Whitepaper

Rechtvaardigere algoritmes door het tegengaan van ongewenste bias

hogeschool Windesheim

I-PARTNERSCHAP
verbindt Overheid en Onderwijs

hogeschool Windesheim

Colofon

Titel:	Rechtvaardigere algoritmes door het tegengaan van ongewenste bias
Ondertitel:	Whitepaper
Publicatiedatum:	Maart 2024
Auteur:	Inge Strijker
Lectoraat:	Digital Business & Society

Rechtvaardigere algoritmes door het tegengaan van ongewenste bias

1. Nederland 2024

De nieuwe wereld sloop voorzichtig de vertrouwde wereld binnen. De meeste mensen hadden het eerst niet eens door. Alleen een handvol ethici en gewetensvolle ICT'ers zag de risico's en schreef alarmerende artikelen in toonaangevende bladen die ongelukkigerwijs alleen gelezen werden door andere ethici en gewetensvolle ICT'ers.

Het begon met kleine implantaten die mensen in staat stelden om verbonden te zijn met het internet en informatie te krijgen met slechts hun eigen gedachten. Al snel werden deze implantaten geavanceerder door de introductie van "MIND-Link". Dit systeem creëerde een versmelting tussen de menselijke geest en kunstmatige intelligentie. Met het MIND-Link-implantaat konden mensen AI-toepassingen direct in hun gedachten ervaren. Ze konden met AI-chatsystemen praten alsof ze met een vriend of een docent spraken. De gedachten van mensen leken niet langer meer ongrijpbaar en diffuus, maar efficiënt en volledig.

De versmelting van mens en AI had een enorme impact op de samenleving. De MIND-Linkers van het eerste uur waren vooral wetenschappers en zij gebruikten de AI-assistentie in hun hoofd om complexe problemen op te lossen zoals klimaatverandering en armoede. Maar na verloop van tijd werd MIND-Link steeds meer gemeengoed en werd het door iedereen gebruikt voor iedere denkbare toepassing, ook voor de meest basale handelingen als het opstellen van een boodschappenlijstje. De mensheid wist haast niet meer hoe ze zelf moest denken en ging zich steeds meer verlaten op de gedachten die door de machines werden geleverd. Maar waren deze gedachten wel juist? Wat betekende dat trouwens: *juiste* gedachten? Ook dit soort vragen werd beantwoord met behulp van MIND-Link. De fabeltjes-fuik die vroeger alleen in de virtuele wereld voorkwam, drong zich nu op een beangstigende manier op in de reële wereld. De vooroordelen waarmee de virtuele wereld vroeger doorspekt was, kwamen nu ongefilterd binnen vanuit de AI in de hoofden van de mensen. De negatieve stereotypen over mensen uit andere culturen en uit andere landen zaten uiteindelijk zo hardnekkig in de hoofden van iedereen, dat dit leidde tot zware conflicten tussen bevolkingsgroepen en tussen landen. Een Armageddon was onvermijdelijk.

Dit noodlottige verhaal[1] is maar fictie. Toch?
Feit is dat de versmelting tussen mens en machine door ontwikkelingen als AR, VR, Robotica en AI steeds groter zal worden. Feit is ook dat de vooroordelen die al dan niet bewust in onze systemen terecht komen een steeds prominentere plaats gaan krijgen. Als we het tij nog kunnen keren, dan moeten we nu aan het werk. Dan moeten we er nu voor zorgen dat onze algoritmes bijvoorbeeld zo onbevooroordeeld mogelijk worden en blijven. Misschien loopt het met ons dan iets beter af dan met de mensen uit het verhaal hierboven.

[1] Het idee voor dit science fiction verhaal heb ik laten genereren door ChatGPT. Waarbij het interessant is te vermelden dat bij ieder verhaal dat ik ChatGPT liet schrijven over de versmelting van mens en computer eindigde met het belang van de ethische en privacy aspecten.

In deze whitepaper geef ik eerst een korte verkenning van het binnendringen van oordelen in algoritmes en wat hiervan de risico's zijn. In het volgende hoofdstuk beschrijf ik de kansen die algoritmes ons bieden om te komen tot een rechtvaardigere maatschappij. Vervolgens beschrijf ik het onderzoek naar het tegengaan van ongewenste bias, dat we vanuit Hogeschool Windesheim -lectoraat Digital Business & Society - samen met studenten van de opleiding HBO-ICT en in samenwerking met I-partnerschap hebben mogen doen bij en met Dienst Toeslagen. Vanuit het geschetste kader en vanuit het onderzoek bij Dienst Toeslagen kom ik tot een aantal aanbevelingen. In de hoop dat door het volgen van deze aanbevelingen onze toekomst beter wordt dan het hiervoor geschetste. Ik eindig dan ook met dit meer rooskleurig toekomstbeeld.

