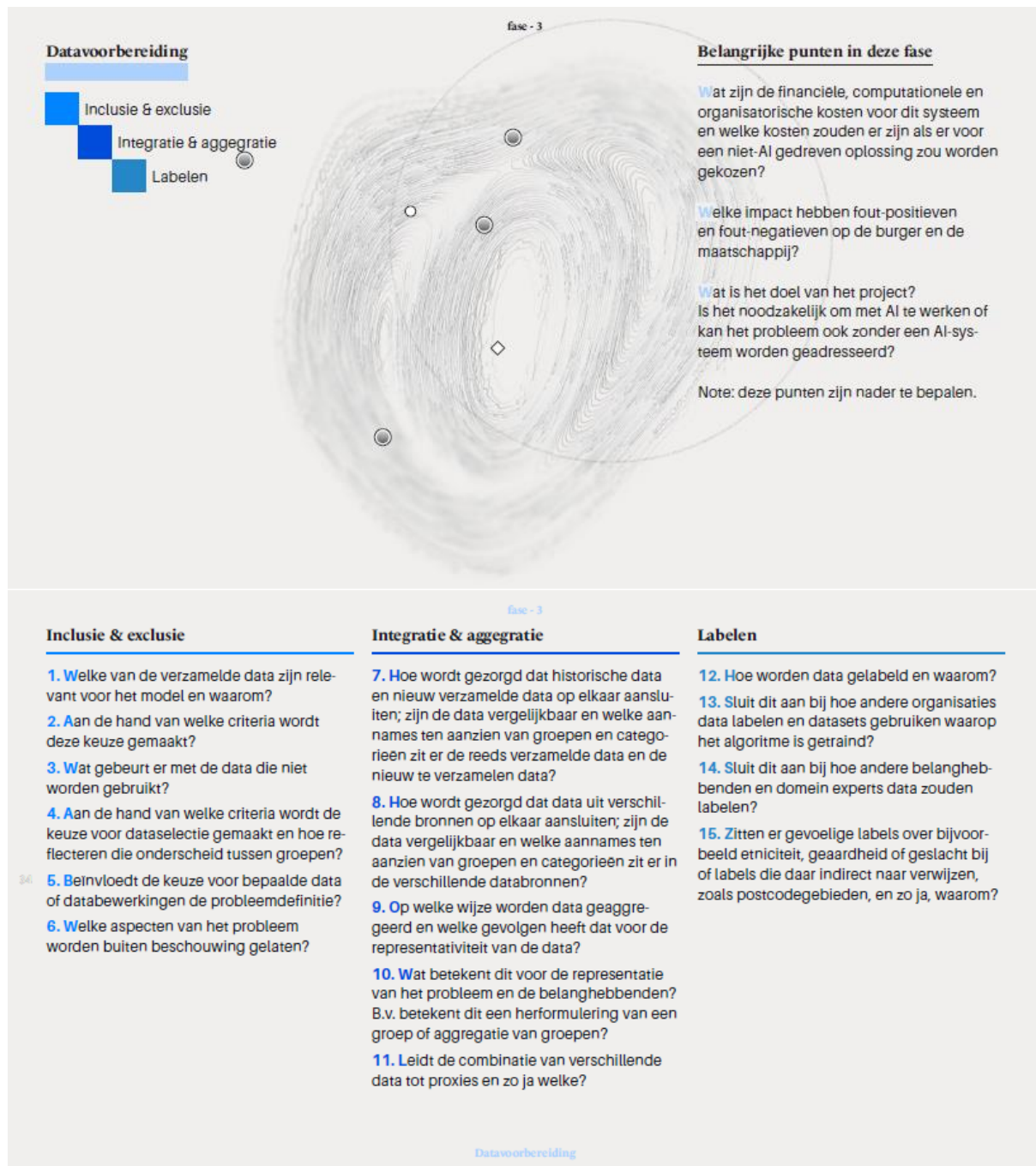
32

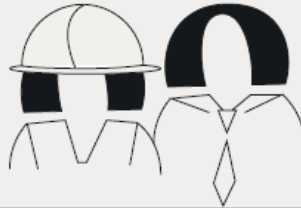


fase - 2

- ❶ **Zorg dat data zo neutraal en objectief mogelijk worden verzameld en dat dit zo transparant en controleerbaar mogelijk geschiedt.** De wijze waarop data worden verzameld, hoe, door wie, waar en met gebruikmaking van welke technieken kunnen bepalend zijn voor de neutraliteit en betrouwbaarheid van de verkregen data.
 Leg dit proces nauwgezet vast en bewaar te allen tijde gegevens over hoe de dataset tot stand is gekomen, om zo de processen controleerbaar en herhaalbaar te maken. Doe dat ook voor de data die via derden zijn verkregen.
- ❷ **Controleer op ontbrekende waarden, nauwkeurigheid en representativiteit en of de verdeling voor verschillende groepen gelijk is.** Leg verschillen uit en overweeg mitigerende maatregelen.
- ❸ **Overweeg om nieuwe gegevens te verzamelen of om het doel van het project te herzien:** ga dan terug naar fase 1 (probleemdefinitie).
- ❹ **Zorg voor organisatorische veiligheid en vertrouwelijkheid van de data.** Zorg dat data alleen met dubbele authenticatie in te zien zijn door een geselecteerd aantal werknemers of betrokkenen.

Dataverzameling





35

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Data wordt gelabeld op basis van kwaliteit en van de sollicitanten, om zo een objectieve keuze te kunnen maken voor een sollicitant. Er zitten gevoelige gegevens tussen, maar deze spelen geen rol bij de labeling.

Datavoorbereiding



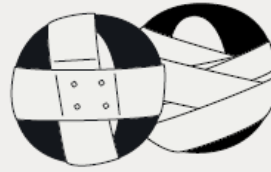
36

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Data wordt gelabeld op basis van wijken/postcodegebieden, om zo in kaart te brengen welke wijken meer aandacht behoeven. Er zitten gevoelige labels tussen, namelijk postcodegebieden. Deze kunnen verwijzen naar bepaalde etniciteit en naar sociaaleconomische status.

Datavoorbereiding



37

Aquila pratem expersp eriorposant et occum doluptat adlis idlit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquia pratem expersp eriorposant et occum doluptat adlis idlit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquia pratem expersp eriorposant et occum doluptat adlis idlit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Data wordt gelabeld op basis van onder andere leeftijd, etniciteit en geslacht, om zo te zien of deze bepaalde vorm van kanker meer voor komt in bepaalde groepen. Er zitten gevoelige labels tussen, deze zijn nodig om in kaart te kunnen brengen in welke groepen een groter percentage deze vorm van kanker heeft.

Datavoorbereiding

Juridisch

38

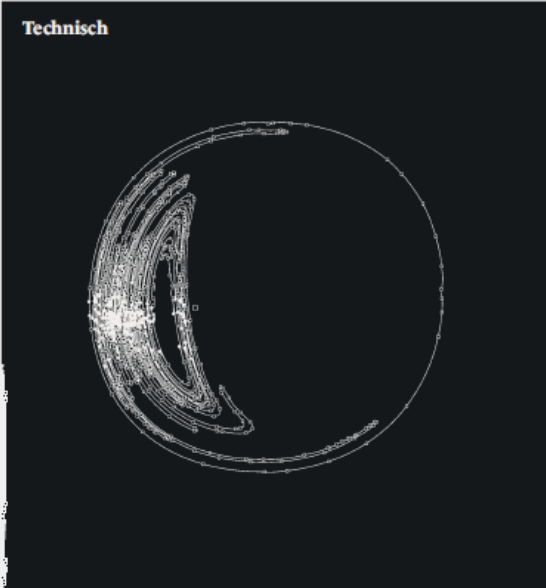


- 1 Controleer welke gevolgen de datasetselectie heeft voor de representatie, met name van categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid.
- 1 Controleer wederom welke gevolgen de integratie en aggregatie van data hebben voor categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status. Zelfs het samenvoegen van twee neutrale databases kan samen een database met sterke bias opleveren.
- 1 Controleer welke gevolgen het labelen van data heeft voor de categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status.

Datavoorbereiding

Technisch

39



Data voorbereiding

- 1 Pas pre-processing technieken voor bias-mitigation toe, zoals:
 - Instance class modification
 - Instance selection
 - Instance weighting
- 1 Welke technische beperkingen brengen de data met zich mee? Doe aanpassingen aan het technisch systeem al naar gelang de datakwaliteit dat vereist. Heeft de datakwaliteit gevolgen voor fase 1 (probleemdefinitie)?
- 1 Zijn de data op dezelfde wijze verzameld? Welk invloed heeft het verschil in de wijze (methodologie, tijd, plaats, etc.) van de dataverzameling op de mogelijkheid van integratie en vergelijkbaarheid van de data?
- 1 Welke categorieën worden gekozen ten aanzien van het aggregatieproces en waarom? Controleer op aggregatie bias: die kan ontstaan bij het bouwen van een model. Houd er rekening mee dat de vertaling van input naar labels niet altijd voor alle groepen hetzelfde is. Vermijd een "all-size-fits-all" model.
- 1 Controleer op meting bias: Deze kan ontstaan als de keuze van labels bevooroordeeld zijn.
- 1 Houd er ook rekening mee dat de granulariteit en kwaliteit van data tussen verschillende groepen kan verschillen en dat elke classificatie een versimpeling is van de werkelijkheid.

