Wat is rechtvaardige AI?

Een kader voor het ontwikkelen en toepassen van algoritmes voor automatische besluitvorming

Tjerk Timan & Francisca Grommé[*]

Samenvatting

Vragen over eerlijkheid, rechtvaardigheid en gelijke behandeling (*fairness*) in kunstmatige intelligentie zijn een punt van aandacht in recente debatten over mogelijke negatieve gevolgen van de toepassing van artificial intelligence (AI) in de samenleving. Veel van deze zorgen zijn echter niet nieuw – ze komen voort uit maatschappelijke en politieke discussies over digitalisering van de samenleving in het algemeen. In de kern draaien ze om eerlijkheid, toegankelijkheid en exploiteerbaarheid van digitale diensten en big data: wie heeft de middelen, de expertise en de feitelijke gegevens om maximaal gebruik te maken van digitalisering, en ten koste van wie of wat? Automatisering van besluitvorming door middel van algoritmische besluitvorming (ADM) is een toepassing van AI die wordt gezien als bedreiging voor de rechtvaardigheid van beleid en bestuur, vooral omdat geautomatiseerde besluitvorming vormen van reeds bestaande ongelijkheid in de samenleving versterkt. Echter, het kan ook gezien worden als kans om bestaande oneerlijkheid juist te beteugelen door het vermogen van AI om objectievere en dus meer rechtvaardig beslissingen te nemen. Op basis van recente literatuur uit verschillende domeinen binnen de sociale wetenschappen stellen we een kader voor dat kan helpen bij de ontwikkeling en de toepassing van AI binnen de publieke sector.

Inleiding: waarom vormt AI een probleem als het gaat om rechtvaardigheid?

In de context van digitalisering van dienstverlening vanuit de overheid richting burgers en bedrijven zien we steeds vaker dat er geëxperimenteerd wordt met verschillende vormen van AI. AI-applicaties worden in toenemende mate ontwikkeld en toegepast als oplossing voor diverse vraagstukken binnen de publieke dienstverlening, van het efficiënter opereren intern tot het beter en persoonlijker maken van overheidsdiensten. Zo zien we de implementatie van *chatbots* op overheidswebsites die burgers sneller naar het juiste loket moeten leiden, of gebruik van AI op de achtergrond om problematische schulden te voorspellen en daarmee ook te voorkomen. Waar dit soort toepassingen een duidelijk doel nastreeft met

* Dr. Tjerk Timan is onderzoeker bij de afdeling Strategy, Analysis & Policy van de Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek (TNO). Dr. Francisca Grommé is assistant professor aan de Erasmus School of Social and Behavioural Sciences/Organizational Dynamics in a Digital Society, Erasmus Universiteit Rotterdam.

betrekking tot de taken die een overheid heeft, zien we andersoortige AI-toepassingen die aanleiding geven tot bezorgdheid, zoals de voorspelling van fraude (Van Veenstra e.a., 2019). Een berucht voorbeeld is SyRI (Systeem Risico Indicatie), een 'datamining'-applicatie die risicoprofielen genereert uit een brede selectie gegevens. Deze digitale dienst is ontstaan uit een samenwerkingsverband van onder andere de Belastingdienst, de Inspectie van het ministerie van Sociale Zaken en gemeenten. Er zijn zorgen geuit over de mogelijke vertekening die inherent is aan de datasets die worden gebruikt, als gevolg van de focus op lagere-inkomenswijken (Tweede Kamer, 2015[1]). Naar aanleiding van deze zorgen hebben burgergroeperingen een juridische procedure aangespannen bij de Nederlandse Staat; de zaak is in oktober 2019 voorgekomen.

We zien in dergelijke voorbeelden de contouren van een grote verzameling van problemen rondom eerlijkheid en gelijke behandeling als gevolg van automatische en in toenemende mate on-uitlegbare AI. Ondanks de verschillende publieke waarden die in het geding zijn, zoals beschreven in onder andere het kader van de AIHLEG-groep rondom ethiek in AI op Europees niveau,[2] zien we dat vragen rondom eerlijkheid, gelijke behandeling en rechtvaardigheid recent het meest prominent naar voren komen. In het Engels worden deze verschillende begrippen (eerlijkheid, gelijke behandeling en rechtvaardigheid) samengevat in de term *fairness*. Voor de leesbaarheid van dit document, en de EU-richtsnoer volgend,[3] zullen we de term 'rechtvaardigheid' toepassen als vertaling van *fairness*.
Als het gaat om de vraag waarom rechtvaardige AI een probleem is, zullen we moeten kijken naar wie er op dit moment de touwtjes in handen heeft als het gaat om het bepalen van de toekomst van AI. Op dit moment zijn dat de technologiebedrijven die we kennen uit Silicon Valley of China. Veel van deze bedrijven hebben expliciete uitspraken gedaan waarin onder andere wordt gezegd dat ze 'geen kwaad willen doen'.[4] Toch hebben we gezien dat zodra dergelijke bedrijven en platformen internationaal gaan opereren en opschalen, hun eerste ideeën en kaders van wat rechtvaardigheid is, of wat gelijkheidsprincipes zijn, steeds moeilijker te handhaven zijn. Wanneer marktkrachten worden geconfronteerd met ethische principes, hebben de eerste de overhand, zo lijkt het. Of het nu gaat om de fraude bij de recente verkiezingen in de Verenigde Staten en Brexit, of om het gebrek aan ingrijpen van socialemediaplatforms bij door fake nieuws veroorzaakte wreedheden,[5] veel op algoritmes gebaseerde socialemediaplatforms (zijnde de belangrijkste ontwikkelaars en gebruikers van AI zelf) zijn niet bereid of in staat om complexe sociaal-politieke processen te begrijpen. Ook hebben ze tot nu toe weinig actie ondernomen om verantwoordelijkheid te aanvaarden voor door algoritmes aangerichte schade. In dat opzicht beginnen we pas nu de volledige impact en gevolgen te zien van big data en algoritmische besluitvorming in de maatschappij. Het opschrijven van een handjevol AI-principes lijkt een magere reactie om de onrechtvaardigheid veroorzaakt door big data, AI en ADM te beteugelen. Zelfs het installeren van onafhankelijke ethische commissies die daadwerkelijk beslissingsbevoegdheid hebben, lijkt te veel gevraagd: een recente poging dit te doen leverde veel kritiek op met betrekking tot de eerlijkheid van de panelformatie an sich,[6] en de beweegredenen om zo'n commissie te installeren werden