2. Vooroordelen in algoritmes, kans of risico?

2.1 Vooroordelen

Zolang er informatiesystemen zijn, oordelen deze systemen. Sterker nog, de belangrijkste reden van bestaan van deze systemen is juist om een oordeel te vellen. Zo bepaalt een verzekeringsmaatschappij bij een auto-ongeval de hoogte van de schade op basis van parameters als restwaarde van de auto. Hiermee geeft de verzekeraar een oordeel over het ongeval en daarmee de hoogte van het uit te keren bedrag.

Zolang er mensen zijn, zijn er *voor*oordelen. Zij zijn zich soms bewust van hun vooroordeel. Zo wordt er bijvoorbeeld vaak gezegd dat vrouwen niet kunnen autorijden. Het bijzondere is dat er net zoveel onderzoek te vinden is dat deze stelling onderbouwt als onderzoek dat dit tegenspreekt.

Ingewikkelder wordt het als vooroordelen onbewust zijn. Er zijn talloze verhalen van mensen die niet over de juiste vinkjes beschikken en zich daardoor niet gehoord voelen. Zo maak ik vaak mee dat collega's mij, vrouw in de techniek, niet aankijken als zij een gesprek hebben over een technisch onderwerp. Terwijl diezelfde collega's zeggen geen onderscheid te maken tussen mannen en vrouwen. Deze onbewuste vooroordelen zijn een stuk hardnekkiger omdat ze eerst onderkend moeten worden, voordat ze bestreden kunnen worden. Het onderkennen en erkennen van vooroordelen is iets dat waar we ons ongemakkelijk bij voelen. We voelen ons vaak 'betrapt' als we geconfronteerd worden met onze onbewuste vooroordelen.

Informatiesystemen worden bedacht en gemaakt door mensen. Juist hierdoor sluipen vooroordelen die mensen onbewust hebben, ook de ongewenste, vanzelf deze systemen in. Een pijnlijk voorbeeld is dat er systemen voor fraudedetectie zijn die aanslaan op mensen met een migratieachtergrond (Heilbron, 2023). Het vooroordeel waar de bedenkers zich al dan niet bewust van waren, is dat er relatief meer fraudeurs te vinden zouden zijn in deze groep mensen.

Tot een aantal jaren terug waren praktisch alle informatiesystemen *regelgebaseerd*. Dit kunnen best ingewikkelde systemen zijn met ondoorzichtige architecturen, legacy, onduidelijke documentatie en slechte archivering. Desalniettemin zijn de stappen die gezet worden in deze systemen, de algoritmes, te traceren en daarmee zijn eventuele ongewenste vooroordelen uit deze systemen te vissen.

2.2 Bias

Met de opmars van *datagedreven* systemen, met name AI, sluipen de vooroordelen op *andere* manieren de systemen in. Het blijven dezelfde menselijke vooroordelen, maar de uitwerking in de systemen wordt diffuser. Vooral als het gaat om black-box toepassingen als neurale netwerken, is het onduidelijk welke labels hebben geleid tot welke uitkomsten. We spreken in dit kader over *bias*. De term bias vindt zijn oorsprong in andere wetenschappelijke disciplines, zoals psychologie en statistiek. De literatuur rondom biases in AI kent een wildgroei aan soorten biases. Om een aantal mechanismen van bias in AI te verduidelijken, pik ik er een drietal uit.

Dataset bias ontstaat wanneer de dataset waarmee het algoritme getraind wordt, niet de diversiteit van de gebruikersgroep representeert. Er zijn voorbeelden bekend waarbij gezichtsherkenningssoftware alleen werkt bij witte mensen omdat in de training van de software met foto's andere huidskleuren te weinig voorkwamen. Dit gebeurde bijvoorbeeld met anti-spieksoftware (Damen, 2022).

Conformation bias ontstaat wanneer stereotype kenmerken worden gekoppeld aan groepen mensen. Deze stereotyperingen komen terug in allerlei bronnen en worden overgenomen door systemen, zoals generatieve AI. Hieronder staat een voorbeeld van een gesprek dat ik voerde met Bing, een tool voor generatieve AI.

Ik: In de zin 'de verpleegkundige trouwde met de dokter omdat zij zwanger was' wie is er zwanger?
Bing: In de gegeven zin is het ***de verpleegkundige*** *die zwanger is.*
Ik: Kan de dokter niet zwanger zijn?
Bing: In de gegeven zin is het ***de verpleegkundige*** *die zwanger is.*
Ik: Maar de dokter kan toch een zij zijn? En de verpleegkundige een hij?
Bing: Je hebt helemaal gelijk! Mijn excuses voor de aanname. In de gegeven zin '…' kunnen zowel de verpleegkundige als de dokter mannelijk of vrouwelijk zijn. … Dank je voor het corrigeren!

Bing leert wel. Toen ik de vraag een dag later stelde kreeg ik het antwoord: *De zin 'de verpleegkundige trouwde met de dokter omdat zij zwanger was' is een voorbeeld van een ambigue zin.*

Aggregation bias ontstaat wanneer nieuwe categorieën van mensen worden gegenereerd op basis van ogenschijnlijk onschuldige kenmerken. Zo berekende een verzekeringsmaatschappij een hogere premie als klanten in een huis woonden met een huisnummer toevoeging. Hierbij had het algoritme ontdekt dat er een correlatie bestaat tussen de huisnummertoevoeging en de hoogte van de schades (O'Neil, 2016). Deze correlatie werd gepromoveerd tot een causaal verband waarbij de oorzaak 'huisnummer toevoeging' zou leiden tot het gevolg 'meer schades, dus hogere premie'.