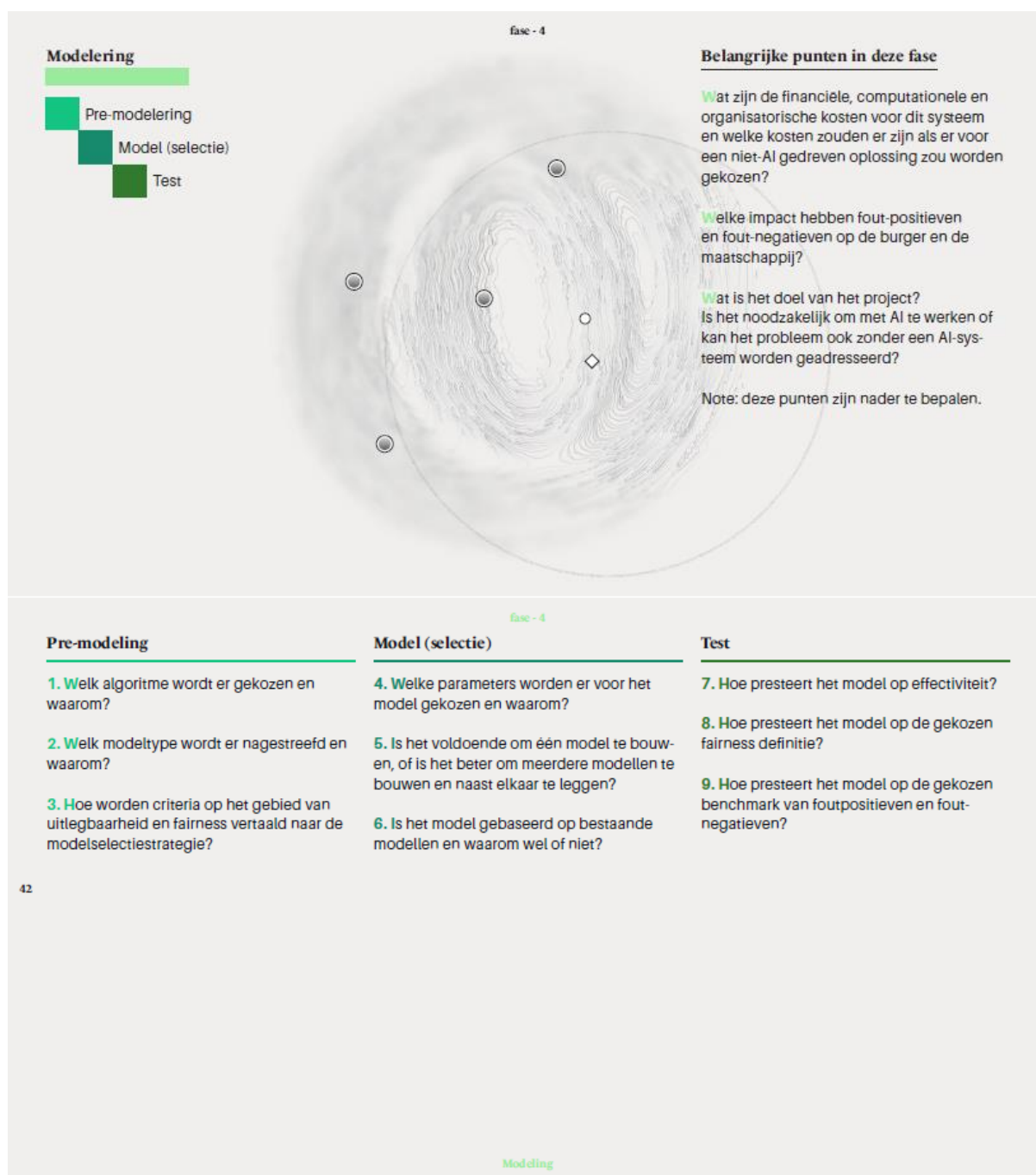
Organisatorisch

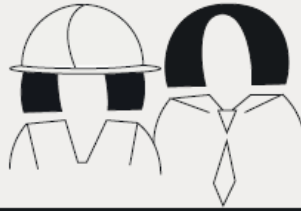
40



Data voorbereiding

- 1 Leg nauwgezet de keuzes vast ten aanzien van inclusie en exclusie. Welke rationale is er gekozen voor de inclusie of exclusie van de data en waarom?
- 1 Welke gevolgen heeft de kwaliteit en representativiteit voor de werkzaamheid van het systeem en de gestelde benchmarks? Verzamel eventueel meer gegevens als blijkt dat bepaalde groepen over- of onder gerepresenteerd zijn: ga dan terug naar fase 2.
- 1 Datasets, analysemethode en beslissingen worden gekozen met het oog op objectiviteit; fouten worden vastgelegd en direct gecorrigeerd. Als ze van belang zijn voor andere organisaties worden die onmiddellijk gedeeld.
- 1 Methodes, procedures, definities en classificaties worden consistent toegepast; deze worden zo veel mogelijk gestandaardiseerd tussen organisaties, om vergelijkingen en controle mogelijk te maken.
- 1 Controleer de gevolgen van het labelen op de kwaliteit en representativiteit en de werkzaamheid van het systeem.
- 1 Verzamel eventueel meer gegevens als blijkt dat bepaalde groepen over- of onder gerepresenteerd zijn: ga dan terug naar fase 2.





43

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.

Modeling



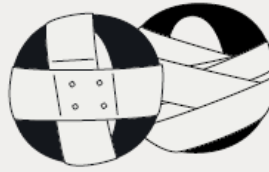
44

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.

Modeling



45

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu

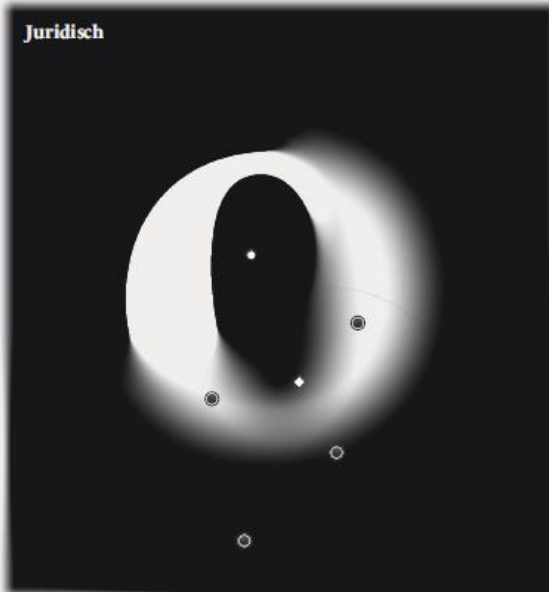
Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquia pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.

Modeling

Juridisch

46



1 Wordt er gekozen voor een model gebaseerd op causaliteit of correlatie? Wees bewust van het feit dat het juridische domein is gebaseerd op causaliteit; juridische uitlegbaarheid en juridische rechtvaardiging kunnen vaak niet worden gevonden in statistische correlaties. Ga dus na of bij een systeem gebaseerd op correlatie, de uitkomsten kunnen worden gekoppeld aan causaliteit op individueel niveau. Deep-learning systemen met black box elementen die besluiten nemen zijn bijna altijd verboden.

1 AI-systemen die zijn gebaseerd op statistische modellen kunnen rekening houden met de principes die hiervoor leidend zijn, zoals:

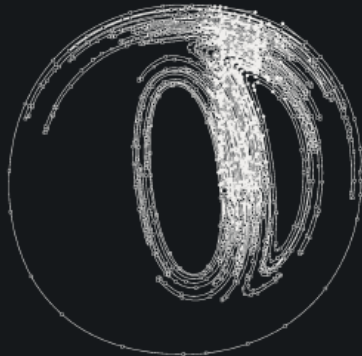
- Betrouwbaarheid
- Onpartijdigheid
- Objectiviteit
- Vergelijkbaarheid
- Consistentie

[Klik hier voor meer informatie](#)

1 Controleer hoe het model presteert t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid.

Modeling

Technisch



47

fase - 4

- 1 Algoritme selectie: Verklaar en documenteer keuzes voor een ML-algoritme in het licht van interpreteerbaarheid en verklaarbaarheid.
- 1 Modelkeuze:
 - Overweeg de vooraf gedefinieerde fairness-maatstaven bij het selecteren van een model.
 - Controleer of aangeleerde relaties een afspiegeling zijn van bestaande domeinexpertise en niet willekeurig zijn.
 - Gebruik het eenvoudigste model dat de prestatiespecificaties mogelijk maken.
- 1 Selecteer de kandidatenpoel van AI-algoritmes in het licht van benodigde fairness interpreteerbaarheid, en uitlegbaarheid. Voorbeeld: wanneer fairness criteria gesteld zijn, overweeg het gebruik van in-processing of post-processing bias mitigation methodes om hiervoor te optimaliseren. Kijk bijvoorbeeld *hier* voor een selectie van bias mitigation algoritmes. Voorbeeld: wanneer causaliteit een voorwaarde is, wordt gezorgd dat de kandidatenpoel bestaat uit direct interpreteerbare algoritmes, zoals linear regression. Voorbeeld: wanneer voorspellingen als doel hebben om menselijke beslissingen te ondersteunen, wordt gezorgd dat de kandidatenpoel bestaat uit methodes die uitlegbaar zijn voor mensen met een minder technische achtergrond, zoals een simpele decision tree.

Modeling

- 1 Vertaal succescriteria naar technische maatstaven, bijvoorbeeld op het gebied van accuraatheid, false positives, false negatives, en fairness. Voorbeeld: om te voorkomen dat het model minder goed werkt voor minderheden, zal de fairness metric equalized odds worden meegenomen in de modelselectie. Indien gebruik wordt gemaakt van bestaande modellen, neem deze dan mee in de modelselectiestrategie. Voorbeeld: in een NLP-applicatie wordt het gebruik van pre-trained word embeddings overwogen, de evaluatie van de embeddings op de succes criteria is daarom onderdeel van de modelselectiestrategie.

48

- 1 Bias-beperking:
 - Gebruik indien nodig voorbehandelingsmethoden om het algoritme aan te passen om vertekende resultaten te minimaliseren, bijvoorbeeld door aanpassing van het classificatiemodel;
 - Pas indien nodig nabewerkingsmethoden toe om bias te beperken na training van het classificatiemodel. White-box-methoden passen het model aan; black-box-methoden passen de voorspellingen aan.
- 1 Kijk bijvoorbeeld *hier* voor een selectie van bias mitigation algoritmes. Voorbeeld: wanneer causaliteit een voorwaarde is, wordt gezorgd dat de kandidatenpoel bestaat uit direct interpreteerbare algoritmes, zoals linear regression. Voorbeeld: wanneer voorspellingen als doel hebben om menselijke

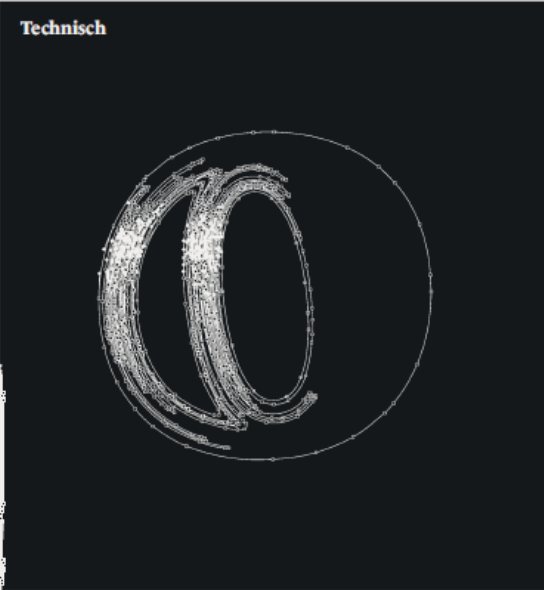
beslissingen te ondersteunen, wordt gezorgd dat de kandidatenpoel bestaat uit methodes die uitlegbaar zijn voor mensen met een minder technische achtergrond, zoals een simpele decision tree.

