Samenvatting van het rapport *Non-Discriminatie by design*

Het rapport geeft een wetenschappelijk verslag van het onderzoeksproject, met daarin een beschrijving van de verschillende onderzoeksstappen, de bestudeerde literatuur, de gekozen methodologie, de uitkomsten en bevindingen en de interpretaties daarvan. Beoogd lezerspubliek zijn wetenschappers vanuit de diverse relevante disciplines (juristen, AI-specialisten, ethici, etc.) en beleidsmakers die achtergrondinformatie zoeken over hoe de handreiking tot stand is gekomen, waarom bepaalde keuzes zijn gemaakt en hoe die worden verantwoord. Naast dit onderzoeksrapport is een handreiking worden ontwikkeld, met een toelichting daarop. De handreiking en de toelichting is gericht op personen en organisaties die AI-systemen bouwen of in de praktijk toepassen en eventueel op opdrachtgevers die relevante criteria zoeken voor gunning van een project. De opzet van het rapport is als volgt.

Hoofdstuk 0 (Inleiding)		
Het inleidende hoofdstuk geeft een: - Beschrijving van de aanleiding voor het onderzoek - Beschrijving van de aard en het doel van het onderzoek - Analyse van de behoefte aan het onderzoek - Beschrijving van de gebruikte definities voor dit onderzoek - Weergave van de gekozen opzet en methodologie voor dit onderzoek		
Hoofdstuk 1 (Probleemanalyse)	**Hoofdstuk 2 (Analyse bestaande standaarden)**	**Hoofdstuk 3 (Mapping standaarden op problemen)**
Door literatuuronderzoek, aangevuld met de uitkomsten van twee enquêtes, wordt in kaart gebracht welke problemen er zijn beschreven met betrekking tot AI en non-discriminatie	Door literatuuronderzoek, aangevuld met de uitkomsten van de enquêtes, wordt in kaart gebracht welke standaarden t.a.v. non-discriminatie en AI er zijn ontwikkeld in wetgeving, jurisprudentie, richtsnoeren en literatuur	Door middel van drie testgroepen, aangevuld met de uitkomsten van de enquêtes, wordt nagegaan in hoeverre de bestaande standaarden voldoen en op welke punten aanvulling/ verandering gewenst is
Hoofdstuk 4 (Ontwikkeling concepthandreiking)	**Hoofdstuk 5 (Validatie concepthandreiking)**	**Hoofdstuk 6 (Finaliseren handreiking)**
Op basis van de eerste drie stappen zal het onderzoeksteam een eerste concept handreiking ontwikkelen	Dit concept wordt getest in drie workshops. Een aangaande de inzet van AI in de strafrechtketen en een tweede over de inzet van AI in de gezondheidszorg. Een derde workshop richt zich op een interdisciplinaire discussie tussen experts.	Op basis van stap 5 en input van externe experts wordt het concept aangepast, de toelichting bij de handreiking uitgewerkt voor de drie contexten en de handreiking vormgegeven.

Hoofdstuk 0 – Inleiding

Beschrijving van de aanleiding voor het onderzoek

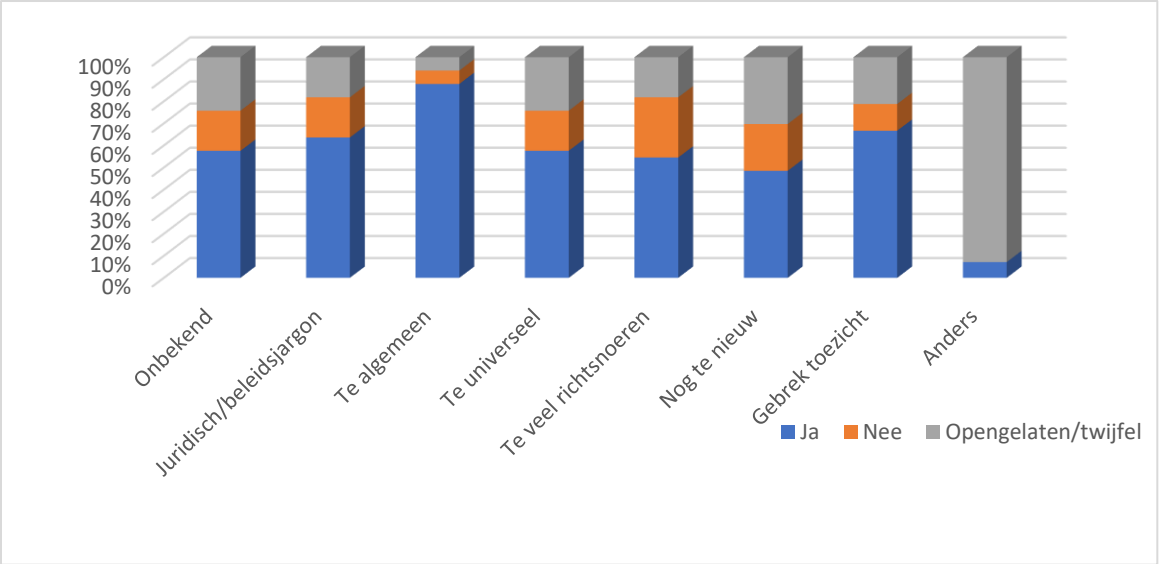
Algoritmische besluitvorming raakt steeds meer ingebed in organisatiestructuren; dergelijke besluitvorming kan grote (rechts)gevolgen hebben voor individu en samenleving. Met de komst van Big Data en de daarmee samenhangende opleving van Artificiële Intelligentie (AI) en Machine Learning (ML) vindt er 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 patronen te ontdekken; deze patronen kunnen vervolgens worden gebruikt in besluitvormingsprocessen.