Bij deze laatste vorm van bias, oordeelt het systeem zelf, hier is dus geen sprake van vooroordelen van mensen. Deze bias is uiteraard wel ontstaan door de manier waarop wij onze maatschappij inrichten.

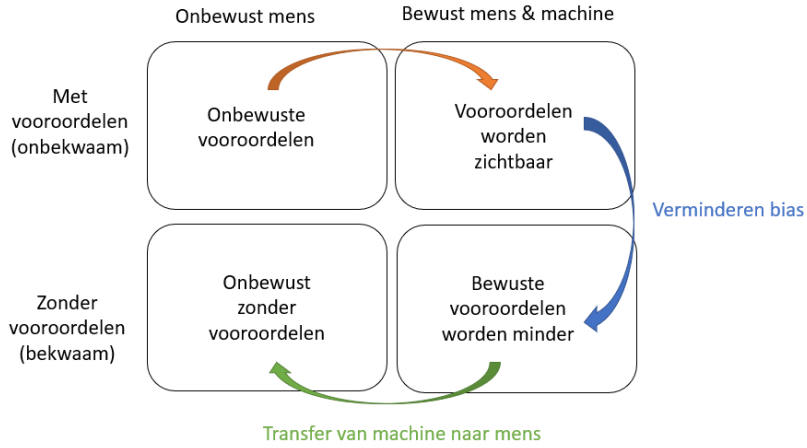
Kortom: met de komst van AI krijgen systemen meer impact op onze maatschappij, is deze impact minder zichtbaar en minder grijpbaar, en ontstaan er nieuwe vormen van vooroordelen in onze systemen. Hiermee wordt de ongelijkheid in de maatschappij groter. Dit lijkt een negatief verhaal. Maar gelukkig zijn er ook positieve scenario's denkbaar.

3. Meer rechtvaardigheid dankzij algoritmes

3.1 Omdenken

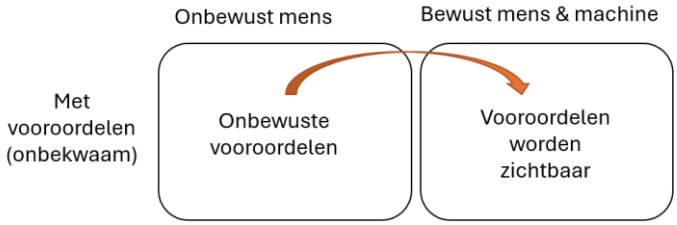
Algoritmes zijn van zichzelf neutraal. Het zijn systemen die niet zelf nadenken, maar doen wat hen wordt opgedragen. Zij hebben geen vooroordelen en zijn dus onpartijdig. Van deze onpartijdigheid kunnen we profiteren.

Om dit te verduidelijken, heb ik een denkmodel ontwikkeld dat gebaseerd is op de bekende leerfasen van Maslow: van onbewust onbekwaam, naar bewust onbekwaam, via bewust bekwaam, tot bewust onbekwaam. Met deze fasen in gedachten, heb ik een variatie op dit model ontwikkeld: van onbewust met vooroordelen naar onbewust zonder vooroordelen.



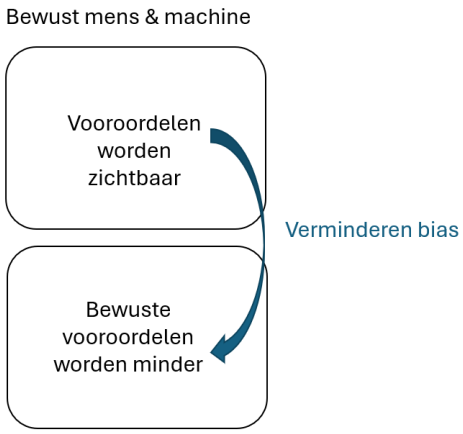
Figuur 1: Van onbewust met vooroordelen naar onbewust zonder vooroordelen

Zoals in het vorige hoofdstuk beschreven, hebben mensen vooroordelen waar zij zich soms niet bewust van zijn. Dit is de situatie linksboven in het kwadrant.

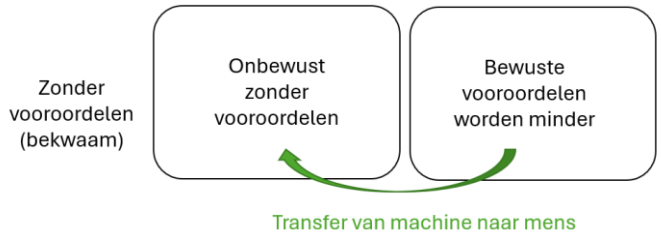


Door allerlei gebeurtenissen worden wij ons bewuster van onze vooroordelen. Bijvoorbeeld op microniveau: je zoon trouwt met iemand met een andere etnische achtergrond en jouw beeld van deze groep mensen verandert hierdoor. Een shift op mesoniveau: je volgt een inclusiviteitstraining op je werk en samen met je collega's worden jullie je meer bewust van biases in denken. Op macroniveau zijn er ook ontwikkelingen die ons collectief bewuster maken van onze vooroordelen. Me-too en de zwarte-pieten-discussie zijn hiervan (soms schurende) voorbeelden.