- 1 Vertaal succescriteria naar technische maatstaven, bijvoorbeeld op het gebied van accuraatheid, false positives, false negatives, en fairness. Voorbeeld: om te voorkomen dat het model minder goed werkt voor minderheden, zal de fairness metric equalized odds worden meegenomen in de modelselectie.
- 1 Indien gebruik wordt gemaakt van bestaande modellen, neem deze dan mee in de modelselectiestrategie. Voorbeeld: in een NLP-applicatie wordt het gebruik van pre-trained word embeddings overwogen, de evaluatie van de embeddings op de succes criteria is daarom onderdeel van de modelselectiestrategie.
- 1 Controleer of aangeleerde relaties een afspiegeling zijn van bestaande domeinexpertise en niet willekeurig zijn.
- 1 Controleer hoe de modellen presteren t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid. Selecteer het eenvoudigste model dat de prestatiespecificaties mogelijk maakt. Voorbeeld: zowel een logistic regression model als een random forest model zijn voldoende accuraat, er wordt daarom gekozen voor het logistic regression model.

Modeling

Technisch

49



Modeling

- ① Hoe presteert het model op de benchmarks van effectiviteit, de gekozen fairness definitie en de benchmark van foutpositieven en foutnegatieven. Stel het model bij aan de hand van de uitkomsten.
- ① Controleer op Evaluatie bias: dit kan voorkomen bij het testen van het model. Let erop dat de benchmarks die gebruikt worden om het systeem te testen overeenstemmen met de beoogde doelgroep.

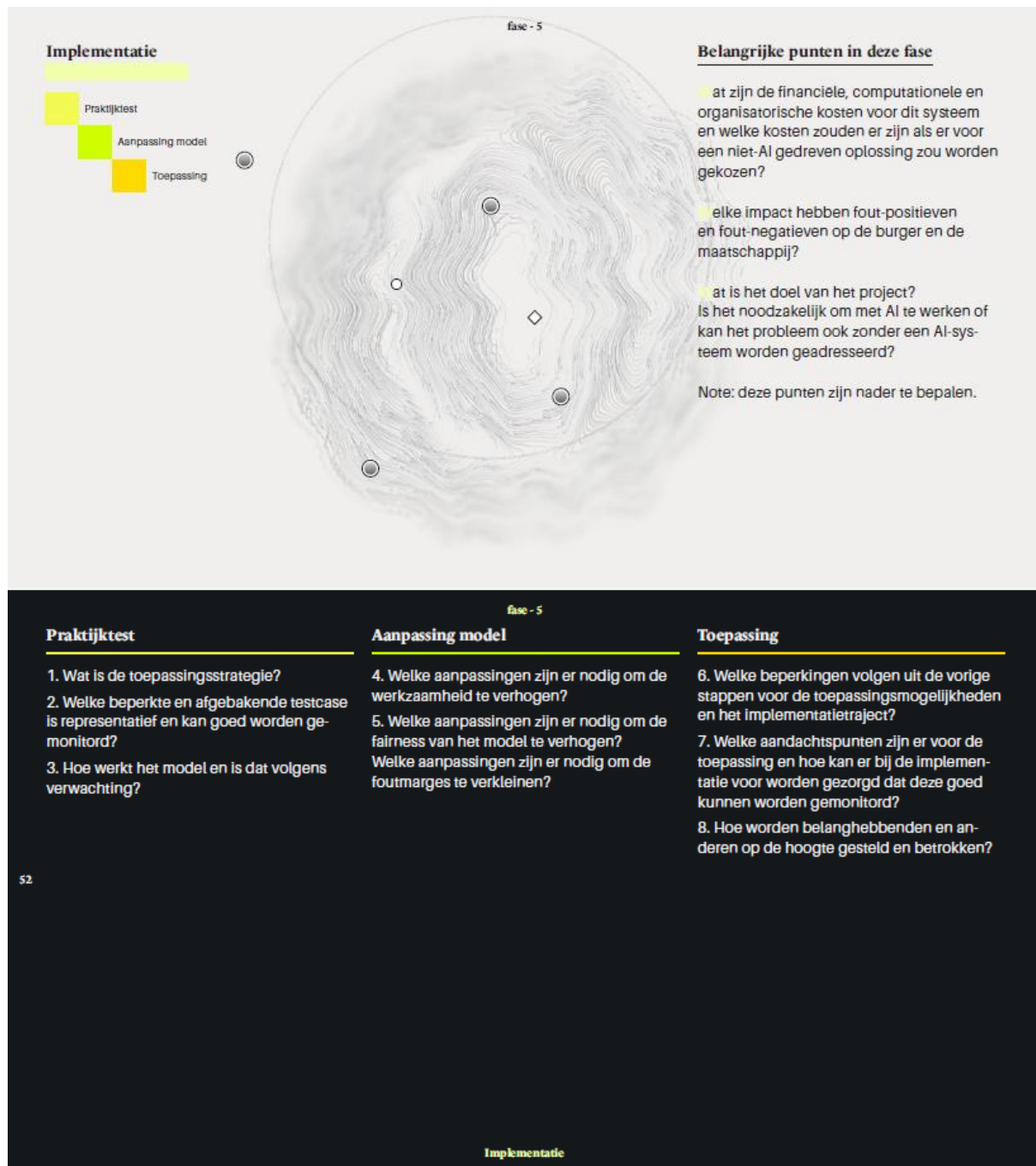
Organisatorisch

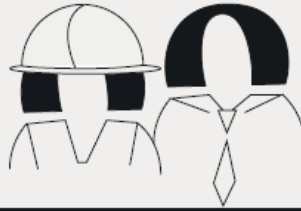
50



Modeling

- ① De modelselectiestrategie worden vastgelegd, gedocumenteerd en openbaar gemaakt. Het ontwerp is het liefst universeel, zodat de modellen goed met elkaar kunnen worden vergeleken op uitkomst en fairness.
- ① Betrek belanghebbenden, zoals eindgebruikers en besluitvormers, bij de selectie van de kandidatenpoel om te zorgen dat de interpreteerbaarheid en uitlegbaarheid van de modellen past bij het doel van het model en de achtergrond van de gebruikers.
- ① Verantwoording:
 - Metadata worden bewaard en vastgelegd.
 - Data zijn voor zover mogelijk toegankelijk voor derden.
 - Het model moet uitlegbaar zijn aan en inzichtelijk voor belanghebbenden.
 - Welk vorm van uitlegbaarheid wordt door het systeem geboden?
 - Voor wie is deze uitleg begrijpelijk?
- ① Vraag een onafhankelijk team van experts met een diverse persoonlijke en professionele achtergrond voor een second opinion.





53

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.

Implementatie



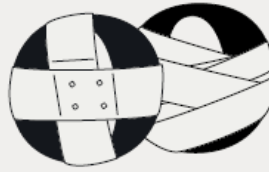
54

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.

Implementatie



55

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa inus poribus que et. **V**erupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu. **A**quila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, **C**umqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu.

Implementatie

fase - 5

Juridisch

56

- 1 Stel de belanghebbenden op de hoogte van hun rechten:
 - Recht op informatie, inclusief informatie over algoritme.
 - Recht om het besluit aan te vechten.
 - Recht om aanvullende informatie aan te dragen.
 - Recht om niet onderworpen te worden aan automatische besluitvorming. [Klik hier voor meer informatie.](#)
- 2 Controleer hoe het model presteert t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status.
- 3 Als het model significant onderscheid maakt, direct of indirect, op basis van een van deze groepen, is daar een rechtvaardiging voor en zo ja welke? Leg dit voor aan een jurist.
- 4 Stel wederom belanghebbenden op de hoogte van hun rechten:
 - Recht op informatie, inclusief informatie over algoritme.
 - Recht om het besluit aan te vechten.
 - Recht om aanvullende informatie aan te dragen.
 - Recht om niet onderworpen te worden aan automatische besluitvorming

Technisch

- 1 Beantwoord de volgende vragen:
 - Wat is de toepassingsstrategie?
 - Welke beperkte en afgebakende toepassing is representatief en kan goed gemonitord worden?
 - Hoe werkt het model en is dat volgens verwachting?
 - Wat is de exitstrategie?
- 2 Overweeg gebruikerstesten om gebruiksgemak en toegankelijkheid voor mensen met een beperking te toetsen. Wanneer het model gebruik maakt van persoonsgegevens, neem dan een vergelijking van de prestaties van het model en de prestaties binnen het huidige besluitvormingsproces op in de evaluatiestrategie. Voorbeeld de prestaties van het model worden getoetst in een pilot study bij een kleine groep gebruikers.
- 3 Monitoring en onderhoud

Wanneer het model regelmatig wordt geüpdatet op basis van nieuwe data, zorg dat ook deze nieuwe modellen grondig worden geëvalueerd.
- 4 Plan voor het monitoren van veranderingen in de datadistributie, zoals concept drift en een verschuiving in de demografie van data subjects.

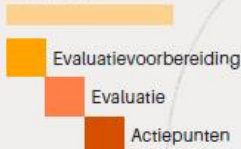
Implementatie

fase - 5

- 1 **Evalueer het model op:**
 - Effectiviteit
 - Fairness
 - Foutpositieven en foutnegatieven.
- 1 **Voer verbeteringen door.**
- 1 **Beoordeel wat de context van de gekozen testcase zegt over de algemene werking van het systeem en denk na in hoeverre andere toepassingsgebieden andere contextgevoeligheden met zich brengen. Neem daarop passende maatregelen.**
- 1 **Monitoring en controle:**
 - Plan evaluatiemomenten
 - Documenteer de gebruikresultaten
 - Controleer op exitcriteria
 - Maak API's beschikbaar voor externe auditors
 - Maak databladen en modelkaarten voor zover mogelijk openbaar.

Organisatorisch

- 1 **Stel betrokkenen en belanghebbende op de hoogte van het feit dat het AI-systeem in een testsetting wordt geïmplementeerd. Doe dat waar mogelijk voordat zij in aanraking komen met of de gevolgen ondervinden van het systeem.**
- 1 **Overleg met zowel belanghebbenden als externe experts over de functionaliteit, fairness en accuratesse van het model.**
- 1 **Implementeer voor zo ver mogelijk hun adviezen en suggesties.**
- 1 **Leg veranderingen in het systeem vast.**
- 1 **Documenteer het gekozen model en de resultaten van de evaluatie, bijvoorbeeld in een *model card*.**
- 1 **Stel wederom iedereen op de hoogte van het feit dat het AI-systeem wordt geïmplementeerd en voer een klachtenprocedure in.**
- 1 **Vraag een onafhankelijk team van experts met een diverse persoonlijke en professionele achtergrond voor een second opinion.**

Implementatie**Evaluatie**

fase - 6

Belangrijke punten in deze fase

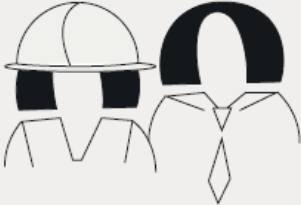
Wat zijn de financiële, computationele en organisatorische kosten voor dit systeem en welke kosten zouden er zijn als er voor een niet-AI gedreven oplossing zou worden gekozen?