Tabel 1 Typen ongelijkheid en uitdagingen vanuit verschillende onderzoeksvelden

Onderzoeksveld	Rechtvaardigheidstype	Geïdentificeerde uitdagingen
Sociaal/bestuurskundig	Procedureel, distributief	Gelijke weergave in inzicht in gegevens (informatiesymmetrie), toegang tot expertise en AI-instrumenten (software, algoritmes), democratisering van AI, intentie van AI & ADM-systeem of -dienst, indirecte effecten en meting van impact op lange termijn
Juridisch/ethisch	Procedureel	Vooroordelen in trainingsgegevens, gebrek aan betwistbaarheid door de betrokkene, ondoorzichtigheid van beslissingen, ongelijke toegang tot informatie, (grenzen aan) het inbouwen van de wet in algoritmische systemen, (grenzen aan) privacy en antidiscriminatiewetgeving om systemische ongelijkheid aan te vechten
Informatica/ontwerpgericht	Procedureel	Selectie van datasets, afweging van nauwkeurigheid versus vertekening van data, featureselectie, onomkeerbaarheid in geaggregeerde data, 'aanpasbaarheid' van het resultaat, visualisatie van – en interactie met – de resultaten

afgedaan als ethisch witwassen.[7] Een recente kritiek stelt dat er door de 'techies' simpelweg niet wordt geluisterd naar de lessen vanuit de sociale wetenschappen.[8] Een eerste stap om de vele niveaus van rechtvaardigheid met betrekking tot AI en algoritmische besluitvorming te ontwarren, is het probleem te beschouwen door een multidisciplinaire bril. Een overzicht van de literatuur waar wij naar keken, die in basis vaak al multidisciplinair is, hebben we in zeer grove categorieën weergegevens in tabel 1.

Rechtvaardigheid als publieke waarde en bestuurlijke uitdaging

Eerlijkheid en gelijkheid zijn in de eerste plaats subjectieve waarden. Hoewel ze kunnen worden gemeten als individuele ervaring, zijn ze geen geheel individuele aangelegenheid. Dit kan gemakkelijk worden begrepen uit de recente openbare debatten die in heel Europa worden gevoerd. Er is een groeiend gevoel dat samenlevingen ongelijker worden in termen van inkomensverdeling en dat structurele ongelijkheden langs de lijnen van opleidingsniveau, etniciteit en klasse voorkomen dat maatschappelijke groepen grote uitdagingen van de 21ste eeuw, zoals bijvoorbeeld veranderende industrieën, arbeidsmarkten en ecologische veranderingen, aankunnen (Piketty, 2015; Joint Research Centre, 2017). Rechtvaardigheid is dus ook een collectief gedeeld gevoel over hoe kansen moeten worden verdeeld. In democratieën zijn noties van wat rechtvaardig is, vastgelegd in een grondwet – die op haar beurt weer is uitgewerkt in andere regelgeving die alle onderdelen van het sociale en economische leven bestrijkt. Deze regelgeving bepaalt vervolgens weer gedragscodes in verschillende sectoren die ons dagelijks

leven beïnvloeden (bijvoorbeeld gedragscodes voor de werving van nieuwe werknemers). Ideeën omtrent rechtvaardigheid verschillen per maatschappij of domein en tussen politiek en praktijk. Zo beschouwen Nederland en de Verenigde Staten beide discriminatie op grond van ras als onrechtvaardig; dit staat in de Grondwet. Echter, het Nederlandse welvaartsbeleid legt veel meer dan in de VS de nadruk op inkomensgelijkheid als rechtvaardigheidsbeginsel.