Beschrijving aard en doel 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 2019 heeft het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties 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 te borgen in een samenleving waarin gebruik van AI steeds meer wijdverbreid zal raken. 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).

Analyse van de behoefte aan het onderzoek

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. Uitgangspunt voor dit rapport is de constatering dat de bestaande richtsnoeren nog onvoldoende effect hebben op de praktijk. 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, een vraag opgenomen. De resultaten laten zien dat bij alle mogelijke antwoordcategorieën vaker eens dan oneens is aangekruist. Duidelijk is ook 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.



Beschrijving van de gebruikte definities voor dit onderzoek

Met het oog op het voorkomen van discriminatie door AI-gebaseerde systemen is een aantal kenmerken van belang. Ten eerste leiden met name de nieuwe vormen van onzekerheid die lerende systemen met zich mee brengen tot zorgen over (onbedoelde) discriminatie. Als een taak vooraf niet volledig en expliciet gespecifieerd 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 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 te 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.

Methodologie voor dit onderzoek

Hoofdstuk 1	Hoofdstuk 2	Hoofdstuk 3
➢ Literatuuronderzoek ➢ Enquêtes	➢ Literatuuronderzoek ➢ Enquêtes	➢ 3 testgroepen ➢ Enquêtes
Hoofdstuk 4	**Hoofdstuk 5**	**Hoofdstuk 6**
➢ Eigen reflectie onderzoeksteam	➢ 3 Workshops	➢ Eigen reflectie onderzoeksteam ➢ Creative design

Hoofdstuk 1 – Probleemanalyse

Bouwstenen

Een aantal bouwstenen waarop de samenleving en data-gedreven werken als zodanig zijn gebaseerd veronderstellen categorisering en onderscheid op basis van die categorieën.

Taal	Data	Dataficatie
Taal is een samenbundeling van categorieën en begrippen die afgeleid zijn van de werkelijkheid; talige begrippen en categorieën zijn op hun beurt weer bepalend voor hoe mensen de wereld percipiëren.	Net als taal vormen data een representatie van de werkelijkheid. Wie data verzamelt, waarover, op welke manier, etc., is sterk bepalend voor hoe de data er uit komen te zien.	Personalisatie is een misleidend begrip in de AI-context. Er worden categorieën en correlaties ontdekt en vervolgens worden personen, objecten of fenomenen in die categorieën geplaatst.
Redlining	**Discriminatiegronden**	**Trade-off**
Alle data, categorieën en correlaties wijzen indirect naar de door de wet genoemde bijzondere categorieën (ras, geloof, etc.) – dit doen ze per definitie, de vraag is alleen hoe groot de afgeleide correlatie is.	De vraag is of de bestaande categorieën uit de anti-discriminatie- en gelijkebehandelingswetgeving en -jurisprudentie (ras, geloof, gender, etc.) nog de juiste zijn voor AI-besluitvorming.	AI draait om het maken van onderscheid. Het verbieden van voorspellingen op grond van bv ras or seksuele oriëntatie heeft ten gevolg dat voorspellingen minder accuraat worden.