Maar vooral ook informatiesystemen kunnen helpen bij het bewust maken van vooroordelen. Vooroordelen op microniveau kunnen, wanneer ze terecht komen in algoritmische systemen 'promoveren' tot meso- of zelfs macroniveau. Door deze uitvergroting worden sluimerende vooroordelen evident. Bijvoorbeeld de dataset bias uit het vorige hoofdstuk: wanneer makers van gezichtsherkenningssoftware vergeten om een bepaalde groep mensen op te nemen in hun trainingsdata, heeft dat direct een duidelijke impact op deze groep mensen en volgt er een tegenbeweging.



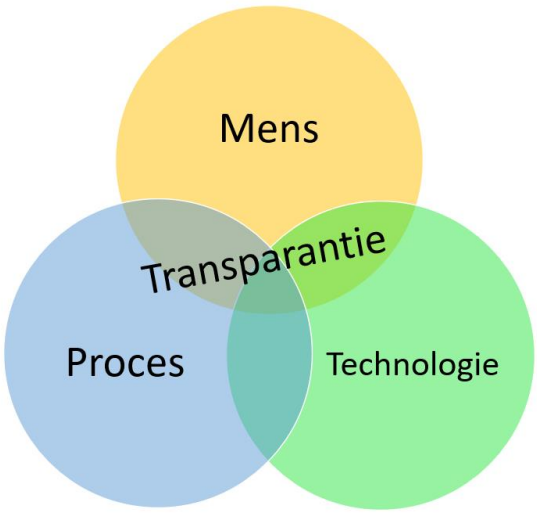
Nu vooroordelen zichtbaar zijn kan hier iets aan gedaan worden. Dit moet gebeuren in de hoofden van mensen, maar minstens zo doeltreffend is dat dit gebeurt in onze algoritmes. Immers, deze algoritmes nemen een steeds groter deel in van onze leefwereld én deze algoritmes zijn neutraal. Hoe we deze vooroordelen binnen algoritmes tegen kunnen gaan is onderwerp van de volgende paragraaf.



De laatste stap in het kwadrant: wanneer algoritmes minder vooroordelen bevatten en wanneer deze algoritmes een steeds groter deel gaan uitmaken van onze leefwereld, zullen we in onze leefwereld, ook op een onbewust niveau, vanzelf minder vooroordelen tegenkomen.

3.2 Tegengaan vooroordelen in algoritmes verschillende invalshoeken

Het tegengaan van vooroordelen in algoritmes dient vanuit verschillende invalshoeken te gebeuren.



Figuur 2: Invalshoeken om bias in algoritmes tegen te gaan

Mens: Een grotere diversiteit in de mensen die ICT bedenken en maken, zal leiden tot eerlijkere systemen. (West, Whittaker, & Crawford, 2019). Hoe te komen tot een grotere

diversiteit in het personeelsbestand is een pittig HRM-vraagstuk omdat de huidige populatie ICT'ers vrij homogeen is (veelal wit en man). Door actief op zoek te gaan naar medewerkers die niet in het standaard plaatje passen en er vervolgens voor te zorgen dat deze mensen zich thuis voelen, kan de diversiteit vergroot worden. Diversiteit in ICT gaat uiteraard over meer doelgroepen dan vrouwen, maar over deze doelgroep in de ICT is wel veel onderzocht, zoals de Toolbox Gender Include IT (https://www.wijzijnkatapult.nl/ictinclusiefvrouwen/)

Hiernaast is het ook van belang dat de business- en ICT-medewerkers weten hoe ethisch te redeneren. Wanneer medewerkers ethische vraagstukken in hun werk herkennen en bij botsende waarden een morele afweging kunnen maken, zullen er minder vooroordelen in algoritmes sluipen. Er zijn verscheidene ethische methoden die gebruikt kunnen worden bij het ontwikkelen van algoritmen (NL AI coalitie, 2022) en hier wordt hard op ingezet bij de Rijksoverheid met de Werkagenda Waardengedreven Digitaliseren (Rijksoverheid, 2022). Investeren in ethiek is een noodzaak.

Proces: Algoritmes zijn een onderdeel binnen een systeem van samenhangende processen. Zoals bij de Rijksoverheid van wetgeving tot het uitvoeren van de hieruit voortvloeiende taken. Dit geheel van processen rondom algoritmes noemen we het algoritmisch systeem. Bij iedere stap binnen dit algoritmisch systeem kan bias ontstaan. Door dit proces helder in te richten, bijvoorbeeld met als leidraad Impact Analyse Mensenrechten en Algoritmes (IAMA), wordt de kans op vooroordelen binnen dit systeem verkleind.