Met betrekking tot het beleid en de besluitvorming van de overheid en maatschappelijke organisaties worden meestal twee dimensies onderscheiden waarmee rechtvaardigheid kan worden beoordeeld: distributieve rechtvaardigheid en procedurele rechtvaardigheid. Omdat deze begrippen in de informaticaliteratuur anders worden gebruikt (vergelijk Grgić-Hlača e.a., 2018), bespreken we ze eerst om aan te tonen dat ze verder reiken dan duidelijk omschreven juridische vragen of kwesties. In het geval van distributieve (*verdelings*)rechtvaardigheid gaat het om de uitkomsten van beleid en beslissingen: de verdeling van inkomsten, middelen en kansen. Kwesties van rechtvaardige verdeling staan hoog op de beleidsagenda's, zoals het probleem van ongelijke toegang tot onderwijs (onder andere bepaald door de opleiding van de ouders) en ongelijke kansen op werk. Kansen worden onder andere bepaald aan de hand van (onuitgesproken) sociale klasse, etniciteit en geografie. Wat betreft distributieve rechtvaardigheid zijn verschillende *scholen* te onderscheiden, waaronder radicale gelijkheid, beperkte ongelijkheid (Rawls), verdienstelijkheid (meritocratie), gelijke kansen, collectief welzijn en libertarisme. Bovendien weten we dat de perceptie van rechtvaardigheid wordt beïnvloed door het eigen inkomensniveau, de maatschappelijke ongelijkheid, het welzijnsbeleid en het culturele karakter van een gemeenschap (zie Joint Research Centre, 2017). Procedurele rechtvaardigheid heeft betrekking op de kenmerken van het proces dat leidt tot een beslissing of beleid. Wat als rechtvaardig wordt ervaren, is niet alleen gebaseerd op de uitkomsten van een beleidsbeslissing, maar ook op de toegepaste procedures. Over het algemeen wordt procedurele rechtvaardigheid in nauwe zin en in brede zin beschouwd. De nauwe zin verwijst naar de kenmerken van de procedure, waaronder neutraliteit, transparantie en de mogelijkheid om in te grijpen. De brede zin omvat de ervaring van een besluitvormingsproces op basis van interactieve aspecten, zoals de beschikbaarheid van informatie (Van Velthoven, 2011). Leventhal, Karuza en Fry (1980) stellen de volgende – algemeen aanvaarde – dimensies van interactieve procesrechtvaardigheid voor: correctheid, consistentie, nauwkeurigheid, ethiek, representativiteit en vooringenomenheid. Net als bij distributieve rechtvaardigheid zijn de beginselen van procedurele rechtvaardigheid contextueel bepaald, zodat elk van deze aspecten verschillende vormen kan aannemen, afhankelijk van het domein, het tijdstip en de plaats. Sommige aspecten van procedurele rechtvaardigheid zijn algemeen aanvaard en vastgelegd in de wet, terwijl andere onderwerp zijn van discussie (denk aan het voorbeeld van positieve discriminatie bij werving- en selectiebeleid). Waar veel bestuurlijke processen en wetgeving zoals de Algemene Verordening Persoonsgegevens (AVG) gebaseerd zijn op procedurele rechtvaardigheid, is het de vraag of we in de context van AI niet veel meer moeten kijken naar verdelingsrechtvaardigheid. Hiervoor is wel een breder begrip van ethische en praktische overwegingen nodig als vertrekpunten voor beleid.

(On)rechtvaardigheid beschouwd vanuit ethische kaders

Om beter te begrijpen hoe AI, en meer in het bijzonder ADM, ingrijpt op rechtvaardigheid moeten we weten hoe rechtvaardigheid en AI zich tot elkaar verhouden. Vanuit een abstract oogpunt kan het gebruik van AI worden 'getoetst' aan een aantal principes. Floridi e.a. (2018) vergeleken een hele reeks terugkerende principes in de recente literatuur over AI, ethiek en rechtvaardigheid: 'Overal in de documenten heeft rechtvaardigheid betrekking op:

a het gebruik van AI om misstanden uit het verleden te corrigeren, zoals het uitbannen van discriminatie en oneerlijke behandeling;
b ervoor zorgen dat het gebruik van AI voordelen creëert die worden gedeeld (of ten minste kunnen worden gedeeld); en
c het voorkomen van het ontstaan van nieuwe problematiek, zoals het ondermijnen van bestaande sociale structuren.'

Onrechtvaardigheid in AI en ADM kunnen enerzijds worden benaderd vanuit een technologie-agnostisch gezichtspunt, wat betekent dat AI slechts bestaande maatschappelijke processen expliciet maakt. AI vergroot en versterkt in die zin reeds bestaande onrechtvaardigheid, niet in de laatste plaats als gevolg van de digitalisering van de samenleving. Debatten over rechtvaardigheid rond digitalisering in het algemeen draaien onder meer om onderwerpen als toegang tot big data en *digital divides* (Van Dijk & Hacker, 2003), de toegankelijkheid en controleerbaarheid van databronnen en digitale tools, de kracht en macht van geopolitieke digitale infrastructuren en de opkomst van 'splinternetten' (Malcomson, 2016). Bovendien zijn er uitdagingen op het gebied van rechtvaardigheid te vinden in (het gebrek aan) expertise om digitalisering te benutten en er voordeel uit te halen, en inbreuken op privacy en autonomie als gevolg van data en AI. Deze raken bovenmatig vaak de armen (bijvoorbeeld Madden e.a., 2017); zij die zich niet of in veel mindere mate digitaal kunnen weren. Aan de andere kant kan de technische logica van AI-systemen nieuwe vormen van onrechtvaardigheid opleveren die losstaan van bestaande sociaal-maatschappelijke scheidslijnen, en die een specifieke aanpak vereisen. Hoewel veel waarschuwingen en risico's betrekking hebben op de gevolgen van AI in het algemeen (AGI/ASI[9]), zijn toepassingen van geautomatiseerde besluitvorming (ADM) in bepaalde sectoren of delen van de samenleving misschien meer urgent. Beruchte voorbeelden zijn gezichtsherkenning op basis van bevooroordeelde trainingsgegevens (vergelijk Magnet, 2011), of oneerlijke behandeling en discriminatie in geautomatiseerde rekruteringssoftware. Dergelijke toepassingen lijken op het eerste gezicht alleen maar te benadrukken dat impliciete discriminatie en oneerlijke behandeling helaas nog steeds wijdverspreid in de samenleving aanwezig zijn.[10] Ze werken echter ook als een vergrootglas, al was het maar door het normaliseren van ongelijke behandeling door middel van de ogenschijnlijke objectiviteit van big data en algoritmes (zie Van Dijck, 2014). Dit roept de vraag op welke ethische en wettelijke grenzen moeten worden gesteld aan het gebruik van ADM, met name in gevoelige contexten zoals functiescreening of politieonderzoek.