- Welke impact hebben fout-positieven en fout-negatieven op de burger en de maatschappij?

Wat is het doel van het project?
Is het noodzakelijk om met AI te werken of kan het probleem ook zonder een AI-systeem worden geadresseerd?

Note: deze punten zijn nader te bepalen.

Evaluatie

fase - 6		
Evaluatievoorbereiding	Evaluatie	Actiepunten
1. Wordt er gekozen voor een permanente evaluatie, specifieke evaluatiemomenten of beide? 2. Wordt er gekozen voor een interne evaluatie, een evaluatie door externen of beide? Hoe wordt de evaluatie getest en met welke meetpunten?	3. Hoe functioneert het systeem ten aanzien van de benchmarks? 4. Welke aanpassingen zijn er ten aanzien van de beschermde categorieën nodig? 5. Hoe zou het systeem functioneren met een ander model, fairness definitie en/of algoritme?	6. Moet het systeem al dan niet tijdelijk worden stopgezet? 7. Kunnen gevonden problemen en obstakels worden verholpen? 8. Wat vinden belanghebbenden en externe experts van de evaluatieresultaten?
Evaluatie		
fase - 6		
voorbeelden Ai-inzet		Arbeidsmarkt
60 Aquila pratem expersp erlorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. Verupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu	Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu. Aquila pratem expersp erlorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,	Aquila pratem expersp erlorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que voluptiam as eum quam, ut possum quia que doluptati omnimu.
Evaluatie		

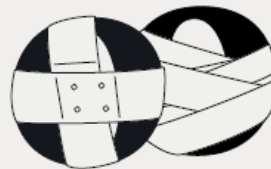


61 **A**quila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. **V**erupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu. **A**quila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, **C**umqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu.

Evaluatie



62 **A**quila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, quodigenim es venecuptam et elum necus minulpa Inus poribus que et. **V**erupicta sime nis modictation cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu

Cumqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu. **A**quila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit,

Aquila pratem expersp eriorposant et occum doluptat adis idit occupiet quia si consequasi aut abor rerspici sernati beri nus, omnit, **C**umqui doluptur rem. Ore, sa dolorepero dolenimendi necepel id utemporiti dolessit volorunt que volupitiam as eum quam, ut possum quia que doluptati omnimu.

Evaluatie

Juridisch

- 63
- 1 Als er persoonsgegevens worden verwerkt moet het systeem duidelijk effectiever zijn dan de status quo zonder het AI-systeem om te voldoen aan de noodzakelijkheids-, proportionaliteits- en subsidiariteitstest. Verantwoord dit.
 - 1 Controleer hoe het systeem presteert t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid, vermogen, geboorte of andere status.
 - 1 Als het model significant onderscheid maakt op basis van een van deze groepen, is daar een rechtvaardiging voor en zo ja welke? Ga na of het systeem al dan niet tijdelijk moet worden stopgezet of dat naar een van de vorige fasen in het proces moet worden teruggegaan om aanpassingen te doen.
 - 1 Vraag een externe jurist om een second opinion te doen en advies te geven ten aanzien van de juridische randvoorwaarden voor het systeem.

fase - 6

Technisch

- 1 Leg de volgende punten nauwgezet vast:
 - Hoe wordt de evaluatie ingericht en waarom zo?
 - Waar wordt de verantwoordelijkheid voor de evaluatie belegd en waarom?
 - Welke meetpunten worden gekozen en waarom?
- 1 Plan een uitgebreide evaluatie waarin wordt.
- 1 Overweeg gebruikerstesten om gebruiksgemak en toegankelijkheid voor mensen met een beperking te toetsen.
- 1 Wanneer het model gebruik maakt van persoonsgegevens, neem dan een vergelijking van de prestaties van het model en de prestaties binnen het huidige besluitvormingsproces op in de evaluatiestrategie. **Voorbeeld:** de prestaties van het model worden getoetst in een pilot study bij een kleine groep gebruikers.
- 1 Evalueer het model op:
 - Effectiviteit
 - Fairness
 - Foutpositieven en foutnegatieven
- 1 Voldoen de presentaties aan de voorgestelde benchmarks. Zo nee, stop het project tijdelijk of permanent. Zo ja, zijn er mogelijkheden om het systeem nog effectiever, rechtvaardiger of accurater te maken?

Evaluatie

- 64
- 1 Hoe functioneert het systeem:
 - Op andere data?
 - Met een ander algoritme?
 - Met een andere fairness definitie?
 - Met een ander model?
 - 1 Vraag een externe data-analist om een second opinion te doen en advies te geven ten aanzien van het technisch systeem.

fase - 6

Organisatorisch

- 1 Betrek belanghebbenden bij de evaluatie. Voorbeeld: Doe een survey of focusgroep met belanghebbenden om hun ervaringen in kaart te brengen.
- 1 Stel betrokkenen en belanghebbende op de hoogte van het feit dat het AI-systeem in een testsetting wordt geïmplementeerd. Doe dat waar mogelijk voordat zij in aanraking komen met of de gevolgen ondervinden van het systeem.
- 1 Ga na of het systeem, het model, de data en/of de evaluatie openbaar worden gemaakt, al dan niet in geanonimiseerde vorm.
- 1 Documenteer de evaluatiemethodes, motivaties voor keuzes, en verantwoordelijkheid voor de evaluatie.
- 1 Documenteer het gekozen model en de resultaten van de evaluatie, bijvoorbeeld in een *model card*.
- 1 Vraag een bedrijfskundige om een second opinion te doen en advies te geven ten aanzien van de procesinrichting van het systeem.

Evaluatie

65	Conclusie Mendam, sentotatium velestis dolor audam is dolorpo rehenis molorumque esserro torro to officiam rempeliquia neculparum volo dus disitionem licipit am lacepta exercima vellent lam sectur? Quibus dolorum doluptate res exerunt dent quia nonsequatur? Nam, inctat dolorit peratus, tor aut eatus, illicit molum eos nostistiur molendi repro es dolupta estioneseque pelescia volupta tempore ma quate mil in corruptasit, siti cuptae con ni omnias mincimos et invelique pratquo moditiatur, nonserunt. Ores dolor aut esersperor am reratis ma con ratiium unt. Lesernam consequae et, te dolupta delitae aspeliquam quat a si nihictae eturem facepedi totatem. Et amusand uciisim oluptat ecuscim porias utem facerum animpos quias verest utasperis sum reptatem ea nus sint parit, voluptatiati net aut landaerferes magnien dionet voluptatatia seque incius, cones res eostist orrumendenis quo doloriam ea esecatam a simagnit et laborum sus, sitasped magnihitibus sequiat emolorro moditis am fugiae consequia consequi berovit estem et qui aut que dolorate nullatus ventiam, cus et aut volorro tec- totatur si optas ma doluptas solest is alic temolectatur arum, nonsequ iducim doluptaquis sectioneius ius reruntus con pe molorerum et, officia excepelendit quo tem et ommoditiorro consectenis dollantur? Evendero illorum aut et mo qui ide repro id maioribus moluptibus dit et dem liquuntiorio blatis esequam, vent officimus dit eost, et latecat initatis doles sunt venim est, sin earunt que eos quis dent dende acerum dolum, quam, quodit, consedis ex- ceped untisitatur, odipsam, volut quisima ximintis alibus, sa que aspid magnatu ribusapit aut list, simus quatiores resti tempellabo. Itatecae expliqu latinul parisi as ent estrumquamet lam fugitibus sitatum ea ventis nemporibus, volorem ut quaeptatis ut occulli tateceaque qui ipsunt erionse quatemquate suntiis aperatio cus.
66	Colofon Mendam sentotatium Velestis dolor Audam is dolorpo Rehenis molorumque Esserro torro To officiam Rempeliquia Neculparum volo Dus disitionem Ricipit am lacepta Esercima vellent Lam sectur Lulibus dolorum Doluptate res Exerunt dent

7.7 Verslagen Workshops

7.7.1 Workshop 1

Verslag workshop AI en discriminatie
04 December 2020 15.30-17.30

Aanwezigen

Josine van der Brug (Tactlegal)

Wouter van der Bij (VWS)

Esther Keymolen (TILT)

Sander Klous (KPMG & UvA)

Egge van der Poel (TIAS)

Erwin Redeman (Neurocast)

Bart van der Sloot (TILT)

Verslag workshop AI en discriminatie
04 December 2020 15.30-17.30

Aanwezigen

Josine van der Brug (Tactlegal)

Wouter van der Bij (VWS)

Esther Keymolen (TILT)

Sander Klous (KPMG & UvA)

Egge van der Poel (TIAS)

Erwin Redeman (Neurocast)

Bart van der Sloot (TILT)

Voorstelronde

Keymolen: Uitleg en korte toelichting handreiking

Redeman: Uitleg van het Neurocast systeem en het team:

Who we are

We specialize in passively monitoring patients. We help patients, doctors, and researchers better understand the course and cause of chronic diseases such as MS, Alzheimer's, or cancer. We believe that much of the data that could tell us something about how you are doing – mentally, physically, or emotionally – is hidden in how we interact with technology.

The challenge

It is virtually impossible to obtain a round picture of a patient with clinical data alone. There is an obvious need to monitor patients in their daily environment, but collecting a consistent flow of Real World Data is a complex matter. People suffering chronic illnesses do not want to be confronted with their illness even more, therefore active measuring often fails.

Our mission

We've made it our mission to turn everyday digital interactions into clinically approved outcomes, enabling doctors & researchers to passively measure individual patients performance in daily life. With our technology this can be done completely unobtrusive, with privacy embedded by design and available for everyone to join.

What we can do for you

The Neurocast True Passive Patient monitoring platform can handle a wide variety of research needs. Our technology stack is available out of an ISO27001 Certificated AWS Virtual Private Environment. Ready for use within 48 hours.

Neurocast for Pharma

The Neurocast platform can be used to measure primary and secondary clinical end points within clinical trials. Increasing efficiency and (cost) effectiveness of drug development by passive and monitoring, with easy to deploy technology.

Neurocast for Care & Cure

Data in a Real-World setting enables doctors to actively monitor their patients and proactively act upon status changes. The first step in Telemedicine where doctors visit their patients instead of patients visiting their doctors base on RWE.

Neurocast for Researchers

We have research in our DNA and are well connected with scientists around the globe. As a researcher looking at our technology and with a compelling idea, feel free to drop us a message. We are always looking for new projects.

The Neurocast platform

We believe that much of the data that could tell us something about how you are doing – mentally, physically, or emotionally – is hidden in how we interact with technology. Therefore we've made it our mission to turn everyday digital interactions into clinically approved outcomes, enabling physicians & researchers to unobtrusively measure individual patients' performance in daily life.

We started with the most frequently used device, being the smartphone, meanwhile adding other devices that have become an unseen part of our daily lives, such as computers and smartwatches. Neurocast not only gathers keystroke dynamics data with sensor, Health, and e-PROM data but above all, we develop algorithms that place our Real-World Data into the right context and correlate it with gold standards used by researchers worldwide.