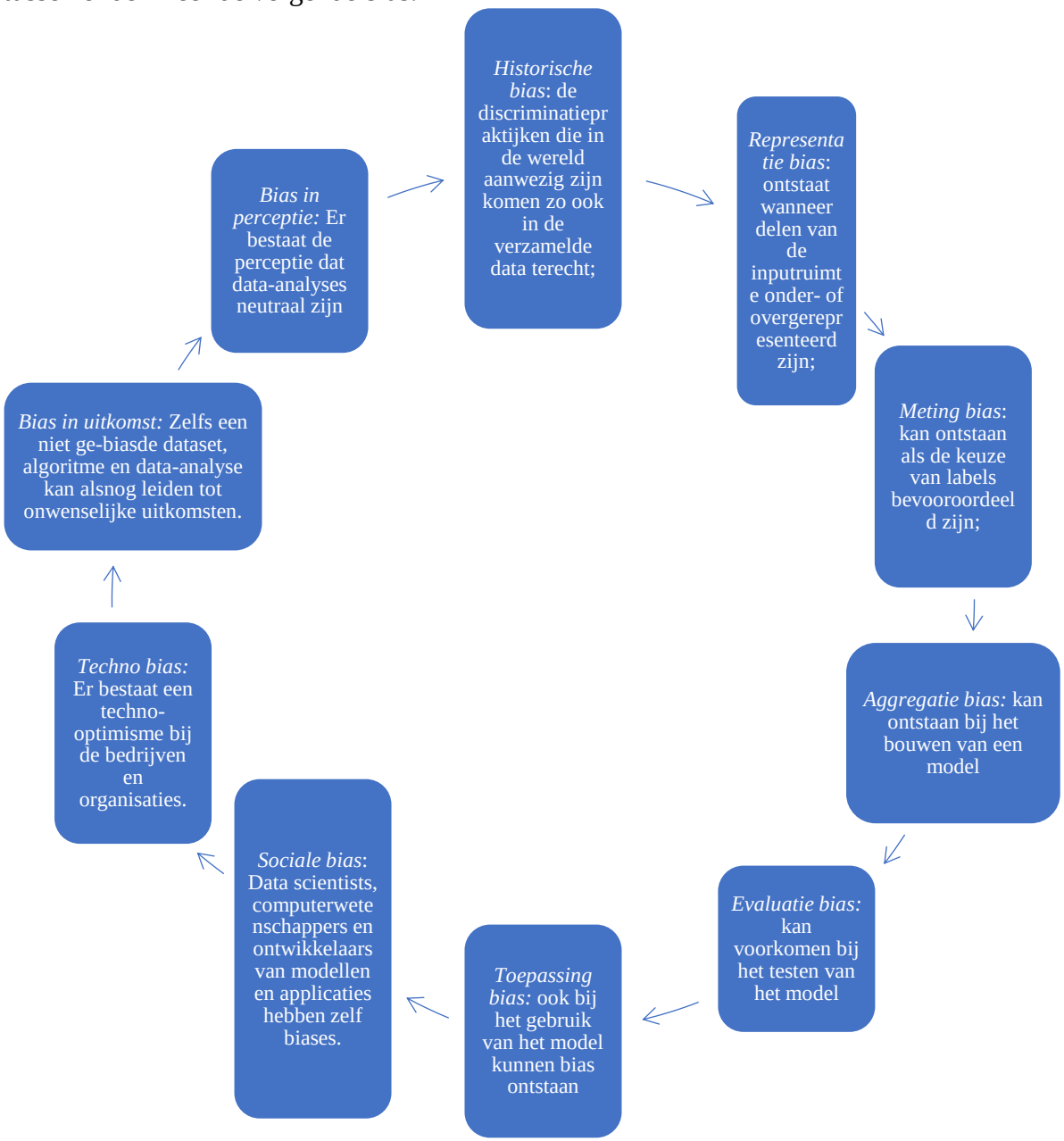
Explainability

Uitlegbaarheid van AI-systemen is nodig om besluitvormingsprocessen te kunnen evalueren, te controleren en te bevragen op eventuele discriminatoire uitkomsten. Er is een aantal obstakels:

- Ten eerste is er vaak geen zicht meer op wat zich tussen input en output afspeelt. 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 uitlegbaarheid te bevorderen. Deze oplossingen kennen echter beperkingen.
- 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 bepaalde instrumenten een uitleg aan technische experts. Dit houdt echter niet in dat deze technische 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.
- Ten slotte, een derde obstakel voor de uitlegbaarheid en transparantie van AI-systemen is dat er kosten mee gemoeid zijn.

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. Bias doen zich noodzakelijkerwijs voor bij machine learning systemen. Er valt een onderscheid te maken tussen onder meer de volgende bias.



Fairness

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 groepsgelijkheid (*group parity*) of als individuele gelijkheid (*individual parity)*. 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 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:

- *Portability Trap*: Het niet begrijpen hoe herbestemde algoritmische oplossingen die voor één sociale context zijn ontworpen, misplaatst zijn in een andere context;
- *Formalism Trap*: Het vangen van sociale concepten (bv eerlijkheid) in wiskundige modellen, terwijl die concepten procedureel, contextueel en betwistbaar zijn;
- *Ripple Effect Trap*: Het niet begrijpen hoe de integratie van technologie in een bestaand sociaal systeem het gedrag en de ingebedde waarden van het systeem verandert.

Technische knelpunten

Er bestaat een aantal additionele technische knelpunten ten aanzien van het voorkomen of verminderen van discriminatoire uitkomsten van AI.

- Bias-in-bias-out fenomeen: er zijn simpelweg meer gegevens over bepaalde groepen dan over andere. 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.
- Vertekening modelontwikkeling: De gegevens en fenomenen die modellen weerspiegelen kunnen vrij complex en heterogeen zijn. Traditionele ML-routines zijn gericht op het minimaliseren van de gemiddelde fout ten aanzien van de meerderheidspopulaties. 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.
- Formaliseren fairness: Het formaliseren van fairness in technische systemen is problematisch, enerzijds omdat er geen uniforme definitie van fairness bestaat, anderzijds omdat context-gebonden interpretaties van fairness zich niet lenen voor een AI systeem, waarbij om een algemeen model te bereiken doorgaans voor een vorm van abstractie en formalisering wordt gekozen.