Als we kijken naar de transparantie van dergelijke systemen bij de overheid, dan blijkt dat technische specificaties van applicaties en diensten, bijvoorbeeld SyRI, vaak niet openbaar worden gemaakt. Dit maakt het lastig om vanuit een technologisch oogpunt te analyseren of een systeem rechtvaardig is of niet. We kunnen echter wel het een en ander distilleren vanuit andere reeds bestaande voorbeelden waarbij de uitkomsten bekend en bediscussieerd zijn. Deze discussies gaan vaak over de aannames, vooringenomenheid en vooroordelen (hierna geschaard onder de Engelse term *bias*) in data die ten grondslag lag aan een digitale dienst. Onduidelijkheid over de oorsprong en de kwaliteit van data in combinatie met ontransparante ontwerpkeuzes rondom weegfactoren in AI-modellen kan leiden tot ingebouwde en geautomatiseerde ongelijkheid en discriminatie. De COMPAS-casus is een van de bekendste gevallen: een Amerikaans systeem dat recidive voorspelt en waarvan is aangetoond dat het via indirecte variabelen racistisch is (Larson & Angwin, 2016). Een ander recent voorbeeld dat de publieke opinie zorgen baart, is het Netflix-aanbevelingssysteem dat racistische getinte inhoud leek aan te bevelen op basis van voorspellingen van gebruikers (Iqbal, 2018). Soortgelijke bezwaren zijn geuit tegen applicaties die worden ingezet om kandidaten voor een bepaalde positie te selecteren en te screenen (Bogen & Rieke, 2018; Zuiderveen Borgesius, 2018; Krishnakumar, 2019). Dergelijke voorbeelden hebben de publieke en academische aandacht voor 'ethische AI' aangewakkerd, een uitdaging die een grote verscheidenheid aan disciplines doorkruist, waaronder informatica ('datascience'), juridisch onderzoek, mediastudies, bestuurskunde en wetenschaps- en technologiestudies.[11] De roep om een 'verantwoorde datawetenschap' op basis van de waarden eerlijkheid, nauwkeurigheid, vertrouwelijkheid en transparantie (FACT) is door Europese onderzoeksinstituten opgepakt (Van der Aalst, Bichler & Heinzl, 2017). Veel instituten en organisaties hebben kaders voor verantwoorde datawetenschap en AI ontwikkeld. Zo heeft TNO verantwoordelijke AI gedefinieerd als 'systemen die ethische beslissingen kunnen nemen en uitleggen' (TNO, 2018). Een veelgebruikt startpunt hierbij is het door de *digital society* voorgestelde kader,[12] waarin zij verschillende aspecten van de bestaande FACT-en fair-principes combineren om de volgende lijst van aandachtspunten voor te stellen:

- betrouwbare en te vertrouwen benaderingen voor data-engineering, databeheer, datawetenschap, *machine-learning* en AI, interoperabiliteit en herbruikbaarheid;
- mens-centrische oplossingen voor de digitale samenleving;
- algoritmische rechtvaardigheid, transparantie en uitlegbaarheid;
- sociale, juridische en ethische aspecten van verantwoorde datawetenschap.

Rechtvaardigheid is vaak een aspect van deze kaders. In de FACT-beginselen wordt rechtvaardigheid omschreven als 'data science without prejudice – how to avoid unfair conclusions even if they are true'.[13] Rechtvaardigheid gaat over het risico dat de uitkomsten van algoritmes vooroordelen bevatten, maar ook over de balans tussen het gelijk behandelen van elk item of onderwerp versus het rechtvaardig behandelen ervan (dit laatste betreft situaties of beslissingen die een 'oneerlijke' lokale situatie vereisen om een grotere mate van eerlijkheid op grotere

schaal te bereiken). Hoewel er grote vooruitgang is geboekt op het gebied van databeheer en algoritmes, alsmede op het gebied van strategieën om bias te elimineren, wijzen onderzoekers ook op de beperkingen van deze benaderingen wanneer datasets inherent in onbalans zijn als gevolg van maatschappelijke vertekening of misrepresentatie tijdens de dataverzameling (Mittelstadt e.a., 2016). Sommige filosofische en juridische auteurs stellen dat rechtvaardige AI om deze reden onmogelijk is. Toch worden er op dit moment AI-gebaseerde diensten ontwikkeld en geïmplementeerd, wat vragen oproept over waar de grootste uitdagingen – en oplossingsrichtingen – liggen.

Technologische ontwerpuitdagingen en bestaande raamwerken rondom het inbouwen van rechtvaardigheid