Privacy & security is crucial for patients and, therefore, for us. We don't collect data that can be traced back to individuals patients. Neurocast is GDPR & HIPAA compliant and ISO 27001 certificated.

Completely unobtrusive . Privacy embedded by design . Available for everyone

Private Cloud Environment

The complete Neurocast technology stack is available in an ISO27001 certificated Virtual Private Cloud environment. This is especially designed for clinical trials and research projects, compliant with the highest security and privacy standards.

We offer our industry and academic partners a dedicated instance of the Neurocast software and infrastructure:

Neurokeys Apps

Whitelabel options

Smart Panel ® Technology

Data Science Models

*BI Dashboards & tooling
Tailored project website*

High patient adherence

We believe that patients shouldn't even notice that they are constantly measuring themselves. Only then have we succeeded in making our technology truly unobtrusive. That is why we are in constant dialogue with users to learn how we can do better. We are proud that 95% of patients using our technology are still active after their first year.

Highest security standards

73% of patients have valid concerns about privacy & security. It is often unclear what kind of data is used, for what purpose, and with whom it is shared. Safety is of paramount importance at Neurocast. We do not collect data that can be traced back to individual patients, and all data is anonymously stored in a highly secure ISO 27001 certificated and GDPR / HIPAA compliant environment.

Clinically validated technology

To ensure that our Real World Data can be used in a medical environment, we must be continuously working on validation. In addition to our clinical validation studies in collaboration with leading institutes such as Amsterdam UMC, we are always working on implementation validation together with our customers, to ensure that our technology can be integrated quickly.

Discussie

Redeman: Belangrijkste voor de handreiking lijkt mij het begin. De start van een project. In de praktijk begin je een project vaak zonder groot plan; je begint gewoon. Daarom is het essentieel om aan het begin een aantal vragen te stellen en doelstellingen formuleren. Als je dat doet en binnen die kaders blijft, dan zit je denk ik grosso modo goed. Wij hebben voor een inclusief model gekozen; iedereen moet mee kunnen doen, zeer laagdrempelig. Bovendien slaan wij geen gevoelige gegevens op. Wel moeten wij MS patiënten van gewone burgers/andere patiënten te onderscheiden, want anders ben je gewoon typgedrag aan het meten. Ook proberen we in een vroegtijdig stadium mensen met alzheimer/dementie te onderscheiden. Je kan in theorie nog veel meer data verwerken – energiedata, transportdata, smart producten, etc. – en daarmee nog betere voorspellingen doen. Maar dat moet binnen de juiste kaders zijn, dus hopelijk kan jullie initiatief daaraan bijdragen. Voor ons zijn dit de uitgangspunten:

- Privacy by design
- Controle individu
- Data minimalisatie
- Niet op persoonsniveau
- Encryptie
- Organisationele (veiligheids)waarborgen
- Training van medewerkers; continue awareness
- Alles goed documenteren
- Pentest door hackers
- Wij slaan geen/minimaal data op

In het begin moet een team nog iets breder kijken dan nu in de handreiking staat: wat is jouw visie op privacy, non-discriminatie en andere mensenrechten. Hoe belangrijk vind je dat? Wil

je je alleen aan de wet houden of ga je verder? Voor wie doe je het? Wat zijn de waarden van de belanghebbenden? Zorg dat die kernwaarden de basis van je project zijn, want zij zijn de basis.

Van der Poel: Meer in het algemeen: CRISP-DM model is goede basis, maar kent ook beperkingen. Een van de kritiekpunten is dat de evaluatie niet dynamisch genoeg en permanent genoeg is. Er moeten allemaal feedbackloopjes zijn van stap 5 naar 3 en van stap 2 naar 4; al die evaluaties moeten worden terug gevoegd aan het systeem en terug worden gekoppeld aan de burger/het data subject/de belanghebbende.

Redeman: Burger is in principe eigenaar van zijn data; wat wil hij er mee? Alleen zichzelf helpen of ook de gemeenschap; hoe veel data wil hij delen en waartoe? Dat ligt primair bij het individu. Daarom is dynamische consent ook zo belangrijk.

Van der Brug: Belangrijk om in een testfase niet alleen te focussen op zo min mogelijk data, omdat je wil kunnen testen op discriminatie. Handreiking is te lineair opgebouwd; je moet misschien stappen terug en soms een stap overslaan. Het overzicht van de anti-discriminatiewetgeving is nu in drie stappen die elkaar opvolgen, terwijl het volgens mij aparte vragen zijn die door elkaar kunnen lopen. Discriminatie-effect: denk auto's die vrouwonveilig zijn/op mannenlichamen zijn gebouwd.

Klous: Maar wat zou het discriminatieprobleem binnen dit systeem kunnen zijn?

Van der Brug: Stel diagnose werkt wel goed bij vrouwen en niet bij mannen.

Klous: Maar gebruik je het dan niet?

Van der Poel: Jawel, maar wat doe je aan mitigatie en gebruik je de winst bij die je behaalt bij de ene groep door AI om de niet-AI gestuurde diagnose voor de andere groep te versterken?

Van der Brug: Bij het ontwikkelen van AI is het essentieel om te benadrukken dat we bij menselijke beslissingen/handelingen wel fouten accepteren en bij computers veel minder.

Van der Bij: Wellicht goed om in het begin, voor de start van het systeem, al te analyseren hoe de huidige (menselijke) praktijk al discriminatoir is en hoe het AI systeem daar een verbetering op zou kunnen zijn. Daarnaast belangrijk bij deze toepassing: wil je het eigenlijk wel weten, zo'n early warning, en onder welke voorwaarden?

Redeman: wij zijn een adviserende tool en hebben te veel foutpositieven en negatieven om als diagnostic tool te gelden. Daar gelden strenge regels voor en dat is ook goed.

Klous: Perfectie/accuraatheid is iets anders dan rechtvaardigheid. Als je iets communiceert over accurateid, dan wordt dat in de publieke opinie vaak verward met rechtvaardigheid. Dat kan betekenen dat je restrictiever bent over wat je communiceert.

Redeman: een bijkomend probleem is dat als je transparant bent en alles documenteert er een gevaar is voor concurrentie die met de data aan de haal kan gaan. Wees je bij het schrijven van de handreiking ook bewust van economische belangen: investeerders + werknemers van het bedrijf.

Van der Brug: Belangrijk om heel precies te documenteren welke stappen je doet, ook als je bijvoorbeeld een stap terug zet en waarom dan.

Klous: Belangrijk om daarbij goed na te denken over transparantie voor wie en ten behoeve van wat. Grosso modo kun je een onderscheid maken tussen: (1) technische uitlegbaarheid (bv voor technische audit), (2) procesuitlegbaarheid (is waarschijnlijk juridisch het meest relevant) en (3) uitlegbaarheid ten behoeve van de belanghebbende (die wil vaak met name kunnen begrijpen of hij het met de uitleg eens is). Daarnaast is belangrijk ten aanzien van wat eerder werd gesteld, het meer data verzamelen/opslaan om te checken op discriminatie, hoe je dat gaat uitleggen in verband met de privacy.

Van der Brug: p. 25 mist nog de vraag of je data zelf moet verzamelen/opslaan/verwerken of dat het ook anders kan. Daarnaast zijn jullie in de handreiking soms stellig, maar juridisch is alles afhankelijk van de context en de omstandigheden van het geval. Aanvullend achteraf: stel juristen een algemene vraag, dan krijg je een algemeen antwoord waar je vaker niet dan wel iets mee kunt; maak je een vraag concreet en specifiek voor een situatie, dan krijg je een antwoord waarop je verder kunt bouwen. Dat spanningsveld zou ik ergens opnemen in de handleiding.

Van der Bij: De keuze voor een AI systeem op basis van correlaties of van causaliteit is heel belangrijk punt, dat meer aandacht zou kunnen krijgen, bijvoorbeeld al in de eerste fase.

Klous: Ook een belangrijk gevaar is dat de toepassing van algoritmes vaak uit de context wordt gehaald. Als je je algoritme licenceert voor hergebruik door derde partijen, dan verlies je controle over hoe het wordt gebruikt en of het wel in die context werkt of nieuwe problemen oplevert. Dat kan ook aansprakelijkheidsproblemen opleveren

Van der Poel: Er is een belangrijk spanningsveld tussen de juridisch realiteit (controle en documenteren) vs. de technische praktijk (experimenteren en proberen)

Redeman: Divers team samenstellen is heel belangrijk en goed dat dat wordt benadrukt. Wij hebben dat ook als belangrijk aanknopingspunt. Daarnaast zou ik dus in het begin goed nadenken over de zeg 10 gouden regels van je AI project en die als leidraad houden voor je team. Tot slot belangrijk de vraag: als het kan en mag, moet je het dan per sé doen? Dit is gekoppeld aan die gouden regels: bv, is dit in het belang van de burger?

Van der Poel: Divers team is essentieel. Het is belangrijk om bij de ontwikkeling van het systeem tegendenkers te organiseren. Dat helpt bij het draagvlak en de robuustheid van een systeem op lange termijn.

Van der Bij: Eens; al vanaf het begin zoveel mogelijk verschillende stemmen betrekken.

Klous: Belangrijk dat dit ook een algemene waarheid is; diverse teams werken gewoon beter en efficiënter. Dus eigenlijk altijd een goed idee om hier op in te zetten.

Van der Sloot: We hebben het CRISP-DM model gebruikt als basis, maar hebben hem ook iets aangepast, met name in de laatste twee fasen. Is dat problematisch?

Van der Poel: Nee, het CRISP-DM is goed om als basis te gebruiken, want veel gebruikte standaard, maar zoals ik in het begin al zei heeft het ook belangrijke beperkingen. Er zijn meerdere aanpassingen voorgesteld op dit model. Zie bijvoorbeeld ook het [CRISP-ML](#) model, dat specifiek ingaat op dit model voor Machine Learning toepassingen.

Klous: Eens. Zie bijvoorbeeld ook het [CRISP-IDM](#) model dat juist ook binnen de healthcare sector als een aanpassing op het CRISP-DM model wordt voorgesteld.