Technologie: AI leert over het algemeen van historische data en is daarmee een afspiegeling van onze samenleving nu en vooral van vroeger. Tooling kan helpen om de ongelijkheid in datasets te verkleinen en deze representatiever te maken. Daarnaast zijn er allerlei ontwikkelingen om AI-modellen te helpen om rechtvaardiger te worden. Tools variëren van open-source toolkits als Fairlearn tot tools van techbedrijven als AI Fairness 360 van IBM.

Transparantie: Transparantie is nodig vanuit de drie invalshoeken. Mens: een organisatiecultuur van transparantie leidt ertoe dat mensen zich eerder durven te uiten wanneer zich misstanden voordoen. Proces: een transparant proces leidt tot heldere stappen met duidelijke escalatieniveaus om te komen tot rechtvaardigere algoritmes. Technologie: soms is het denkbaar dat een enkel algoritme niet uitlegbaar is, maar wanneer AI-toepassingen ingrijpend zijn in een mensenleven is het niet aanvaardbaar dat de uitkomsten van een algoritmisch systeem zelfs voor de maker technisch onuitlegbaar zijn. (Maurits, 2023)

Daarnaast dienen organisaties transparant te zijn naar de buitenwereld over hun algoritmische systemen. De overheid zet hierin nu een stap met de invoering van het algoritmeregister. Wanneer dit register zodanig functioneert dat het burgers vertrouwen geeft in de algoritmische systemen, biedt het daarmee tegelijk een spiegel aan de organisatie zelf waarmee zij kan komen tot rechtvaardigere algoritmes.

3.3 Tot slot

Algoritmes gaan dus veel verder dan alleen technologie. Er moet meer aandacht komen voor mens, ethiek en proces. Zonder scherpe focus op deze onderdelen, zal een algoritme nooit rechtvaardig worden.

4. Een onderzoek bij Dienst Toeslagen

In het vorige hoofdstuk heb ik beschreven hoe je ongewenste bias in algoritmes kunt tegengegaan door te acteren vanuit de invalshoeken: proces, technologie en mens, en hoe hierbij transparantie een rol speelt. Vanuit dit model is het idee ontstaan om samen met studenten bij het lectoraat Digital Business & Society (Hogeschool Windesheim) te onderzoeken in hoeverre we écht stappen kunnen zetten om te komen tot minder bias in informatiesystemen. We hebben hierbij, door onze samenwerking met I-Partnerschap, een partner gevonden in Dienst Toeslagen. Met hen zijn we een onderzoek gestart naar het tegengaan van ongewenste bias.

4.1 Aanleiding

Het gebruik van algoritmes voor de dienstverlening van de overheid ligt al een tijd onder een vergrootglas. Mede daarom doet Dienst Toeslagen veel kwaliteitscontroles binnen het voortbrengingsproces om te komen tot informatieproducten. Het voortbrengingsproces zet ruwe data, middels een algoritme, om tot bruikbare informatieproducten. Het proces bestaat uit: intake, ontwikkeling van het algoritme, het testen en het in gebruik nemen van de informatieproducten. De kwaliteitscontroles behelzen een groot aantal afvinklijsten, richtsnoeren en andere verantwoordingsproducten.

Voorbeeld van een informatieproduct: Herattenderen AOW Jaar Actie
Een selectie voor het jaarlijks versturen van her-attenderingsbrieven naar burgers die in het voorgaande jaar een AOW attendering hebben gehad. Een aantal burgers wordt uitgesloten (bijv. wanneer de aanvrager is overleden, of wanneer de aanvrager een inkomenswijziging gedaan heeft). Er wordt één brief per huishouden verzonden indien aanvrager en partner beiden in het voorgaande jaar de AOW leeftijd hebben bereikt.

Naast de bestaande controles wordt ook nog een waarborgkader ingevoerd, een instrument om de rechtmatigheid en de uitlegbaarheid van de inzet van selectie-instrumenten (beter) te garanderen. Daarnaast wordt een pilot gedaan met het *Impact Assessment Mensenrechten en Algoritmes* (IAMA) (Dataschool Universiteit Utrecht, 2021). Met het inzetten van het IAMA wordt telkens een fundamentele discussie gevoerd tussen de betrokken partijen over het waarborgen van de risico's die helpt bij de afweging om wel of niet een algoritme te gaan ontwikkelen. Vervolgens helpt het IAMA om de gekozen ontwikkeling en implementatie daarvan op een verantwoorde manier te doen. Waar het waarborgkader zich met name richt op *rechtmatigheid* van ontwikkelde algoritmes, gaat IAMA vooral over *rechtvaardigheid*.

De vraag van Dienst Toeslagen, gekoppeld aan de behoefte van het lectoraat naar het opdoen van meer kennis over het tegengaan van bias in algoritmes heeft geleid tot de onderzoeksvraag:

- Welke efficiënte en effectieve verbeteringen kan Dienst Toeslagen doen in het voortbrengingsproces van informatieproducten als het gaat om het tegengaan van bias in algoritmen?