Onrechtvaardigheid als gevolg van AI kan zich manifesteren in databronnen die gebruikt worden voor het trainen van algoritmes (zogenaamde *trainingsdata*), de gekozen kenmerken en de bijbehorende labels en/of de aannames die binnen het algoritmische ontwerp worden gemaakt (wie stelt de prestatiewaarden vast, worden er statistische back-upcontroles gedaan op de uitkomsten, et cetera). Chavalarias en Ioannidis (2010) geven 'een taxonomie van 235 *biases* die kunnen ontstaan bij het uitvoeren van onderzoek met grote datasets'. Deze biases verwijzen grotendeels naar mentale processen of gedrag van programmeurs. Vooringenomenheid speelt ook een rol bij prestatieproblemen die direct samenhangen met de balans tussen vaak uiteenlopende databronnen die gebruikt worden om een hypothese te toetsen of om algoritmes te trainen (Chawla e.a., 2002) (geparafraseerd in Howard & Borenstein, 2018). Kortom, bias kan optreden bij vele stappen in het onderzoeks- en ontwerpproces van AI-gebaseerde toepassingen: van selectie van databronnen tot en met de interpretatie van het toepassingsgebied en aannames over wat gebruikers of burgers willen, wat ze nodig hebben of waar ze bezorgd over zijn. Dit laatste is niets nieuws, maar is een bekende uitdaging bij het ontwerpen van producten en interfaces.[14] Een nieuw element dat algoritmes en ADM hieraan toevoegen, is dat interfaces ondoorzichtiger worden en dat alternatieve manieren om dezelfde dienst of toepassing te leveren niet altijd mogelijk zijn. De paradox van gepersonaliseerde digitale dienstverlening zowel in de private als in de publieke sector is dat het juist leidt tot singuliere oplossingen die weinig ruimte laten voor alternatieven. Wanneer algoritmes worden toegepast in besluitvormingsprocessen die rechtstreeks van invloed zijn op burgers of eindgebruikers, worden zij leidend voorwerp van een geautomatiseerd proces waarvan de uitkomsten steeds moeilijker te verklaren zijn. Hierdoor neemt de mogelijkheid om de door dat systeem genomen beslissingen te beïnvloeden, of er bezwaar tegen te maken, af. Een belangrijk risico van ADM is dan ook dat het kan leiden tot bevooroordeelde of anderszins schadelijke uitkomsten die het gevolg zijn van schijnbaar objectieve gegevensmanipulaties en -combinaties: negatieve impact die een ontwikkelaar van een digitale dienst niet noodzakelijkerwijs kan overzien of waarop hij kan anticiperen. In een recente studie beschrijft Hacker (2018) twee soorten vooroordelen die in AI-toepassingen sluipen, met tegengestelde effecten

op de nauwkeurigheid en bruikbaarheid van een model: 'Aan de ene kant, als de trainingsgegevens bevooroordeeld zijn, zullen *machine learning* algoritmes deze vooroordelen in hun besluitvormingsmodellen opnemen. Hier verbetert het verminderen van bias de voorspellende nauwkeurigheid. Aan de andere kant, in gevallen van indirecte (proxy) discriminatie, zijn kenmerken die het algoritme oppikt significant gecorreleerd met het lidmaatschap van een beschermde groep. De vermindering van discriminatie impliceert dan een afname van de voorspellende nauwkeurigheid' (Hacker, 2018, 34).
Sommigen informatici beargumenteren, in tegenstelling tot de filosofen, dat we AI daadwerkelijk kunnen gebruiken om de oneerlijkheid te verminderen. Zo stelt Chayes (2017) dat 'met een zorgvuldig algoritme-ontwerp computers eerlijker kunnen zijn dan typische menselijke besluitvormers, ondanks de bevooroordeelde trainingsgegevens', en hebben anderen erop gewezen dat AI zou kunnen helpen om ervoor te zorgen dat historische vooroordelen de veranderende opvattingen over sociale rechtvaardigheid niet overschaduwen (Howard & Borenstein, 2018, 1526). Hacker waarschuwt echter voor een dergelijk solutionisme door erop te wijzen dat op de keuze om bepaalde mechanismen, benaderingen en instrumenten in te zetten om oneerlijkheid aan te pakken op zichzelf al een normatieve beslissing is: 'Hoewel algoritmen kunnen worden gebruikt om de afweging tussen verschillende rechtvaardigheidscriteria te formaliseren en in verschillende mate af te dwingen, blijft de selectie van de adequate maatstaf een zeer normatieve en uitdagende vraag. Ongetwijfeld zullen verschillende contexten verschillende maatstaven vereisen' (p. 211).

Van ethische kaders naar de praktijk: hoe en waar kan onrechtvaardige AI herkend en aangepakt worden?

Aangezien ADM en andere vormen van automatisering die onder de generieke paraplu van 'AI' vallen steeds vaker worden toegepast, zijn het de minder voor de hand liggende vormen van gegevensmanipulatie waarover we ons zorgen moeten maken als het gaat om rechtvaardige AI. Het ligt niet voor de hand om precies aan te geven waar en hoe oneerlijkheid plaatsvindt binnen een algoritmisch ontwerp, hoewel hiertoe wel pogingen zijn gedaan. Studies hebben bijvoorbeeld gekeken naar geautomatiseerde bewerkingen door *edit-bots* op Wikipedia (zie Niederer & Van Dijck, 2010), of hoe online controverses geautomatiseerd worden binnen eenvoudige sorteer- en labelalgoritmes (zie bijvoorbeeld Marres, 2015). Automatisering van gegevensselectie, -categorisering, -labeling en -bewerking oogt eenvoudig en ongevaarlijk, maar draagt het risico in zich dat modellen worden gebouwd bovenop databronnen die op hun beurt weer aan de basis liggen van geautomatiseerde processen die oneerlijk kunnen zijn. Om een voorbeeld te geven uit de marketingwereld: als een reclamealgoritme wordt getraind om te zoeken naar de meest populaire kleur kledingstukken op een webshop, en de webshop biedt een tijd lang alleen maar rode kleding aan, dan zal het algoritme 'leren' dat mensen rode kleding prioriteren, met als gevolg dat de inkoopafdeling op basis van de output van het algoritme meer rode kleding gaat kopen en dit weer

aanbiedt op de webshop, wat statistisch de kans op een voorkeur voor rode kleding weer verhoogt, et cetera. Als we dit projecteren op domeinen zoals veiligheid en zorg, dan wordt duidelijk hoe ingrijpend, maar ook gecompliceerd het gebruik van AI kan worden. De grenzen van ADM zijn moeilijk te trekken, aangezien er talloze op algoritmes gebaseerde keuzes worden gemaakt voordat een digitale dienst een eindgebruiker bereikt.