Maatschappelijke knelpunten

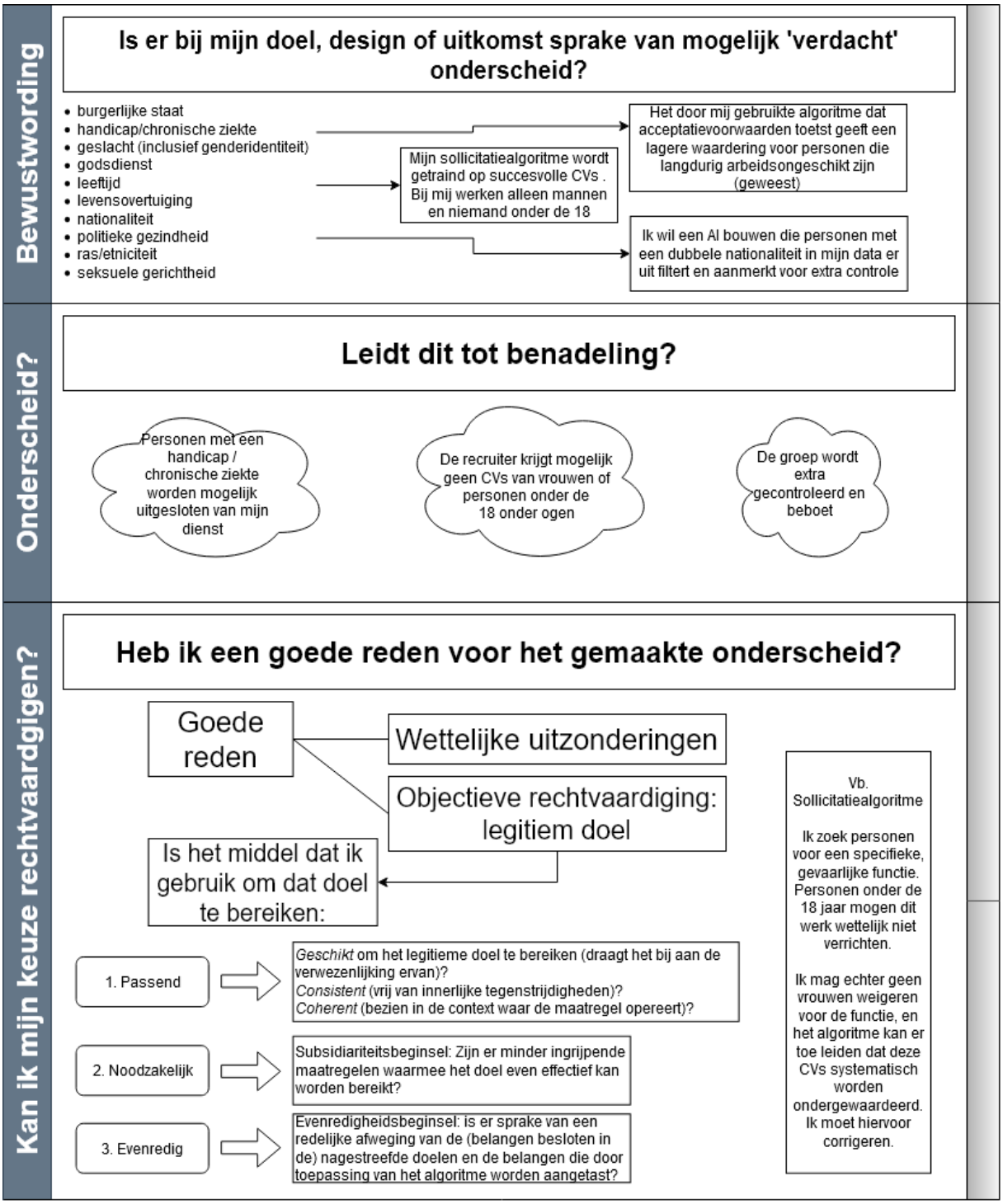
In beleidsrapporten en literatuur is tot slot een aantal bredere maatschappelijke knelpunten geïdentificeerd als het gaat om de gevaren van AI in relatie tot discriminatie en ongelijkheid:

- Effectiviteit en betrouwbaarheid: 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.
- Filter bubble en Mattheus effect: oude patronen kunnen in AI bepalend worden voor de mogelijkheden en kansen van groepen in de toekomst. Bestaande sociale ongelijkheid in de samenleving wordt door AI, zonder ingrijpen, herhaald en daarmee versterkt.
- Legitimiteit en vertrouwen: Interne audits en extern toezicht op AI laat te wensen over. Dit ondermijnt het vertrouwen van burgers in AI-besluitvormingsprocessen. Contextualiteit van data en besluiten is voor burgers intuïtief belangrijk voor acceptatie, terwijl AI juist de-contexualisering mogelijk maakt.

Hoofdstuk 2 – Bestaande Standaarden

Discriminatie

Het gelijkebehandelingsrecht is een complex rechtsgebied dat bestaat uit internationaal, EU en nationaal recht. Het toetsingskader 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 en een proportionaliteitanalyse. Een en ander kan als volgt worden samengevat.



Richtsnoeren, CoC en praktische tools

De afgelopen jaren hebben nationale en internationale bedrijven, overheden en organisaties een grote hoeveelheid aan ethische richtlijnen ontwikkeld voor AI. Er bestaan zes verschillende soorten aanpakken en strategieën voor het tegengaan van bias.

1. **Statische aanpakken en software toolkits**: 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. Verder onderscheid kan worden gemaakt tussen:
   - ***pre-processing* gericht op data**: methodes die zich met name richten zich op het samenstellen van evenwichtige datasets en het minimaliseren van bias in de dataset;
   - ***in-processing* gericht op algoritmes**: methodes gericht 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.
   - ***post-processing* gericht op ML modellen**: methodes 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.