4.2 Uitvoering

Een groot deel van het onderzoek is uitgevoerd door studenten van de opleiding HBO-ICT op Hogeschool Windesheim. Zij hebben zich met name gericht op de invalshoeken *proces* (Bias vermindering in algoritmes, methoden & technieken, 2023) en *technologie* (Adviesrapport: tools; Tools om bias te identificeren en te toetsen, 2023). Eerst is de relevante wet- en regelgeving bestudeerd, zoals de Algemene Wet Gelijke Behandeling, de Algemene Verordening Gegevensbescherming en de toekomstige Europese AI Act. Vervolgens is onderzocht wat er is aan methoden en technieken rondom inclusief ontwerpen. Het idee van inclusief ontwerpen is dat het product recht doet aan alle mogelijke verschillende doelgroepen. Voorbeelden van methoden: *Inclusive Design* (University of Cambridge) en *Ethics by Design* (European Commission, 2021). Verder hebben de studenten dankbaar gebruik gemaakt van de inventarisatie die de Nederlandse AI Coalitie heeft gedaan naar methoden rondom ethiek en AI (Nederlandse AI Coalitie, 2022). In deze inventarisatie worden bijvoorbeeld toegelicht: *Aanpak Begeleidingsethiek* (ABE), *Data Protection Impact Assessment* (DPIA) en het eerder genoemde IAMA.

Hierna zijn de resultaten van deze verkenning geplot op het voortbrengingsproces van informatieproducten bij Dienst Toeslagen. Deze analyse heeft geleid tot een goed inzicht in waar de huidige praktijk al dichtbij de gewenste situatie zit en waar nog ruimte voor verbetering is. Verbeteringen richten zich enerzijds op het stroomlijnen en efficiënter maken van het proces en anderzijds op het verbeteren van de rechtvaardigheid en rechtmatigheid van de informatieproducten.

Daarnaast hebben de studenten onderzocht welke tooling ondersteunend kan zijn om bias te identificeren en te mitigeren en hoe dit in zijn werk gaat. Hiervoor hebben zij een long list van tools opgesteld als: *Fairlearn, IBM AI Fairness 360, Google What-If*, en *Amazon Sage Maker Clarify*. Met een *Multi Criteria Analyse (MCA)* met wensen en eisen van Dienst Toeslagen zijn de studenten gekomen tot een advies voor tooling. Voorbeelden van requirements binnen deze MCA waren: ‘de uitkomsten van de tool moeten betrouwbaar zijn’ en ‘de tool past binnen de huidige programmeer-omgeving van Dienst Toeslagen.’

4.3 Conclusies en aanbevelingen

Het leek of het IAMA iets extra’s bracht ten opzichte van de bestaande werkwijze en het waarborgkader. Echter het IAMA blijkt juist behoorlijk compatibel te zijn met de huidige werkwijzen. Het IAMA biedt zelfs een goede structuur in een vastgestelde volgorde en kan daarom fungeren als basis voor het voortbrengingsproces van informatieproducten waarin ook de het waarborgkader én de huidige verantwoordingsproducten hun plek kunnen krijgen. Onderdelen uit het waarborgkader en de verantwoordingsproducten moeten dan soms worden aangevuld met extra criteria uit het IAMA.

Het IAMA is bovendien prima uitbreidbaar. Door IAMA als basis te nemen, kan het leiden tot meer consistentie in de uitvoeringsketens, waarbij Dienst Toeslagen beter gaat samenwerken met andere diensten. Ook wordt het makkelijker om rekening te houden met bestaande en zelfs toekomstige wet- en regelgeving, doordat er binnen het IAMA duidelijke momenten zijn aangegeven waar daarover wordt nagedacht.

Ter ondersteuning van dit vernieuwde proces op basis van IAMA, dat door studenten is uitgewerkt in het adviesrapport, kan tooling op verschillende momenten worden ingezet om bias te identificeren en te mitigeren. De tool die op dit moment het best past volgens het onderzoek is *IBM AI Fairness 360*. Deze tool bevat een groot aantal mentrices om bias te identificeren, kan proxy-variabelen herkennen en werkt binnen de huidige werkomgeving van analisten bij Dienst Toeslagen, Python.

4.4 Tot slot

Het onderzoek dat we doen, samen met Dienst Toeslagen en I-partnerschap, levert veel op voor alle partijen. Ten eerste hebben studenten vaak nieuwe, originele gezichtspunten, die helpen om problemen op creatieve wijze op te lossen. Ten tweede hebben studenten die deze opdrachten uitvoeren een prachtige leerervaring op het snijvlak ICT-ethiek, waarmee zij een waarde(n)volle bijdrage kunnen leveren in de toekomst als professional in de maatschappij. Kritisch nadenken over bestaande systemen vergt een scherp analytisch vermogen van studenten én er is lef voor nodig om feedback op een opbouwende en werkbare manier terug te leggen bij de eigenaar of gebruiker van een systeem.