Om toch grip op AI na te streven is een bepaalde vorm van het in kaart brengen van het gehele proces, van dataverzameling tot verwerking tot uitkomst, een belangrijke stap. Dit soort processen kan grofweg in kaart worden gebracht langs de lijnen van de levensloop van data (ook wel de *data-lifecycle* genoemd[15]). Een dergelijke indeling, hoewel sterk vereenvoudigd, biedt een basis voor het in kaart brengen van rechtvaardigheidsuitdagingen in AI en potentiële oplossingsrichtingen zoals die in de hierboven besproken literatuur aan bod komen. Als we uitgaan van een lineair proces of 'datastroom', kunnen we een aantal verschillende fasen onderscheiden, zijnde datageneratie en -verwerving, analyse en verwerking, opslag en beheer, en visualisatie en interfaces (zie Cavanillas, Curry & Wahlster, 2016). Toegepast op AI kunnen we data die verkregen worden uit de echte wereld, data die voor een bepaald AI-proces worden gegenereerd, de AI-manipulatie zelf, de uitkomsten van dit proces, de context van het gebruik (datavisualisatie, datainteractie) en ten slotte het effect dat dit heeft op de echte wereld onderscheiden.[16] Als het gaat om dataverzameling, is het belangrijk om te weten waar de gegevens vandaan komen, of de spreiding representatief is en of er toestemming is om deze gegevens te gebruiken voor een te ontwikkelen dienst of applicatie. Oplossingsrichtingen zijn onder andere te vinden in actieve en dynamische toestemming van het datasubject, of het gebruik van geverifieerde open databronnen. Gedurende de fase van het opruimen en gereed maken van data voor analyse is het belangrijk om een balans te vinden tussen compleetheid en bruikbaarheid van de databron: gooi je te veel niet-volledige attributen weg, dan verlies je misschien te veel context; neem je alle onvolledige of half volledige attributen mee, dan kom je waarschijnlijk veel minder ver in de analyse en is daarmee ook de uitkomst van het algoritme minder betrouwbaar. Het begrijpen van datalabels is van cruciaal belang; datalabels met dezelfde naam kunnen per context heel verschillend gebruikt worden of een compleet andere betekenis hebben. Het bespreken van deze labels met domeinexperts en eindgebruikers kan enorm verhelderend werken, alsmede het uitvoeren van kwaliteitsstudies op de gebruikte databronnen. Met betrekking tot de algoritmes zelf zou het helpen om een beter beeld te hebben van de best practices rondom hoe bepaalde algoritmes vooroordelen versterken. Ook zouden mogelijke technische oplossingen, bijvoorbeeld rondom uitlegbaarheid (in de informatica ook wel XAI genoemd: *explainable* AI), en organisatorische oplossingen, zoals het doen van (publiek toegankelijke) statistische controlestudies op de uitkomsten, stappen in de goede richting zijn. Wat betreft de uitkomsten is het van belang om gestandaardiseerde *auditing*-processen toe te passen en om op de lange termijn impact assessments uit te voeren op zowel procedurele als distributieve rechtvaardigheid. Ten slotte is het, logisch maar niet altijd gebruikelijk, van belang om in de context van publieke dienstverlening transparant en helder te communiceren over het doel en de gebruikte data-

bronnen van de AI-gebaseerde dienst.[17] Aan de hand van deze verdeling geeft tabel 2 een eerste indicatie van welke soorten uitdagingen op het gebied van distributieve en procedurele rechtvaardigheid kunnen ontstaan en welke mogelijke oplossingen kunnen worden onderscheiden.

Tabel 2 Distributieve-en procedurele uitdagingen en oplossingsgebieden

Stap in het dataproces	**Sub-stap**	**Distributieve uitdagingen**	**Procedurele uitdagingen**	**Oplossingsgebied**
Data- generatie en verzameling	Datageneratie	Weergave van de bevolking in gegevensbronnen, aannames bij data-verzameling	Bewustwording van het feit dat iets een datapunt is en van het bestaan van soorten gegevensverzameling	Dynamische toestemming, open data
	Collectie & selectie	Balans in soorten en hoeveelheid bronnen, vertekening en vooroordelen in data	Toegankelijkheid, toestemming en herbruikbaarheid van gegevens	Nieuwe vormen van toegang tot en bewijs van authenticiteit van data (via bijvoorbeeld SSI)
Voorbereiding van data voor AI	Dataopschoning	Verlies van context Inherente *proxies*	Gebrek aan transparantie in dataopschoningsmethodieken	Controlestudies, betrouwbare/ gecontroleerde datasets
	Datalabels en koppeling van data	Weging en interpretatie van etiketten/labels en gebruikte kenmerken Vereenvoudiging van variabelen door koppeling	Geautomatiseerde en/of handmatige koppeling van datasets in nieuwe datasets; (on)omkeerbaarheid	Bespreken van labels/features met belanghebbenden Publiceren van labels, co-design methoden toepassen
AI-proces	Selectie van modellen	Publiek begrip van AI-modellering, -expertise en openheid/gesloten heid van modellen en software	(In)transparantie van modellen - inherente technische complexiteit	Kennisgrafieken (*knowledge graphs*), sectoroverzichten van AI-applicaties, XAI
	Testen en interne validatie	Privé-audits, zelfregulering en zelf testen van uitkomsten	Nieuwheid van algoritme en ADM in het domein, gebrek aan benchmarks, gebrek aan standaard- procedures	Statistische controles Publicatie van het validatieproces

 doi: 10.5553/BenM/138900692020047004009 - Beleid en Maatschappij 2020 (47) 4

Tabel 2 (Vervolg)