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 in 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, model cards* en conformiteitsverklaringen. 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. Modelkaarten zijn korte documenten die worden afgegeven bij getrainde ML modellen en bieden een benchmarkevaluatie onder verschillende omstandigheden, zoals tussen 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. Modelkaarten geven ook de context weer waarin modellen moeten worden gebruikt, details van de prestatie-evaluatieprocedures en andere relevante informatie. 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.

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 *crowsourced 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. Een onderscheid kan bijvoorbeeld worden gemaakt tussen (1) institutionele, (2) softwarematige en (3) hardwarematige mechanismen. Relevante mechanismen voor de te ontwikkelen handleiding zijn:

- 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)

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. 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.

6. **Socio-technische methodes**: De beschreven methoden leggen veelal de nadruk leggen op het technische deel van de problematiek waar bias en discriminatie uit voortkomen. Ze hebben weinig oog voor de culturele, organisatorisch en politieke context waarin AI-algoritmen worden ontwikkeld en gebruikt. Ook op dit gebied zijn er echter inmiddels meerdere strategieën en methodes ontwikkeld. Een zesde categorie is daarom wenselijk. 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 verschillende initiatieven op het belang van diverse teams, inspraak van betrokkenen en aandacht voor de machtsverhoudingen en -structuren in het socio-technisch systeem.

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.

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.

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. Voor de ontwikkeling van de AI systeemprincipes is het belangrijk om met deze beperkingen en uitdagingen rekening te houden.

Statistische beginselen

Het verzamelen van gegevens en het gebruik van inzichten en analyses voor beleidsdoeleinden en besluitvorming zijn aan juridische banden gelegd, onder meer door het privacy- en gegevensbeschermingsrecht en het discriminatieverbod en recht op gelijke behandeling. Bij de analyse van de data is dat in mindere mate het geval, omdat er geen concrete besluiten worden genomen die een effect hebben op burgers en er niet noodzakelijkerwijs persoonsgegevens worden verwerkt (bijvoorbeeld als de verwerking geschiedt op geaggregeerd niveau en er naar algemene verbanden wordt gezocht). Daarom is het van belang om aansluiting te zoeken bij de diverse principes die zijn ontwikkeld voor statistiek en statistische analyses.

Privacy en gegevensbescherming

De AVG legt een aantal standaarden neer voor AI die gebruik maakt van persoonsgegevens:

Legitiem	Fair	Doel en doelbinding
Overheidsinstanties dienen een wettelijke bevoegdheid te hebben en een publiek belang te dienen. Private organisaties kunnen zich beroepen op toestemming of op een belang dat het belang van het datasubject overstijgt.	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, moet behoorlijk en fair zijn.	Data die voor het ene, specifieke doel zijn verzameld, mogen in principe slechts voor dat doel worden verwerkt. Een uitzondering op het doelbindingsprincipe geldt echter als het gaat om gegevensverwerking voor statistische doeleinden.
Dataminimsalisatie	**Datakwaliteit**	**Transparantie**
Slechts die data mogen worden verzameld die noodzakelijk zijn voor het specifieke doel; ze moeten weer worden verwijderd zodra het doel is bereikt.	Data moeten correct en up to date zijn; burgers hebben het recht om aanvullende gegevens aan te dragen.	De burger heeft recht op informatie over wie welke gegevens verwerkt en waarom en de logica van geautomatiseerde besluit-vorming en profiling.
Gevoelige gegevens	**Geautomatiseerde besluitvorming**	**Verantwoording**
Gegevens over ras, etniciteit, geloof, geaardheid en medische en strafrechtelijke gegevens mogen niet worden verwerkt, tenzij er bijvoorbeeld een groot publiek belang is of expliciete toestemming.	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.	Organisaties hebben de plicht om een register bij te houden van dataverwerkingen, data protection impact assessments te doen bij risicovolle projecten en data protection by design standaarden aan te nemen.

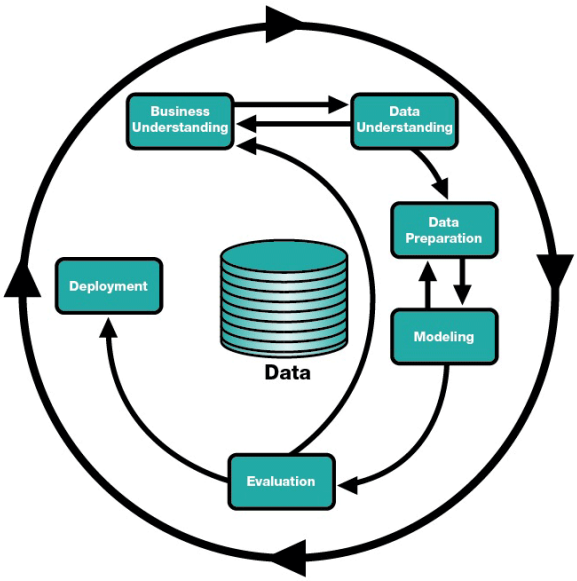
Er zijn procedurele standaarden voor overheidsbesluiten die de persoonlijke levenssfeer raken:

- 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.

