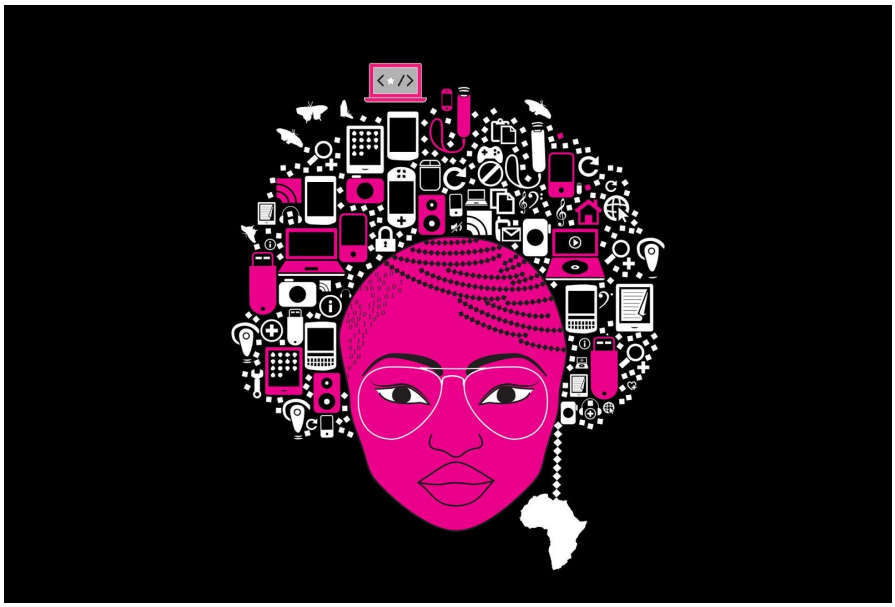
Computer says no tegen zwarte vrouwen

Een interdisciplinair onderzoek naar discriminatie door algoritmes jegens zwarte vrouwen



Black Girls Code (2018)

Scriptie: Interdisciplinair onderzoek II (LA3V11003)
Utrecht, juni 2020

Auteurs

Margot Hoogerwerf	5996740
Eline Kempkes	5958490
Emma Sofie Achterberg	6252389

Begeleiders

Interdisciplinair:	prof. dr. Iris van der Tuin
Geschiedenis:	Feline van den Boogerd
Kunstmatige Intelligentie:	dr. Rianne van Lambalgen
Filosofie:	dr. Sander Werkhoven

Voorwoord

Na maanden van hard werken presenteren wij met trots onze interdisciplinaire scriptie over discriminatie van zwarte vrouwen door algoritmes. Dit onderwerp hebben we onderzocht vanuit de disciplines Geschiedenis, Kunstmatige Intelligentie en Filosofie. Hoewel deze disciplines ver uit elkaar liggen, kwamen wij vanuit onze gedeelde interesse in genderongelijkheid en racisme al snel uit op het gekozen onderwerp. Nadat wij voor ons onderwerp hadden gekozen, laaide het racismedebat wereldwijd eind mei 2020 plotseling sterk op. De aanleiding hiervoor was de moord op de Afro-Amerikaanse George Floyd door een witte politieagent. Niet alleen racistisch gedrag van de politie zelf staat momenteel ter discussie, maar ook de ICT-wereld wordt onder de loep genomen. Zo plaatste de New York Times (Hill, 2020) twee dagen voor het voltooien van dit onderzoek een artikel over hoe Robert Julian-Borchak Williams, een zwarte man, onterecht beschuldigd werd door een gezichtsherkenning-algoritme en werd opgepakt. Techbedrijf Amazon kondigde recent aan een jaar lang geen software voor gezichtsherkenning meer te zullen verkopen aan Amerikaanse politiekorpsen, omdat de software te vaak fouten maakt bij mensen met een donkere huidskleur. Techbedrijf IBM maakte bekend helemaal te stoppen met gezichtsherkenning (Hijink, 2020). Deze recente ontwikkelingen laten zien hoe relevant onze scriptie over discriminatie door algoritmes op dit moment is.

Wij denken met deze scriptie niet alleen oorzaken voor discriminatie door algoritmes en suggesties voor oplossingen te kunnen noemen, maar ook het nut van een interdisciplinaire insteek te illustreren. Het combineren van inzichten uit onze disciplines heeft geleid tot nieuwe, discipline-overstijgende inzichten. In een schematisch overzicht hebben we zo verschillende niveaus en de onderlinge samenhang hiertussen kunnen identificeren. Hiermee wordt het makkelijker om problemen en oplossingen op het juiste niveau te identificeren.