Stap in het dataproces	Sub-stap	Distributieve uitdagingen	Procedurele uitdagingen	Oplossingsgebied
AI/ADM-uitkomsten	Plek van AI in organisatorische processen, inbedding van ADM	Interpretatie van het resultaat en de relatie met de praktijk of het domein	Inbedding in andere, bestaande processen: hoe veranderen de uitkomsten de interne processen	Impact assessment, risicoanalyse, interne auditing, controle door
	Externe validatie (wetenschappelijk, statistisch, sociaal, ethisch, juridisch, organisatorisch etc.)	Controleerbaarheid en kennis van de resultaten van AI en ADM	Invloed op werkstromen en besluitvormingsprocessen, directe en indirecte effecten	journalistiek, ngo's, toezichtsorganen zoals de gegevensbeschermings-autoriteiten
Gebruik van AI-uitkomsten	*Loop*-effecten (effect van AI op de reële wereld creëert nieuwe gegevensinvoer voor AI)	Langetermijneffecten op distributie van middelen en kansen, indirecte en secundaire effecten	Selffulfilling prophecy's, het inbouwen van data-afhankelijkheid in ADM, het standaard onderdeel worden van besluitvormingsprocessen. Uitdagingen rond verantwoording en vertrouwen	Meten en evalueren van de impact van AI en ADM op rechtvaardigheid op de lange termijn

Conclusie

In deze introductie rondom rechtvaardigheid hebben we getracht huidige debatten over procedurele en distributieve rechtvaardigheid in- en rondom kunstmatige intelligentie (AI) en algoritmische besluitvorming te beschrijven. Op basis van recente literatuur uit verschillende disciplines (onder andere juridische wetenschappen, wetenschaps- en technologiestudies en beleidsstudies) hebben we een overzicht gegeven van wat sociale gelijkheid kan betekenen en waarom vragen rond rechtvaardigheid ontstaan in een tijd waarin AI en algoritmische vormen van besluitvorming vanuit laboratoria hun weg naar de maatschappij gaan vinden. Vanuit een kort literatuuronderzoek zijn we op zoek gegaan naar facetten van een onderzoeksagenda, met name in een Nederlandse en Europese context, waarin we (tot nu toe) geen kwantitatieve standaarden hanteren rondom rechtvaardigheid (in tegenstelling tot in de Verenigde Staten, waar er bijvoorbeeld sprake is van een ongelijke impactkwantificering in het arbeidsrecht). Als het gaat om rechtvaardigheid, moet bij het ontwerp van een eerlijk systeem niet alleen rekening worden gehouden met verdelingsvraagstukken, maar moet ook worden gezorgd voor het tijdig updaten van informatie en voor toegankelijkheid en inzichtelijkheid van de overwegingen om AI toe te passen, alsook het verschaffen van inzicht in de gebruikte algoritmes in bijvoorbeeld AI-registers.[18]

Verdere bevindingen zijn dat veel aannames rondom distributieve rechtvaardigheid – wat een eerlijke verdeling is – voortbouwen op een langere geschiedenis van sociale ongelijkheid en deze in vele gevallen ook versterken. Dit laatste blijkt uit recente voorbeelden, zoals het verzet tegen het prioriteren van lagere-inkomenswijken bij het opsporen van fraude of de politieke onrust omtrent automatische risicoprofilering of schuldsanering. Dit toont aan dat sociale ongelijkheid niet alleen betrekking heeft op wettelijk beschermde 'gevoelige' categorieën, maar ook op sociale en politieke opvattingen over wie recht heeft op wat. Recente protesten in het Verenigd Koninkrijk rondom het automatisch genereren van eindexamencijfers[19] laten zien dat zodra algoritmische besluitvorming breed gevoelde grenzen overschrijdt, er een maatschappelijk debat ontstaat over het doel en de effecten van AI-toepassingen. Het voorgestelde raamwerk is een eerste poging om structuur aan te brengen in de mogelijk risico's van AI-gedreven diensten op rechtvaardigheid en mogelijke oplossingsrichtingen.

Noten

1 www.volkskrant.nl/nieuws-achtergrond/rotterdam-stopt-omstreden-fraudeonderzoek-met-syri~becb336a/.
2 https://ec.europa.eu/futurium/en/ai-alliance-consultation.
3 Zie https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60434 , p. 10.
4 Ooit was dit Google's motto: 'do no evil'; recentelijk is die ondertitel verdwenen van de Google-pagina's. https://gizmodo.com/google-removes-nearly-all-mentions-of-dont-be-evil-from-1826153393.
5 www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.
6 https://twitter.com/RobKitchin/status/1121749685545373697.
7 www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation.
8 www.theguardian.com/media/2020/oct/10/cory-doctorow-technologists-have-failed-to-listen-to-non-technologists.
9 Artificial General Intelligence en Artificial Super Intelligence.
10 www.theguardian.com/commentisfree/2019/may/07/racism-politics-reflects-pervasive-prejudice-brexit-britain.
11 Mittelstadt e.a. (2016); Larson & Angwin (2016); Dignum (2017); Selbst e.a. (2018); Taddeo & Floridi (2018); Vetzo, Gerards & Nehelman (2018).
12 www.thedigitalsociety.info/themes/responsible-data-science/ for an overview of responsible data science principles.
13 https://redasci.org/wp-content/uploads/2016/05/pipeline.png.
14 Zie bijvoorbeeld Oudshoorn & Pinch (2003); Latour (1999).
15 Jacob, D. (2019, april). *FAIR principles, an new opportunity to improve the data lifecycle*. https://hal.archives-ouvertes.fr/hal-02070883/.
16 Uiteraard zijn er vele variaties op een dergelijke indeling, zie bijvoorbeeld Curry (2016).
17 Zoals recentelijk gedaan door de gemeente Amsterdam via een publiek toegankelijk AI register. Zie https://algoritmeregister.amsterdam.nl/en/ai-register/.