Hoofdstuk 3 – Evaluatie bestaande standaarden en principes

Op basis van stappen 1 en 2 is een eerste overzicht van principes en standaarden gemaakt. 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 organisatorisch 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 de handreiking aanpassingen gedaan op dit model, om meer ruimte te geven voor permanente evaluatie.

	Juridisch	Organisatorisch	Technisch
Fase 1 - Probleemdefinitie: identificeren van het probleem/ het doel en die vertaling naar AI/ML taak	**Noodzakelijkheid:** 1. Is er een wezenlijke noodzaak om een AI project te beginnen? 2. Is het noodzakelijk en proportioneel om meer data te verzamelen en te verwerken voor het project? 3. Wat is de te verwachten effectiviteit van het AI project? **DPIA:** 1. Wat is impact op alle mensenrechten binnen een project? 2. Hoe kunnen de gevaren worden gemitigeerd? 3. Als er hoge risico's blijven bestaan na aanvullende maatregelen dan moet toestemming van de Autoriteit Persoonsgegevens worden gevraagd **Transparantie:** 1. Leg alle keuzes vast en beargumenteer de keuzes 2. Geef burgers zoveel mogelijk proactief informatie 3. Reageer op verzoeken op informatie	**Team:** 1. Het projectteam wordt gekozen op basis van competenties 2. Het projectteam is divers in culturele/gender/religieuze achtergronden 3. Het projectteam is divers in vakinhoudelijke achtergrond **Mandaat en middelen:** 1. Het team heeft de toegang tot de benodigde middelen 2. Het team heeft de benodigde bevoegdheden **Betrokkenheid:** 1. Betrokkenen of vertegenwoordigers van betrokken worden gedurende het hele proces meegenomen en gehoord 2. Krijg inzicht in de context van het probleem en de stakeholders 3. Betrek stakeholders in de definitie van het probleem, en het opstellen van requirements	**Systeemkeuzes:** 1. Beargumenteer de keuze voor AI/ML systeem in relatie tot probleem/doel 2. Beargumenteer de keuze voor statistiek in relatie tot probleem/doel 3. Beargumenteer de keuze voor methodiek van werken, bijvoorbeeld CRISP DM, in relatie tot probleem/doel 4. Formuleer de logica en het waarom van het AI systeem **Benchmarks:** 1. Formuleer target: wanneer is het AI systeem een succes? 2. Formuleer een acceptabele marge voor foutnegatieven en beargumenteer die marge 3. Formuleer een acceptabele marge voor foutpositieven en beargumenteer die marge **Ethics Canvas:** Voer de Ethics Canvas uit als ontwikkeld door het Open Data Institute
Fase 2 - Initiële dataverzameling en opslag	**Doel- en doelbinding:** 1. Leg een concreet doel vast 2. Verzamel slechts data die voor dat concrete doel noodzakelijk zijn en verwijder ze zodra het doel is bereikt 3. Gebruik de data niet voor andere doeleinden **Legitiem:** 1. Overheidsinstanties moeten een wettelijke grondslag hebben voor een AI project; dat project moet een publiek belang dienen 2. Private organisaties hebben voor het verzamelen van data toestemming van alle betrokkene nodig, tenzij het AI systeem een belang vertegenwoordigt dat al hun belangen overstijgt. **Veilig en vertrouwelijk:** 1. Zorg ervoor dat onbevoegden buiten de organisatie geen toegang tot de gegevens hebben	**Datalast:** Dataverzameling vindt gelijkelijk plaats over de bevolking; dubbelingen in datasets worden vermeden. **Objectieve compilatie:** 1. Welke bronnen worden gekozen en zijn die representatief? 2. Zit er een historische bias in de data en zo ja welke? 3. Welke categorieën labels worden gekozen voor de data en waarom? **Kwaliteitscontroles:** Organisaties en teams evalueren permanent hun werkzaamheden en procedures; waar nodig worden externe experts ingeschakeld	**Bias-in-bias-out:** Controleer op biases in de dataset, zeker wanneer de data van derden/openbare bronnen zijn verkregen **Samplingmethode:** Kies een sampling methode (bv random, stratified, oversampling) en beargumenteer waarom deze **Documentatiestandaarden**: AI algoritmen maken vaak gebruik van veel verschillende soorten data of deels getrainde modellen. Er dient een gestandaardiseerde manier te worden gekozen voor de documentatie van data en keuzes, bijvoorbeeld door *data sheets* en *model cards.*