Terugkijkend is het schrijven van deze scriptie een erg leuke en uitdagende ervaring geweest. We hebben tijdens het schrijven veel steun gevonden bij elkaar, maar ook zeker bij onze begeleiders. Wij willen onze gezamenlijke begeleider Iris van der Tuin bedanken om al haar adviezen over integratietechnieken en haar enthousiasme. Ook onze vakreferenten Feline

van den Boogerd, Rianne van Lambalgen en Sander Werkhoven willen wij bedanken voor hun uitgebreide feedback.

Veel leesplezier,
Margot Hoogerwerf, Eline Kempkes, Emma Sofie Achterberg

Samenvatting

In deze scriptie is interdisciplinair literatuuronderzoek gedaan naar de discriminatie door profilerende algoritmes jegens zwarte vrouwen. De concrete hoofdvraag van dit onderzoek is: *"Welk handelingsperspectief kan worden afgeleid uit de oorzaken van de discriminatie jegens zwarte vrouwen door profilerende algoritmes?"*

Deze vraag is vanuit drie verschillende disciplines behandeld, namelijk de Geschiedenis, de Kunstmatige Intelligentie en de Filosofie, met als doel de drie inzichten te integreren en uiteindelijk een interdisciplinair antwoord op de hoofdvraag te geven.Vanuit Geschiedenis is onderzocht hoe het historisch verklaard kan worden dat juist zwarte vrouwen door profilerende algoritmes worden gediscrimineerd, door historische stereotypes en hun doorwerking in hedendaagse technologieën te bestuderen. Vanuit de Kunstmatige Intelligentie is gekeken naar de technologische problematiek bij het ontwerpen van een eerlijk algoritme, om zo inzicht te verschaffen in de oorzaak en daaruit oplossingen af te leiden. Vanuit de Filosofie is er een handelingsperspectief afgeleid uit de morele bezwaren en morele verantwoordelijkheid omtrent algoritmes die discriminatie veroorzaken.

Uiteindelijk zijn deze drie verschillende inzichten met elkaar geïntegreerd, door eerst een *common ground* te creëren, waardoor conflicten tussen de disciplines werden opgelost. Daarna wordt door de *more comprehensive understanding* een interdisciplinair antwoord op de hoofdvraag gecreëerd. Hiertoe is een schema met drie lagen gevormd waarbij er een analyse van de werking van discriminerende algoritmes wordt gegeven. Uit deze analyse blijkt dat er een positieve terugkoppeling bestaat bij het gebruik van discriminerende algoritmes in de maatschappij, ze kunnen bestaande stereotypen versterken: het zogeheten cybertyping. In dit schema is tenslotte het handelingsperspectief gepresenteerd, waarbij er op meerdere momenten ingegrepen moet worden om zo de positieve terugkoppeling tegen te gaan.

Inhoudsopgave

Inleiding

Hoofdstuk 1: Geschiedenis

Hoofdstuk 2: Kunstmatige Intelligentie

Hoofdstuk 3: Filosofie

Hoofdstuk 4: Integratie

4.1. Disciplinaire conclusies

4.2. *Common Ground*

4.3. *More comprehensive understanding*

4.4. Discussie

Inleiding

Emma Sofie Achterberg, Eline Kempkes, Margot Hoogerwerf

Het probleem

Steeds meer bedrijven en organisaties maken gebruik van algoritmes. Deze algoritmes blijken soms discriminerend te zijn tegenover vrouwen[1] en mensen van kleur. Er zijn hier verschillende voorbeelden van. Zo werd bij een sollicitatieprocedure van Amazon een algoritme gebruikt dat had geleerd om geen vrouwen te selecteren (Dastin, 2018). Daarnaast karakteriseerde een veelgebruikt taalalgoritme Afro-Amerikaanse namen als 'onaangenaam' en namen van witte Amerikanen als 'aangenaam' (Zou & Schiebinger, 2018). Er zijn veel voorbeelden die illustreren dat discriminatie door algoritmes geen incidenteel probleem is, maar een fundamenteel gevolg van institutioneel racisme en seksisme.

Zo bleek uit onderzoek van ProPublica (2016) dat het Amerikaanse algoritme COMPAS, dat de kans op recidive van gevangenen voorspelt, racistisch werkte. Het algoritme voorspelde dat mensen van kleur veel vaker opnieuw de fout in zouden gaan, terwijl dit in de praktijk niet gebeurt. Als gevolg hiervan zullen zwarte criminelen na hun vrijlating strenger gecontroleerd worden dan witte criminelen, hetgeen een vorm van discriminatie is, omdat er onterecht een onderscheid wordt gemaakt tussen groepen mensen.