18 Zie bijvoorbeeld het AI-register van Amsterdam: https://algoritmeregister.amsterdam.nl/en/ai-register/.
19 https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/.

Literatuur

Aalst, W.M.P. van der, Bichler, M., & Heinzl, A. (2017). Responsible Data Science. *Business & Information Systems Engineering*, 59 (5): 311–313. doi.org/10.1007/s12599-017-0487-z.

Bogen, M., & Rieke, A. (2018). *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*. www.upturn.org/reports/2018/hiring-algorithms .

Cavanillas, J.M., Curry, E., & Wahlster, W. (2016). *New horizons for a data-driven economy: a roadmap for big data in Europe*. Berlijn: Springer International Publishing.

Chavalarias, D., & Ioannidis, J.P. (2010). Science mapping analysis characterizes 235 biases in biomedical research. *Journal of clinical epidemiology*, 63 (11): 1205-1215.

Chawla, N.V., Hall, L.O., Bowyer, K.W., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16: 321-357.

Chayes, J. (2017, 23 augustus). How machine learning advances will improve the fairness of algorithms. *Huffington Post* .

Curry, E. (2016). The big data value chain: definitions, concepts, and theoretical approaches. In: *New horizons for a data-driven economy*. Cham: Springer, 29-37.

Dignum, V. (2017). *Responsible Artificial Intelligence: Designing AI for Human Values*. http://dspace.library.daffodilvarsity.edu.bd:8080/handle/20.500.11948/2177 .

Dijck, J. van (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & society*, 12 (2): 197-208. doi.org/10.24908/ss.v12i2.4776.

Dijk, J. van, & Hacker, K. (2003). The digital divide as a complex and dynamic phenomenon. *The information society*, 19 (4), 315-326.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Schafer, B. (2018). AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28 (4): 689-707.

Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Thirty-Second AAAI Conference on Artificial Intelligence*. Presented at the Thirty-Second AAAI Conference on Artificial Intelligence. www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16523.

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 55 (4): 1143-1185.

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and engineering ethics*, 24 (5): 1521-1536.

Iqbal, N. (2018, 20 oktober). Film Fans See Red Over Netflix 'targeted' Posters for Black Viewers. *The Observer*. www.theguardian.com/media/2018/oct/20/netflix-film-black-viewers-personalised-marketing-target .

Joint Research Centre (2017). *What Makes a Fair Society? Insights and Evidence*. European Commission. http://publications.jrc.ec.europa.eu/repository/bitstream/JRC106087/kj0716182enn.pdf.

Krishnakumar, A. (2019). *Assessing the Fairness of AI Recruitment Systems* (Masterthesis). Delft: TU Delft.

Larson, J., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm .

Latour, B. (1999). On recalling ANT. *The Sociological Review*. 47: 15-25.

Leventhal, G., Karuza, J., & Fry, W. (1980). Beyond fairness: a theory of allocation preferences. *Justice in social interaction*, 3 *:* 167-218. New York: Springer.

Madden, M., Gilman, M., Levy, K., & Marwick, A. (2017). Privacy, poverty, and Big Data: A matrix of vulnerabilities for poor Americans. *Wash. UL Rev.*, 95: 53.

Magnet, S. (2011). *When Biometrics Fail: Gender, Race, and the Technology of Identity*. Durham: Duke University Press.

Malcomson, S. (2016). *Splinternet: How geopolitics and commerce are fragmenting the World Wide Web* . OR Books.

Marres, N. (2015). Why map issues? On controversy analysis as a digital method. *Science, Technology, & Human Values*, 40 (5): 655-686.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3 (2): 2053951716679679. doi.org/10.1177/2053951716679679.

Niederer, S., & Dijck, J. van (2010). Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society*, 12 (8): 1368-1387.

Oudshoorn, N.E., & Pinch, T. (2003). *How users matter: The co-construction of users and technologies* . MIT press.

Piketty, T. (2015). About Capital in the Twenty-First Century. *American Economic Review*, 105 (5): 48-53. doi.org/10.1257/aer.p20151060.

Selbst, A.D., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2018). *Fairness and Abstraction in Sociotechnical Systems* (SSRN Scholarly Paper No. ID 3265913). https://papers.ssrn.com/abstract=3265913 .

Taddeo, M., & Floridi, L. (2018). How AI Can Be a Force for Good. *Science*, 361 (6404): 751-752. doi.org/10.1126/science.aat5991.

TNO (2018, 28 november). Lastige Ethische en Legale Vragen bij Kunstmatige Intelligentie. *NRC*. www.nrc.nl/advertentie/tno/kunstmatige-intelligentie-houdt-ons-spiegelvoor.

Tweede Kamer (2015). *Rapportage over het Systeem Risico Indicatie (SyRI)-projecten* (Kamerstukken II 2014/15, 17050, nr. 508). Tweede Kamer der Staten-Generaal. www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2015Z12876&did=2015D25854.

Veenstra, A.F. van, Djafari, S., Grommé, F., Kotterink, B., & Baartmans, R. (2019). *Quik Scan AI in de Publieke Dienstverlening*. www.rijksoverheid.nl/documenten/rapporten/2019/04/08/quick-scan-in-de-publieke-dienstverlening.

Velthoven, B.C.J. van (2011). Over het Relatieve Belang van een Eerlijke Procedure: Procedurele en Distributieve Rechtvaardigheid in Nederland. *Rechtsgeleerd Magazijn THEMIS*, 1: 7-16.

Vetzo, M., Gerards, J., & Nehelman, R. (2018). *Algoritmes en Grondrechten*. Den Haag: Boom juridisch.

Zuiderveen Borgesius, F. (2018). *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*. Council of Europe. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 .