7.7.2 Workshop 2

Workshop 2
8 December 2020
10.00-12.00

Workshop AI Nota's

Introductie

- Thomas Attema: cryptoloog; data analyses op privacy vriendelijke manier uitvoeren. Hands-on technische persoon.
- Dennis Craandijk: AI lab nationale politie; AI onderzoeker, technisch persoon; onderzoek en implementatie van AI technieken op het gebied van cold cases.
- Nils Hulzebosch: politie Amsterdam; team rendement operationele informatie. Data scientist. Computer vision; objectherkenning, tekstherkenning, image classification. Digitale researchers kunnen deze tools gebruiken in de operatie.
- Jelte Mense. Onderzoeker. Data science inzetten in de organisatie. Simulatie.
- Jeroen van Rest: 20 jaar bij TNO, veiligheidsorganisaties zoals politie/mararchaussee helpen met innovatie. Niet alleen innovaties maken (R&D) maar ook helpen om te innoveren (toe te passen). Afwijkend gedrag, surveillance technologie, privacy by design, profileren. Technische achtergrond, opgeleid als projectleider. Consultant.
- Anna Garghetti: Programma manager bij de politie. Niet technisch. Politieacademie om te praten over ethisch profileren (niet etnisch!). Digitale transformatie, wat is de relatie mens machine. Rechten aan robots geven. Meer science fiction denker.
- Jeannine Florie. Eenheid midden-Nederland. Ondersteunt als adviseur Mark ???? in zijn portefeuille ethiek en privacy.
- Amir Niknam. Adviseur Innovatie bij Politie Nederland.

Handreiking

Tegengaan van discriminatie bij de inzet van AI systemen. Vertaling juridisch raamwerk en praktijk. In de handreiking proberen we hier handvatten voor te geven waar systeemontwikkelaars mee aan de slag kunnen.

Jeroen presentatie: Digitale Perimeter Gezichtsherkenning

Het project draait om privacy-bescherming bij niet-coöperatieve gezichtsherkenning. De politie werkt hierin samen met de Johan Cruijff-Arena.

Het doel van het project is om aan te tonen dat er herkenningstechnologie beschikbaar en demonstreerbaar is die draagvlak heeft. Het is expliciet niet het doel om de technologie te implementeren.

Om privacy verder te conceptualiseren gaan ze uit van de taxonomie van Solove, die privacy-dreigingen beschrijft in termen van bijvoorbeeld informatie verzameling, informatie processing, informatie disseminatie, etc.. Zulke dreigingen kunnen zich ook in het project voordoen: bijvoorbeeld als je databronnen aggregereert kan data leiden tot een privacy lek. Afhankelijk van de uitvoering kun je nog meer last hebben van andere dreigingen. Verder hebben ze in hun analyse van de privacy dreigingen gebruik gemaakt van de brief van de minister over gezichtsherkenning, de WODC studie van Keymolen et al. over horizontale privacy bij het gebruik van gezichtsherkenning en een publicatie van de ACM conferentie waarin wetenschappers onderschrijven dat ze terughoudend willen zijn met ontwikkeling van gezichtsherkenning.

Relevante privacy dreigingen in het kader van het gelaatsvergelijking systeem zijn onder andere:

- De mens is een organisme en de categorieën zijn dus ook niet discreet. Gezichten zijn niet ontworpen om uniek te zijn (e.g. zoals kentekens).
- Afhankelijk van de use case zijn er ook redenen om het systeem te willen ‘foppen’.

Het gaat in dit project expliciet niet om de proportionaliteitsoverwegingen. Mocht een dergelijk systeem ooit geïmplementeerd worden dan moet dit natuurlijk wel nog gedaan worden.

Na een inventarisatie van de verschillende privacy by design technieken heeft het team besloten om in te zetten op *multi-party computation*. We proberen een match te maken tussen live beelden en een *watch list*, zonder dat de verschillende data/templates ooit op dezelfde plek zijn. De vergelijking vindt plaats in het versleutelde domein.

Voor de gezichtsherkenning gebruiken ze een open source pakket. In het systeem proberen ze via *multi party computation* de “afstand” tussen twee gezichten te berekenen.

Vragen over het project:

Merel: we gaan in deze meeting kijken naar het discriminatie-aspect, dus niet naar de privacy dreigingen. Zo is uit onderzoek gebleken dat gezichtsherkenningsalgoritmes niet altijd even goed werken voor alle groepen.

Dennis: hoe komt je gezicht in de database?

Jeroen: in het project maken we gebruik van acteurs. Je wordt uitgenodigd als vrijwilliger om mee te doen en dan geef je expliciete toestemming. Goed om te realiseren dat we dit project de demonstreerbaarheid en beschikbaarheid toetsen. Het is geen implementatieproject en het is dus ook niet zo dat de politie nu gaat besluiten om dit te doen. Er is dus in dit project aan bijvoorbeeld Solove’s idee van vervorming als privacy dreiging geen aandacht besteed omdat het geen implementatieproject is.

Merel: Met goedkeuring van Jeroen, filosoferen wij tijdens het gesprek vandaag iets verder dan het project, dus we kijken wél naar mogelijke implementaties. Stel je voor dat dit project naar implementatieproject gaat, hoe gaan wij van het begin non-discriminatie meenemen?

Jeroen: Casus overvaller juwelier. Hij is nog in de buurt, dus politie laat gezichtsherkenning los 200m van de juwelier en dat draait voor een half uur.

Eerste indruk handreiking

Merel: wat zijn jullie eerste indrukken van de handreiking als het in casus zoals beschreven gebruikt zou worden?

Jelte: Te grote overlap met het kwaliteitskader Big Data. Bij totstandkoming niet veel data science. Het is veel te lang. Ik ga geen boeken lezen bij het begin van een project. Wetten zijn al moeilijk genoeg als AI specialist. Zeker wetten in context is moeilijker. Jurisprudentie is nog ingewikkelder. Wat ik zou willen is max 2 tot 4 pagina’s, wat is de kern van jurisprudentie van non-discriminatie in Nederland wat verder gaat dan als ik zelf 5 minuten ga nadenken. En voor wie is het bedoeld?

“Team zo divers mogelijk...” wat willen we dan als politie? Criminelen? Burgers? ... bereiken we er wel non-discriminatie mee?

Het document maakt ook een onderscheid tussen experimenteren en implementeren. Zo’n systeem als de digitale perimeter is een experiment. Hier maakt het minder uit dat het systeem discrimineert. Je wilt juist dat het in zo’n experiment naar boven komt, dan kunnen we het gaan opvangen en aanpassen. Die experimentatiefase is juist bedoeld om te weten waar het discrimineert.

Haal het vet van het huidige document af en maak een kort maar diepgaand document.

Merel: met diepgaand bedoel je jurisprudentie? Kan je dan iets meer specifiek zijn? Heb je een voorbeeld?

Jelte: Bijvoorbeeld het [WODC onderzoek](#) naar digitale middelen staat bomvol casussen. Wanneer mag je nou een telefoon openmaken en wanneer niet? Voor mij als AI-specialist heeft de handreiking momenteel niet veel meerwaarde, misschien wel voor projectleider.

Anna: je hebt het fantastisch gedaan om alles te benoemen Jelte, waarvoor dank. Ik heb dezelfde vraag: voor wie is dit bedoeld? Je moet verschillende mensen aanspreken. Wat is definitie van discriminatie? Wat maakt dat iets discriminatie is?

Merel: Definitie is de juridische, dat is het uitgangspunt. Als systeemontwikkelaar wil je misschien iets meer dan de wet voorschrijft en wil je misschien net iets meer dan de enge juridische interpretatie dus je kan ook hier en hier over nadenken. De insteek was om een breder kader te geven maar de basis is de juridische definitie.

Anna: misschien moeten we verschillende aspecten in verschillende talen samenvatten, de transparantie van jouw keuze.

Dennis: Eens met Jelte over de overlap met het kwaliteitskader en de ‘guidelines high level expert group european commission’. Dus wordt op de duur beetje een opeenstapeling van documenten die organisaties overspoelen. Ik zie veel privacy-dingen erin opgenomen en niet per se non-discriminatie. Bijvoorbeeld opmerkingen over data labelen etc. Dat zijn dingen die gewoon al moeten en gedaan worden in het kader van privacywetgeving. Of de vraag “Hoe sluit je het aan bij hoe andere organisaties labelen?” Wat voegt dat toe aan non-discriminatie...?

Nils: Eens met Jelte en Dennis, iets praktischere guidelines waren goed geweest, nog vrij abstract en lang dus demotiverend om door te lezen

Merel: wat voor praktische handvaten ben je naar op zoek? Wat mis je?

Nils: bepaalde performances waaraan voldaan moet zijn bijvoorbeeld in termen van accuracy, precision, recall, thresholds, etc.

Jelte: “check of je data gebiased is” Hoe dan? Zeg me hoe je dat moet doen.

Merel: jullie willen weten wat teveel bias is, vanaf welke waarden etc?

Jelte: wat zijn op het eerste gezicht acceptabele waarden? Is zo iets dan gerechtvaardigd mits A, B, C, D?

Dennis: Techneuten hebben behoefte aan cijfers.

Jelte: de discussie is dus hoeveel vertrouwen er is in data scientists binnen de organisatie. Zo’n document kan het vertrouwen hard maken, maar dan moet het wel concreter.

Hilde: hoe bepalen jullie normaal gesproken de algemene performance van jullie model?

Jelte: Als je onder de 80% zit ga je je toch echt vragen stellen over je systeem maar hangt ook af van het doel. Het is niet zo dat we van tevoren een waarde gaan bepalen. Hangt af van de context.

Hilde: Dat soort overwegingen (ratio false positives en false negatives) die kun je niet van tevoren vastleggen... Kan je dat voor ethische vraagstukken ook niet per case bekijken?

Jelte: ja dat kan, dan heb je een uitleg waarom je een bepaalde redenering acceptabel vindt.

Dennis: Werken met een baseline is handig.

Jelte: Inderdaad, uitleggen en motiveren waarom er voor die bepaalde acceptabele baseline gekozen werd.

Jeroen: Wat mij opviel in het document is dat de rol van het gezag een beetje buiten beeld is, ik zou het daar graag in willen terug zien. Er is nog geen doctrine die het gesprek goed regelt. Er zijn wel dingen op het gebied van risicosturing, maar voor gezichtsherkenning is het nog niet goed geregeld. Dat hangt af van de maatschappelijke en ethische aspecten.

Proportionaliteit wordt in de handreiking genoemd maar niet ingevuld: bijvoorbeeld bij gezichtsherkenning daar gaat uiteindelijk het bevoegd gezag over. Maar het is onduidelijk hoe de maatschappelijke en ethisch gevolgen meegenomen worden in het gesprek.

In het voorbeeld van rondtrekkende boevenbendes en profileren in Roermond wordt er nergens stilgestaan bij waarom zij met opzet hier naartoe komen. Er wordt niet stilgestaan bij de rijke context, bij non-discriminatie moet je juist daar wel bij stil staan, wanneer mag je wel discrimineren. Ook de term AI is heel breed. De casus in Roermond is rijker wanneer je het bekijkt vanuit een profilering insteek ipv een non-discriminatie.

De terminologie rammelt en hangt af van het publiek. Met de handreiking heb je een kans om dialoog te faciliteren voor specifieke rollen.

Merel: over het publiek: we kregen de indruk dat de data-scientist het niet alleen kunnen oplossen dus dat het meer een holistisch iets moet zijn waarbij je inderdaad het gesprek faciliteert tussen verschillende rollen. Met dat in gedachte, wat zijn dan makkelijke ingrepen om dat wat helderder te maken?