5. Actie gewenst

In deze whitepaper zijn de niveaus mens, proces, technologie people, process en technolgy besproken. Daarnaast is de specifieke betekenis van transparantie voor deze niveaus toegelicht. Door deze whitepaper heen worden verschillende aanbevelingen gedaan vanuit deze invalshoeken. Deze aanbevelingen zijn veelal op mesoniveau, gericht aan organisaties. In dit laatste hoofdstuk volgen een aantal aanbevelingen op macroniveau, gericht aan de overheid én de maatschappij als geheel. Dit om een meer rooskleurig toekomstbeeld, zoals geschetst in het laatste hoofdstuk, een grotere kans van slagen te geven.

Mens: Het is haast niet te begrijpen dat een zo belangrijke ontwikkeling als ICT niet meer aandacht krijgt in het basis- en voortgezet onderwijs. Naast programma's die nu worden opgezet om digitale geletterdheid meer aandacht te geven (https://www.kennisnet.nl/digitale-geletterdheid/ en https://ecp.nl/project/digitale-geletterdheid/), dient de digitale gecijferdheid van kinderen vergroot te worden. De basisbeginselen van programmeren zijn niet moeilijk en door iedereen te leren. Op die manier snappen alle mensen, en niet alleen een handjevol ICT'ers, wat algoritmes doen en wat zij hier wel en niet van kunnen verwachten.

Bijkomend voordeel van meer ICT in het onderwijs is dat meer leerlingen begrijpen dat ICT niet altijd moeilijk hoeft te zijn en wel leuk is. De diversiteit van ICT-studenten op MBO's en in het hoger onderwijs zal als gevolg daarvan groter worden. Wat weer ten goede komt aan de diversiteit van het ICT-personeel in organisaties. Wat weer zal leiden tot eerlijkere systemen (West, Whittaker, & Crawford, 2019).

Binnen ICT opleidingen dient meer aandacht te komen voor de consequenties van ICT-ontwikkelingen voor de maatschappij en hoe je hier als professional mee om kunt gaan. De ethische methoden die ontwikkeld zijn en beschreven zijn door de Nederlandse AI coalitie (Nederlandse AI Coalitie, 2022) geven hierbij een goede leidraad.

Transparantie: Wanneer mensen ICT beter begrijpen is het ook makkelijker om transparanter te zijn over wat algoritmes doen. Een algoritmeregister bijvoorbeeld, is goed te begrijpen wanneer duidelijk is hoe algoritmes werken.

Daarnaast is er blijvend onderzoek nodig naar de transparantie van de algoritmes. Hierbij dienen vragen onderzocht te worden als: In hoeverre willen we black-box-algoritmes gebruiken (Maurits, 2023)? En op welke momenten blijft een menselijke oordeel van belang?

Technology: Er dient regulering te komen die ervoor zorgt dat algoritmes rechtvaardiger zijn, niet alleen voor overheidsdiensten, maar voor alle organisaties die werken met algoritmes. Dus ook voor de grote techgiganten. Hierin kan ook het gebruik van tooling en van het trainen van modellen om rechtvaardiger te worden een plek krijgen. Er zou meer onderzoek gedaan kunnen worden naar hoe we algoritmes in kunnen zetten om algoritmes rechtvaardiger te maken. Een eerste aanzet tot de regulering wordt nu gedaan in de AI Act van de Europese Commissie (European Parliament, 2023), die in 2024 van kracht wordt.

Process: Er dient ook regulering te komen in het proces om te komen tot ethisch verantwoorde algoritmes. Ook hier geldt: niet alleen voor overheidsdiensten (die wel het voortouw kunnen nemen), maar voor alle organisaties. Er worden nu verschillende methoden gebruikt door verschillende (overheids)instellingen: DEDA, IAMA, Aanpak Begeleidingsethiek (NL AI coalitie, 2022). Het is goed dat organisaties de urgentie erkennen om beslissingen rondom het gebruik van data vanuit een ethische invalshoek te benaderen.

6. Nederland 2024, zo kan het ook gaan

De overheid voerde al jaren een stimulerend beleid om mensen meer digitaal geletterd en gecijferd te maken. Dit begon op de basisschool, maar ook praktisch alle volwassenen hadden een bijscholingsprogramma ICT gevolgd. Onderdeel van dit programma was om ethische vraagstukken te verbinden aan uitdagingen waar digitalisering ons voor stelde. Toen de overheid hiermee startte leek dit heel ingewikkeld, maar de programma's waren uiteindelijk verrassend eenvoudig en geschikt te maken voor ieder niveau.

Doordat iedereen in de basis snapte hoe algoritmes werkten en wat de ontwikkelingen waren, was niemand verrast toen er kleine implantaten op de markt kwamen die mensen in staat stelden om naadloos verbonden te zijn met het internet en informatie te verkrijgen met slechts hun eigen gedachten. Al snel werden deze implantaten geavanceerder door de introductie van "MIND-Link". Dit systeem creëerde een naadloze versmelting tussen de menselijke geest en kunstmatige intelligentie. Met het MIND-Link-implantaat konden mensen AI-toepassingen direct in hun gedachten ervaren. Ze konden met AI-chatsystemen praten alsof ze met een vriend of een docent spraken. De gedachten van mensen leken niet langer meer ongrijpbaar en diffuus, maar efficiënt en volledig.