	2. Zorg ervoor dat onbevoegden binnen de organisatie geen toegang tot de gegevens hebben 3. Zorg ervoor dat als onbevoegden toch toegang krijgen, de data onbruikbaar is		
Fase 3 - Data-analyse en voorbereiding	**Datakwaliteit:** 1. Controleer of data correct en up to date zijn 2. Corrigeer en update de gegevens waar nodig 3. Informeer burgers dat zij het recht hebben om aanvullende gegevens aan te dragen **Gevoelige gegevens:** 1. Controleer of er gegevens in de dataset zitten over ras, etniciteit, geloof, geaardheid en medische en strafrechtelijke gegevens 2. Controleer of deze gegevens direct zijn af te leiden uit de dataset of indirect 3. Beoordeel of het mogelijk is om deze gegevens te verwijderen uit de dataset, zonder significante nadelen **Gevoelige gegevens:** 1. Als gevoelige gegevens nodig zijn, leg dat doel vast 2. Beoordeel of er een legitieme grondslag is voor dat doel, zoals toestemming van alle betrokkenen of een groot publiek belang 3. Als de gegevens slechts worden bewaard voor het tegengaan van discriminatoire uitkomsten, leg dat vast	**Objectiviteit:** Datasets, analysemethode en beslissingen worden gekozen met het oog op objectiviteit; fouten worden vastgelegd en direct gecorrigeerd **Kwaliteit:** Methodes, procedures, definities en classificaties worden consistent toegepast; deze worden zo veel mogelijk gestandaardiseerd tussen organisaties, om vergelijkingen mogelijk te maken **Relevante expertise:** Alle vereiste en relevante expertise is aanwezig om de data te analyseren en te prepareren voor modellering; medewerkers krijgen voortdurende trainingen om hun kennis niveau op pijl te houden.	**Controle:** 1. Beschrijf de samenstelling van de dataset 2. Bestudeer de verdelingen in de dataset 3. Ga na of alle relevante groepen aanwezig zijn **Pre-processing:** Datasets worden evenwichtig samengesteld, onder meer door middel van: 1. Instance class modification 2. Instance selection 3. Instance weighting **Double check**: Controleer na de pre-processing methodes of de dataset evenwichtig en representatief; zo niet, voer nogmaals de diverse correctiemechanismen door
Fase 4 - Modelleren	**Statistische beginselen:** 1. Betrouwbaarheid: statistieken moeten zo getrouw, nauwkeurig en consistent mogelijk de realiteit meten en weergeven 2. Onpartijdigheid: statistieken worden op neutrale wijze ontwikkeld, geproduceerd en verspreid 3. Objectiviteit: statistieken moeten op systematische, betrouwbare en onbevooroordeelde wijze worden ontwikkeld, geproduceerd en verspreid; dit impliceert het gebruik van (context-afhankelijke) professionele en ethische standaarden	**Betrouwbaarheid/vergelijkbaarheid:** methodes voor dataverzameling en analyse worden vastgelegd, gedocumenteerd en openbaar gemaakt. Toegankelijkheid en universeel ontwerp staan voorop zodat iedereen de producten kan gebruiken, inclusief mensen met een handicap. Universeel-ontwerp principes moeten worden ingezet om zoveel mogelijk gebruikers te kunnen bedienen.	**In-processing:** algoritme wordt aangepast om gebiasde uitkomsten te minimaliseren, o.a. door: 1. Classification model adaption; 2. Regularisation/loss function s.t. constraints; 3. Latent fair classes. **Post-processing:** De bias wordt beperkt na het trainen van het classificatiemodel. White-box methodes passen het model aan; black-box methodes passen de voorspellingen aan. Voorbeelden van methodes:

	4. Vergelijkbaarheid: de toegepaste statistische concepten, meetinstrumenten en procedures worden vergeleken met en zoveel mogelijk geharmoniseerd tussen geografische gebieden en sectoren 5. Consistentie: het gebruik van concepten, classificaties en methoden is consistent door de tijd heen; afwijkingen en aanpassingen worden vastgelegd en beargumenteerd	**Accountability:** 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 de geïdentificeerde stakeholders **Participatie van betrokkenen:** Bij het modelleren worden burgers, betrokkenen en externen betrokken	1. Confidence/probability score corrections; 2. Promoting demoting boundary decisions; 3. Wrapping a fair classifier on top of a black-box baselearner. **Causaliteit:** Moet AI uitgaan van causaliteit, bijvoorbeeld omdat het voor besluitvorming wordt gebruikt, dan ligt deep learning niet voor de hand. Beargumenteer de keuze.
Fase 5 - Evaluatie (geselecteerde model evalueren aan de hand van succes criteria vastgesteld in stap 1 en een 'test set')	**Het recht op discriminatie:** De organisatie die AI inzet moet aantonen dat het systeem niet direct of indirect discrimineert en als het dat wel doet, dat dit legitiem en noodzakelijk is. **Recht op privacy:** Bij beslissingen die burgers raken: 1. Moeten burgers daarover worden geïnformeerd; 2. Daarbij betrokken worden; 3. De mogelijkheid hebben de beslissing aan te vechten; 4. De mogelijkheid hebben om bijstand te krijgen; 5. De beslissing moet fair en begrijpelijk zijn; 6. De beslissing moet op onafhankelijke wijze tot stand zijn gekomen **Gegevensbescherming:** 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	**Validering:** Statistische uitkomsten worden gevalideerd door: 1. Prior testing 2. Reviewing 3. Monitoring 4. Editing 5. Designing **Verbeteringen:** Fouten in data of modellen die al in de praktijk zijn toepast worden zo snel mogelijk aangepast en publiekelijk bekend gemaakt. **Universeel ontwerp:** Gebruik een toegankelijk en universeel ontwerp zodat iedereen de producten kan gebruiken, inclusief mensen met een handicap.	**Fairness:** Welke definitie van fairness wordt gekozen (bv *individual parity* of *group parity)* en waarom? **Anti-classification:** een model wordt eerlijk geacht te zijn als het beschermde kenmerken uitsluit bij het maken van een classificatie of voorspelling. Sommige anticlassificatiebenaderingen proberen ook proxy's voor beschermde kenmerken te identificeren en uit te sluiten. **Outcome / error parity:** Vergelijk hoe leden van verschillende beschermde groepen door het model worden behandeld. Met uitkomstpariteit is een model eerlijk als het gelijke aantallen positieve of negatieve uitkomsten geeft aan verschillende groepen.