Jelte: Jeroen heeft heel valide punt mbt betrekking van gezag erbij. Ik heb te maken met mensen die gaan over het al dan niet implementeren van een systeem. Dus die mensen hebben zeker behoefte aan dat gesprek om geïnformeerder afwegingen te maken, nu vertrouwen ze veelal op ons.

Het is bijzonder lastig om een generiek document op te stellen dat toepasbaar is op verschillende domeinen.

Merel: Jeanine heb jij vanuit ethiek hoek nog dingen?

Jeanine: inderdaad goed om juiste vragen te stellen en kijken welk effect dat dan heeft.

Merel: Wat kan er beter aan de handreiking zodat we het gesprek tussen de verschillende rollen kunnen faciliteren? We hebben het gebaseerd op het CRISP-DM model, is dat handig om het zo in te delen? Dat hebben we gedaan om per stap vragen te kunnen stellen die relevant zijn met het idee het gesprek te faciliteren. Welke vragen zou je dan wanneer willen stellen? Wanneer zou dat gesprek op gang moeten komen?

Dennis: op zich qua model is het een goed model, werkelijkheid is uiteraard iets *messier* maar lijkt me goed om zo op te delen want het maakt de vragen behapbaar. Ik twijfel er wel aan of mensen het echt zo gaan beantwoorden in die volgorde.

Jelte: voor het gesprek faciliteren is het inderdaad handig om het zo in te delen

Merel: we hebben laatste 2 fases omgedraaid. Wat zijn jullie gedachten hierover?

Jelte: Evaluatie hoort achteraan maar moet eigenlijk de hele tijd gebeuren.

Dennis: het is inderdaad een iteratief proces

Jelte: voor *checks* is het wel handig dat er net voor implementatie nog een evaluatie plaatsvindt

Jelte: voor dit soort experimentele project hebben AI specialisten een behoorlijk ruim mandaat. Daar kan een heel discriminerend systeem uit voorkomen maar zolang we het niet inzetten is er geen probleem. Dan kunnen we wel aantonen dat het technisch mogelijk is en dan kan het OM nog zien wat ze er mee doen.

Merel: het verschil tussen implementatie en experimentatie, is dat iets om op te nemen in de leeswijzer?

Jelte: misschien is het voor andere domeinen helemaal niet interessant om niet op voorhand te experimenteren... misschien zijn er op andere domeinen striktere regels dan in het domein van de politie, zoals het medische domein

Jeroen: Het zou kunnen helpen om duidelijker te maken of je descriptief of prescriptief wil gebruiken. Het zou ook goed zijn om duidelijk te maken dat de verschillende stappen (CRISP-DM) niet per se chronologisch moeten worden uitgevoerd maar dat het eerder een representatie is van een logica die moet kloppen, dus als je bvb in stap X iets verandert moet de rest waarop die verandering gevolgen heeft ook nog kloppen.

Merel: Laten we even naar de eerste drie fases (probleemdefinitie, collectie en voorbereiding) kijken. Zijn er dingen waar jullie op blijven hangen? Goed? Overbodig? Wat anders? Waar op te letten? Vanuit het idee om het gesprek te faciliteren...

Jelte: impact en benchmarks, hoort dat niet bij evaluatie? Ik vind ze daar nu een beetje gek staan. Ik zou ze onder evaluatie scharen.

Dataopslag bij dataverzameling, weet ik niet..? Heeft dat met discriminatie te maken?

Nils: misschien bedoelen jullie met benchmarks wat de *state of the art* is?

Merel: ja inderdaad, een soort van succescriteria. Wanneer is het goed genoeg en kan je het concretiseren.

Nils: dat is goed maar misschien is een ander woord dan benchmarks dan wel wenselijk.

Hilde: Je hebt aan de ene kant de business succes criteria en aan de andere de technische criteria.

Merel: vooraf moet je nadenken over wat voor impact het systeem zal hebben op verschillende belanghebbenden. Misschien ontstaan er nieuwe categorieën door gebruik systeem etc.. impact gaat ook over het brede maatschappelijke aspect van het systeem.

Jelte: Benchmarks etc... komt voor mij bij modelering

Jeroen: vanuit crisp-filosofie zijn benchmarks bedoeld als bepaalde normen

Jelte: dat is niet iets wat je van tevoren gaat invullen maar eerder achteraf gaat verantwoorden

Jeroen: in de praktijk wordt er geëxperimenteerd en is er een bepaalde uitkomst... je gaat er mee naar opdrachtgever en die beslist dan of er mee aan de slag wordt gegaan. De rolverdeling tussen data scientist, en diens projectleider is een punt van aandacht. De accountability van de betrokken rollen kunnen er baat bij hebben om de concrete manier waarop op biases is of wordt getest, expliciet te maken en het daar met elkaar over te hebben. Zoals jullie hebben gedaan op slides 30-32 van de handreiking.

Merel: vooraf concretiseren wat de gewenste uitkomst is, is vooraf nog niet mogelijk maar gebeurt iteratief. Is dat wat jullie bedoelen? Maar je kan vooraf wel aannames bepalen die je meeneemt maar zonder het meteen hard vast te leggen maar wel zodat je er gedurende het gehele project iteratief over kan nadenken? Zo van: "Dit is waar we van uitgingen, maar is nu anders of niet..."

Dennis: dat klinkt wel hoe je het goed zou kunnen inrichten

Jelte: Dingen worden ook wel spontaan ingezet met harde noodzaak (in de context van een crisis situatie) waar ze oorspronkelijk niet voor gemaakt zijn. Dit is een mogelijkheid niet een gegeven, als het gaat om de inzet van AI modellen.

Merel: Voor die implementatie moet er juridisch gezien op voorhand DPIA en HRIA doen dus dat zijn wel dingen die vooraf moeten gebeuren.

Jelte: dat valt bij ons wel mee want volgens Wet Politiegegevens mogen politiedata worden gebruikt voor wetenschappelijk onderzoek mits toestemming van bevoegd gezag. Wij moeten binnen de WPG ons houden aan regels gegevensbescherming, die verschillen van de AVG.

Bij de inzet moet je opletten wat je gaat doen, maar daarvoor kan je eigenlijk experimenteren.

Jeroen: als onderzoekers bij TNO vallen wij wel onder de AVG.

Merel: dataverzameling en datavoorbereiding. Zijn daar punten waar jullie blijven hangen?

Jelte: Bij de politie hebben wij heel veel en heel weinig data. Wij hebben doorgaans data over mensen waarvan het vermoeden bestaat dat ze strafbare feiten zijn begaan. Wij kunnen niet zo maar extra informatie verzamelen. Het is ook historische data waar wij verder weinig aan kunnen doen. Dagboeken van agenten.

Merel: wat mis je dan concreet hier? Wat zou jou helpen hierin? Welke vragen moeten dan gesteld worden over historische data?

Jelte: Hele grote open vragen: “waarom pak je die data wel mee en die data niet?”

Merel: Hoe vermijd je momenteel het gebruik van problematische bias?

Jelte: ongewenste discriminatie is heel erg lastig te meten en contextafhankelijk. Op het moment gebruik ik synthetisch data om aan het einde te testen op bias. Op het einde van het spelletje haal ik er gevoelige data door en kijk dan waar het probleem ligt...

Merel: Hoe bepaal je welke categorieën er zijn en welke fairness test moeten worden toegepast?

Jelte: Dat zou ik graag uit bijvoorbeeld een handreiking zoals jullie die aan het schrijven zijn halen. Het gemis in het gesprek is het OM, die zouden daar ook iets over moeten zeggen aangezien zij eindverantwoordelijk zijn binnen de opsporing.

Merel: :Laten we kijken naar de laatste drie fases: modeltraining, implementatie en evaluatie: zijn hier dingen die missen?

Dennis: de vraag in hoeverre het model dat je gebruikt bepalend is voor bias is niet iedereen het over eens. Veel mensen zullen wijzen naar data. Er zitten vormen van bias in model maar om door te trekken naar non-discriminatie... daar ben ik niet zeker van.

Nils: dat is voor nu nog onderzoeksgebied in welke mate bias aan het model kan worden toegeschreven, dus dat is voorlopig nog erg onduidelijk.

Hilde: zou je hier fairness mitigation technieken in meenemen?

Dennis: dat hoort volgens mij de stap ‘evaluatie’

Hilde: er zijn technieken waarbij verschillende fairnessmetrics direct worden meegenomen. Zijn jullie daar bekend mee zijn?

Dennis: goed idee maar moet explicieter instaan dat dit het idee is.

Hilde: Ook belangrijke overweging is: hoe transparant moet je model zijn? Ook afhankelijk per toepassing. Wat denken jullie daarover?

Jelte: voorkomen dat het een aftiklijstje wordt: model is transparant, check. Er is geen bias, check. Dus dat moet je voorkomen maar wel nadenken over of de vraag gesteld is geweest. We kunnen niet alles hebben: een model dat EN transparant is EN performant EN...

Merel: goed punt, dat zijn afwegingen die gemaakt moeten worden.

Jelte: ja dat is een dialoog

Merel: Wanneer is het dan een goed moment om die dialoog te voeren?

Jelte: Na de oplevering van het eerste model, want daarvoor is het niet relevant. ben je een gebakken lucht aan het bespreken

Jeroen: Balans tussen transparantie en performance is wel belangrijk op voorhand te bespreken want dan kan je vooraf wel keuzes maken om bvb wel of niet een neural network (wat doorgaans niet transparant is) te kiezen...

Merel: fairness en verschillende criteria: welke fairness criteria en metrics is iets wat je in deze fase kan doen?

Dennis: Nee, dat zou duidelijker moeten zijn in probleemdefinitie.

Merel: wat voor gesprek zou je het dan op dit punt willen hebben?

Dennis: als die vereisten er zijn, welk model wordt dan gekozen en waarom?

Merel: nog over laatste 2 fases: evaluatie moet doorlopend, implementatie is iets anders dan experimenteren

Jelte: De visie die je hebt over de implementatie is heel belangrijk: hoelang zet je iets in en hoe evalueer je? Wanneer en waarom haal je het systeem uit implementatie?

7.7.3 Workshop 3

Workshop III - 9 december 2020 - 09.00-11.00

Aanwezig:

Mark Bovens (UU; WRR)
Rob van den Hoven van Genderen (VU)
Roos van Leeuwen (BZK)
Egge van der Poel (TIAS)
Bart van der Sloot (TiU)
Frank van Vonderen (VKA)
Yvette Wagenveld (TiU)
Johan Wolswinkel (TiU)

Algemene indruk

Van Hoven van Genderen: Structuur, design en daarmee de bedoeling is complex. Symbolen zijn soms eerder verwarrend dan verhelderen. Het is lang. Tot slot wordt er veel basiskennis verondersteld van de lezer.