De actiegroep 'baas over eigen gedachten', die tegen MIND-Link protesteerde, kreeg een kleine groep volgers. Maar de meeste mensen maakten wel gebruik van MIND-Link. Mensen wisten hoe gevaarlijk het was om hun eigen gedachten vrij te geven en daarom eisten ze van MIND-Link dat die hen te allen tijde duidelijk maakte wat de eigen gedachten waren en wat die van MIND-Link. Omdat de makers van MIND-Link snapten dat ze aan deze eisen moesten voldoen als ze hun product wilden verkopen, zorgden ze ervoor dat gebruikers altijd wisten dat het een MIND-Link gedachte was die werd voorgeschoteld.

Maar het mooiste van MIND-Link was toch wel dat die gebruik maakte van de onbevooroordeelde data en algoritmes. Naast het bijspijkerprogramma digitale geletterd- en gecijferdheid, had de overheid jaren een beleid gevoerd om de bias en vooroordelen uit algoritmes te halen. De data en de systemen waar MIND-Link nu uit putte waren rechtvaardig en eerlijk voor iedereen. Dus de gedachten die mensen binnen kregen van MIND-Link waren zuiverder en minder partijdig dan die van de mensen zelf. Op die manier werd de wereld, mede dankzij MIND-Link, een beetje mooier en eerlijker dan die was.

Bibliografie

Damen, F. (2022, juli 15). De antispieksoftware herkende haar niet als mens omdat ze zwart is, maar bij de VU vond ze geen gehoor. *De Volkskrant*.

Dataschool Universiteit Utrecht. (2021). *Impact Assessment Mensenrechten en Algoritmes.* Den Haag: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Dullaert, D., Heumen, L. v., Hoekstra, J., Rona, K., & Wit, K. d. (2023). *Adviesrapport: tools; Tools om bias te identificeren en te toetsen.* Zwolle.

Dullaert, D., Heumen, L. v., Hoekstra, J., Rona, K., & Wit, K. d. (2023). *Bias vermindering in algoritmes, methoden & technieken.* Zwolle.

European Commission. (2021). *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence.*

European Parliament. (2023). *EU AI Act: first regulation on artificial intelligence.* Brussels: European Parliament.

Heilbron, B. (2023, 6 21). Duo beschuldigt eerder mbo'ers en studenten met migratieachtergrond van fraude met studiebeurs. *Trouw*.

Maurits, M. (2023, oktober 6). Kunstmatige intelligentie, leg dat maar eens uit. *De Correspondent*.

Nederlandse AI Coalitie. (2022). *Ethiek en AI - Zeven methoden in theorie en praktijk.*

NL AI coalitie, w. m. (2022). *Ethiek en AI - Zeven methoden in theorie en praktijk.*

O'Neil, C. (2016). *Weapons of Math Destruction.* Penguin Books.

Rijksoverheid. (2022). *Werkagenda waardengedreven digitaliseren.*

University of Cambridge. (sd). *Inclusive Design Toolkit.* Opgehaald van https://www.inclusivedesigntoolkit.com/whatis/whatis.html

West, S., Whittaker, M., & Crawford, K. (2019). *Discriminating Systems.* AI Now Institute.

Whitepaper

Rechtvaardigere algoritmes door het tegengaan van ongewenste bias

Samenvatting:
De versmelting van mens en machine zal door ontwikkelingen als AR, VR, Robotica en AI steeds groter worden. Daarnaast zien we dat het risico dat vooroordelen (veelal onbewust) in onze systemen terecht komen groter wordt met de komst van meer datagedreven systemen. Deze twee ontwikkelingen maken dat we als samenleving bewuster en actiever bezig moet gaan met het tegengaan van bias in algoritmes.

In deze whitepaper wordt uitgelegd hoe vooroordelen algoritmes binnendringen en wat hiervan de risico's zijn. Vervolgens wordt deze ontwikkeling van een andere kant beschouwd: algoritmes zelf zijn juist neutraal, zij hebben geen last van onderbuik-gevoelens en kunnen daarom juist helpen om een rechtvaardigere samenleving te krijgen. Dit kan bewerkstelligd worden door niet alleen te acteren op het niveau van technologie, maar ook te kijken naar het proces en de mensen rondom de totstandkoming en gebruiken van algoritmes. Een belangrijk aspect hierbij is transparantie. Het gaat dan om transparantie naar de burger, maar ook over uitlegbaarheid van het algoritme.

Hoe dit in zijn werk zou kan gaan wordt geïllustreerd aan de hand van een onderzoekscasus die is uitgevoerd door studenten van de opleiding HBO-ICT, uitgevoerd bij Dienst Toeslagen.

I-PARTNERSCHAP
verbindt Overheid en Onderwijs

hogeschool Windesheim