Dit overzicht is vervolgens voorgelegd aan drie verschillende testgroepen. Testgroep 1 bestond uit juristen, testgroep 2 uit 'techneuten' en testgroep 3 bestond uit een combinatie van deelnemers met verschillende achtergronden. 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 belangrijkste leerpunten zijn:

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.

2. **Hou het zacht/open:** Een handreiking kan niet zeggen wat wel en niet mag, want dat bepaalt het recht niet zo eenduidig. Bovendien hebben duidelijke principes het nadeel dat eromheen kan worden gewerkt. Het gaat dus om het duidelijk maken van de kernuitgangspunten van het recht, zodat die tussen de oren zitten: bewustwording.

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

4. **Diversiteit:** Het team dat een AI-systeem uitwerkt en test moet divers zijn, zowel qua professionele als qua persoonlijke achtergrond. Het is belangrijk hier ook stakeholders bij de betrekken, het liefst bij alle fasen van het proces.

5. **Intersectionaliteit:** Het is belangrijk om zo veel mogelijk disciplines bij elkaar te brengen, ook omdat uiteindelijk alles met alles samenhangt.

6. **Domeinkennis:** AI-systeembouwers moeten altijd ook kennis hebben van het domein waarin het systeem wordt ingezet.

7. **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.

8. **Documentatie:** Documentatie van alle vragen en stappen is essentieel, omdat het vaak een iteratief proces betreft.

9. **Continue proces:** AI systemen blijven leren, dus is het belangrijk om te blijven testen en te evalueren op mogelijke bias en discriminatoire effecten.

10. **Handreiking nooit af:** Een statische handreiking slaat de plank mis, aangezien AI-systemen, anti-discriminatiejurisprudentie en ook de interactie daartussen permanent in ontwikkeling is.

Tot slot is nog aanvullend onderzoek gedaan naar jurisprudentie van het Europees Hof voor de Rechten van de Mens en het Hof van Justitie om aanvullende aanknopingspunten voor een handreiking te formuleren. Ook is een korte quickscan gemaakt van buitenlandse jurisprudentie ten aanzien van AI, om mogelijke inspiratie op te doen voor mogelijke benaderingen.

Hoofdstuk 4 – Ontwikkeling handreiking

Op basis van het overzicht van de standaarden, de input van de drie testgroepen en het aanvullende onderzoek is een eerste concepthandreiking gemaakt. Daarin is nog steeds een uitwerking te vinden in juridische, technische en organisatorische principes, maar is gekozen om per fase te beginnen met een aantal kernvragen. Deze kernvragen hebben als doel om het gesprek binnen de organisatie te kaderen en te zorgen dat alle relevante aspecten aan bod komen. Ook is een aanzet gedaan voor de uitwerking van die vragen in drie fictieve casus. De vragen die per fase aan bod komen zijn als volgt:

Fase 1 - Probleem definitie	**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?
Fase 2 - Data verzameling	**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?
Fase 3 - Data voorbereiding	**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?	**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	**Labelen** Hoe worden data gelabeld en waarom? Sluit dit aan bij hoe andere organisaties data labelen en

	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?	categorieën zit er 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?	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?
Fase 4 - Modelleren	**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?
Fase 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?
Fase 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?

Hoofdstuk 5 en 6 – Validatie en finaliseren standaarden

De eerste concepthandreiking is besproken met de klankbord van het Ministerie van Binnenlandse Zaken, waarin vertegenwoordigers van tal van Ministeries, overheidsorganisaties en semipublieke organisaties zitting namen. Daarnaast is de concepthandreiking ook voorgelegd in drie workshops. De belangrijkste aandachtspunten die daarin naar voren kwamen zijn:

1. **Start:** 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. Hier dient ook te worden nagedacht over de inbedding van non-discriminatie.

2. **Ownership:** 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. Ook is het van belang om de burger voortdurend bij het proces te betrekken. In de medische sector is daarom ook dynamische toestemming van belangrijk.

3. **Gevoelige data nodig:** Gevoelige gegevens zijn juist nodig om het AI systeem later te controleren op bias/discriminatie. Deze moeten dus niet worden verwijderd.

4. **Verschil AI en menselijke beslissing:** 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.

5. **Behoefte aan beknopte samenvatting:** Data-analisten zullen vaak geen groter document doorspitten voordat zij beginnen met een project. Het is een idee om samenvatting te maken van de handreiking en een 14tje met de belangrijkste punten.

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.

Colofon

Tekst en onderzoek door: Bart van der Sloot (Tilburg University), Esther Keymolen (Tilburg University), Merel Noorman (Tilburg University), Mykola Pechenizkiy (Eindhoven University of Technology), Hilde Weerts (Eindhoven University of Technology), Yvette Wagensveld (Tilburg University), Bram Visser (Vrije Universiteit Brussel) en in samenwerking met het College voor de Rechten van de Mens.

Opdrachtgever: Ministerie van Binnenlandse Zaken.