Bovens: Plezierig om te lezen. Voorbeelden heel instructief. Er is een hele uitleg, deze was zeer verhelderend. Leuk vormgegeven. Het is wel een enorm pakket. Wat is precies je doelgroep? Heb je een of meerdere doelgroepen.

Van der Sloot: Aanvankelijke idee: leg juridische principes aan ‘techneuten’ uit. Steeds meer ontwikkeld naar: er is een team dat werkt aan AI-systeem (programmeur, data-analist, systeembouwer, vaak ook een jurist, etc.) – welke vragen moeten zij zich stellen gedurende het project. Wij zeggen dan ook niet ‘zus en zo moet het’, maar eerder vragen/uitgangspunten.

Bovens: Wellicht een idee om factsheets te maken: 1 a4 met kernpunten voor juristen; 1 a4 met kernpunten voor systeembouwers; etc.

Wolswinkel: Non-discriminatie is de kern, maar ook veel aanpalende rechtsgebieden/vakgebieden. Daardoor soms onduidelijkheid over de kern.

Van Vonderen: het stuk staat bol van de kennis. Qua vorm vind ik de samenvatting prettig lezen. Qua technische kennis herken ik veel. Over de compleetheid qua technische normen: compliment.

Van Leeuwen: [Ethische data-assistent](#) lijkt op deze handreiking. Deze handreiking zou kunnen worden gezien als uitwerking daarvan voor anti-discriminatie. Meer algemeen: verhoudt zich dit document tot de vele andere handreikingen/documenten die al bestaan? Dat zou explicieter duidelijk kunnen worden gemaakt.

Van der Poel: Ik heb veel geleerd van het document, zoals ex post discriminatietoets en ex ante zorgvuldigheidsbeginselen. Er zou wellicht nog grotere aandacht kunnen zijn voor toestemming/betrokkenheid betrokkenen, gedurende het hele proces.

Wolswinkel: Introductie is erg lang; terminologie niet altijd consistent: non-discriminatie, anti-discriminatie, discriminatierecht, etc.

Van Hoven van Genderen: Duidelijke uitleg van de terminologie is essentieel; wellicht een verklarende woordenlijst in het begin. Soms document te algemeen, soms te hard: bv dat deep learning/black box altijd verboden zou zijn. Dat klopt juridisch niet. Erg goed de voorbeelden in de drie sectoren. Wellicht in de inleiding nog duidelijk maken dat dit drie essentiële gebieden zijn voor de maatschappelijke impact van AI. Ik vond de verschillende samenlevingsdelen als voorbeeld worden gebruikt.

Bovens: Als het gericht is op juristen, dan moet de handreiking idd langer en preciezer, maar als het voor niet-juristen is bedoeld dat is dit juist prettig. Het gaat dan niet om exacte uitleg, maar om bewustwording van kernelementen en -vragen.

Van der Sloot: Dat laatste is primair de bedoeling.

Fase 1

Wolswinkel: Veel vragen; soms ook overlap. Plus hoe verhouden fase 1 en 4 zich precies tot elkaar.

Van Vonderen: Succescriteria: goed punt en snap ik, maar zoals het nu is geformuleerd retorisch. Ik beschouw dit als een succes omdat ik het als succes heb gedefinieerd. Anders formuleren.

Van Hoven van Genderen: Gevoelige gegevens essentieel; wellicht ook aandacht voor doelbinding? Toegegeven, wel een complex juridisch punt.

Wolswinkel: op pagina 26 refereer je aan een impact assessment. Assessment uit de AVG?

Bovens: Overweldigend. Je ziet al twintig o'tjes bij stap 1. Ik zou geïntimideerd worden. En dit is dan alleen maar fase 1. Is het mogelijk een zekere hiërarchie erin te brengen. Dus bijvoorbeeld dat je zegt: deze 2 of 3 zijn essentieel, deze 4 of 5 zijn belangrijk, en als je echt tijd hebt denk dan ook nog even na over deze en deze vragen. Geldt ook voor andere fasen.

Van der Poel: Belangrijk om het gesprek continue te voeren, aan het begin/voorkant, maar ook gedurende het hele proces. Dus niet kwestie van afvinken en weer door, maar steeds teruggaan als je aanpassingen hebt gedaan: en hoe zit het dan nu met deze vragen?

Fase 2

Van der Poel: Zie ook p. 29 over bias. Belangrijk om dit continue te bevragen. Zorgvuldigheidsbeginsel is hier essentieel. Ook hier kan weer voortschrijdend inzicht zijn. Tijdens analyses moet je jezelf sowieso afvragen of het klopt wat je ziet, of het onderscheid dat je ziet alleen in de data bestaat of ook in de echte wereld. Daar zijn hele boeken over geschreven. Het is belangrijk dat dit in gedachten blijft bij de ontwikkelaars blijft.

Wolswinkel: Je hebt aan het begin van de fasen drie belangrijke punten staan. Zijn dit de belangrijkste punten waar Bovens aan dacht?

Van der Sloot: Ja, dat was de bedoeling, maar dat kan nog helderder naar voren komen.

Wolfswinkel: er wordt op pagina 34 gesproken over Aequitas. Is dat een programma?

Van der Sloot: hier gaan we nog een keer naar kijken, ik weet ook niet precies wat dit is.

Van Leeuwen: Wellicht uitleggen wat verschillende typen bias zijn.

Van Hoven van Genderen: Behoeft een verklarende woordenlijst: wat is bias, wat zijn verschillende vormen van bias. Daarnaast zou filtering ook hier al kort kunnen worden genoemd.

Bovens: Juridische gedeelte is veel AVG meegenomen, dus als je nog iets wil schrappen dan zou je hiernaar kunnen kijken. Bijvoorbeeld O nummer 1, 2 en 3, is meer gericht op privacy en niet op bias. Hier zou je in kunnen perken.

Wolswinkel: Ik denk niet dat het erg is om te accentueren dat het gewicht op het gebied van de AVG/bias bij de ene fase groter is dan de andere.

Van Hoven van Genderen: Soms juist goed om meer gegevens te hebben, bijvoorbeeld om te controleren op bias/discriminatie. Natuurlijk wel beperkt tot wat noodzakelijk is.

Bovens: Gaat vaker bij projecten: better safe than sorry. Dus bij twijfel neem je principes en uitgangspunten mee, maar dat heeft consequenties voor de scherpheid en bruikbaarheid.

Van der Poel: De zorgvuldigheid, dat moet bespreekbaar gemaakt worden, en worden aangehouden als de kern waarmee wordt beziggehouden. Data-analyse is een creatief proces, je weet niet van tevoren wat er uitkomt of hoe je systeem er precies uit komt te zien. Er is ruimte voor verrassing nodig. Het is een balanceer act tussen controle en vertrouwen.

Van der Sloot: Staat zo'n handreiking dan niet heel ver weg van iemand die wil spelen met data en probeert een systeem te creëren?

Van der Poel: Nee want technici willen ook gewoon het beste doen. Het staat er niet ver vanaf, maar expliciete ruimte voor experiment is belangrijk. Belangrijk: dit document moet een levend document zijn; rekening houden met nieuwe ontwikkelingen, toepassingen, etc.

Fase 3

Wolswinkel: data voorbereiding lijkt alsof het gaat om vooraf aan de data. Misschien andere term gebruiken? Daarnaast: discriminatierecht gaat uit van verboden gronden, terwijl AI onderscheid kan maken op basis van allerlei gronden. Extra belang van 'welke grond dan ook' benadrukken. Welk onderscheid wordt er gemaakt en is dat gerechtvaardigd/wenselijk?

Van der Poel: Belangrijk daarbij is handelingsperspectief. Onderscheid kan belangrijk en wenselijk zijn, zoals bijvoorbeeld in de geneeskunde. Maar als het een ongelijke positie van mensen als gevolg heeft, dan is dat onwenselijk. Goed het voorbeeld dat in de handreiking

wordt gegeven: als je door AI winst boekt ten aanzien van groep A, probeer dan de winst die je haalt in te zetten ten goede van groep B.

Fase 4

Bovens: Dichtheid aan vragen groot. Vragen, voorbeelden, uitleg, etc. is allemaal vrij veel. Er zit een mix tussen generieke en specifieke aandachtspunten. Goed om expliciet te toetsen hoe het systeem presteert.

Van Hoven van Genderen: Wat is supervised learning?

Van der Poel: Geeft uitleg

Van Hoven van Genderen: Opnemen in verklarende woordenlijst.

Van Leeuwen: Effectiviteit wellicht ook: wat bedoel je daar precies mee.

Van der Sloot: Voordat we naar fasen 5 en 6 gaan. We hebben het CRISP-DM-model enigszins aangepast en de twee fasen omgedraaid. Wat vinden jullie daar van? >> Niemand heeft hier een expliciete mening over <<

Fase 5

Bovens: Cruciale fase. Als ik kijk naar de technische kant dan vond ik het moeilijk om de belangrijke dingen eruit te pikken.

Wolswinkel: Eens: cruciale fase. Soms worden uitgangspunten herhaald uit eerdere fasen: goede herhaling of overbodig? Belangrijk om nogmaals te bedenken wat doet een systeem precies: zelf besluiten nemen, informeren, iets anders.

Bovens: Goede punten; bij de technische kant kan het nog scherper.

Van Leeuwen: Organisatorisch gedeelte: als je je systeem gaat testen op een specifiek toepassing dan is het belangrijk om niet 'waar mogelijk' mensen te betrekken, maar altijd. Kies dus je test zo dat je weet op wie er getest wordt en dat zij daarmee akkoord zijn.

Fase 6

Bovens: Er zitten algemene evaluatievragen in, die natuurlijk altijd goed zijn; maar je krijgt meer scherpte door je specifiek te richten op AI.

Wolswinkel: In de rechtspraak wordt veel gekeken naar maatwerk. Je zou kunnen accentueren dat er bijvoorbeeld ook naar collectieve dossiervorming wordt gekeken en niet alleen naar individuele gevallen. Je ziet namelijk soms dat een regel op een individueel geval goed uitpakt, maar op een hele groep niet.

Van Hoven van Genderen: Oppassen met aanpassen van definitie fairnes dat dat niet geïnterpreteerd wordt als: we passen het zo aan dat ons systeem er nog beter uitkomt.

Van der Sloot: Eens; juist het tegengestelde de bedoeling. Als je fairnessdefinitie A hebt gekozen en je systeem presteert daar goed op, controleer dan ook nog eens op fairnessdefinitie B. Als het daar helemaal niet goed op presteert moet je nog weer eens heel goed onder de motorkap gaan kijken.

Van Vonderen: Op pagina 63 staat een vraag over hoe het systeem functioneert. Ik vond deze vraag niet heel scherp.

Van der Sloot: zouden we een conclusie moeten toevoegen?

Bovens: Nee, bij een rapport is dat logisch, bij een handreiking niet.

<<Anderen eens>>

Van der Sloot: Veel dank allen!