Uit een ander onderzoek (Bolukbasi, Chang, Zou, Saligrama & Kalai, 2016) bleek dat een veelgebruikt taalalgoritme seksistisch werkte: Wanneer aan het algoritme de analogie werd voorgelegd "Man is to computer programmer as woman is to 'X'", werd de 'X' door het algoritme ingevuld met 'homemaker'. Wanneer bedrijven zoals Google dit soort seksistische taalalgoritmes gebruiken, zal de taal die online circuleert ook seksistisch blijven. Dit kan een discriminerende uitwerking hebben op vrouwen. Het online taalgebruik houdt bijvoorbeeld denkbeelden over de rol van vrouwen als 'huisvrouw' in stand, waardoor het moeilijker voor hen zal zijn om uit deze rol te stappen.

[1] Met 'vrouwen' doelen wij op iedereen die zichzelf als vrouw identificeert (genderidentiteit) en niet alleen op mensen die bij de geboorte als vrouw zijn aangeduid op basis van seksekenmerken.

Dit roept een aantal belangrijke vragen op. Hoe kan worden verklaard dat algoritmes discrimineren? Waarom is het zo moeilijk om discriminatie door algoritmes op te lossen? En hoe kunnen we deze discriminatie in de toekomst voorkomen? In dit onderzoek wordt gestreefd naar meer inzicht in de rol van algoritmes bij institutioneel racisme en seksisme.

Met name vrouwen van kleur blijken vaak gediscrimineerd te worden door algoritmes. Zij vallen namelijk zowel onder de groep 'vrouwen' als onder de groep 'mensen van kleur'. Vrouwen van kleur worden daarom gediscrimineerd als vrouw én als persoon van kleur. Algoritmes voor gezichtsherkenning die gebruikt werden door techgiganten zoals IBM, Microsoft en Amazon bleken bijvoorbeeld de gezichten van vrouwen van kleur het slechtst te herkennen. Dit kan verstrekkende gevolgen hebben: wanneer gezichtsherkenning bijvoorbeeld wordt gebruikt bij het opsporen van misdaad, kan een vrouw van kleur hierdoor onjuist worden beschuldigd, omdat haar gezicht niet goed is herkend (Buolamwini & Gebru, 2018: 1). Discriminatie van vrouwen van kleur is een vorm van 'intersectionaliteit', het idee dat individuen gediscrimineerd worden op grond van meerdere factoren die met elkaar interfereren. Het resultaat van deze interferentie is dan groter dan de som der delen (Foulds, Islam, Keya & Pan, 2018). Zo kan een vrouw van kleur last hebben van seksisme en racisme, maar ook van de interferentie van deze twee factoren.

Discriminatie door algoritmes kan plaatsvinden als het algoritme 'profilerend' werkt. Profilerende algoritmes creëren profielen die bestaan uit een verzameling kenmerken van een groep, waar individuen vervolgens onder geplaatst worden (Mittelstadt, 2016: 8). Een individu krijgt zo kenmerken toegeschreven die gekoppeld zijn aan de groep waarin hij/zij ingedeeld is, om voorspellingen over dit individu te maken. De voorspellingen die de profilerende algoritmes doen kunnen vervolgens worden gebruikt voor het opstellen van risicoprofielen, bijvoorbeeld ten aanzien van de kans dat iemand een terrorist zal zijn, of een waardeloze werknemer.

Een fundamenteel probleem van algoritmes is dat zij profileren op basis van essentialistische kenmerken (Cheney-Lippold, 2017). Deze kenmerken doen geen recht aan de verschillende ervaringen die bijvoorbeeld 'vrouwen' of 'mensen van kleur' hebben. Hoewel individuen uit dezelfde groep perspectieven en ervaringen kunnen delen, bestaat er geen essentie

van *de* identiteit van vrouwen van kleur (Harding, 2004; Squires, 2002: 454). Om dit te erkennen zullen wij ons in deze scriptie voornamelijk richten op Amerikaanse zwarte vrouwen. De historische en daardoor ook hedendaagse ervaringen van zwarte vrouwen in Amerika verschillen namelijk sterk van de ervaringen van bijvoorbeeld Amerikaanse latina's of Amerikaans-Aziatische vrouwen evenals van zwarte vrouwen elders op de wereld. Dit kan verklaard worden vanuit de verschillende geschiedenissen van deze groepen en vanuit het colorisme: het idee dat dat er een hiërarchie in huidskleur bestaat waarbij geldt "hoe dichter bij wit, hoe beter." De ervaringen van vrouwen met een zwarte huidskleur zijn hierdoor anders dan de ervaringen van vrouwen met een licht getinte huidskleur. Wij hebben ervoor gekozen ons te richten op de Verenigde Staten, omdat de rassenproblematiek en daarmee het colorisme diep verankerd is in de samenleving. Bovendien worden in de Verenigde Staten de belangrijkste algoritmes ontworpen, dus zou dit ook het land zijn waar veranderingen gemaakt kunnen worden in het ontwerp van algoritmes.

Uiteraard moet worden opgemerkt dat ook Amerikaanse vrouwen met een zwarte huidskleur, de groep die wij in deze scriptie zullen bestuderen, verschillende individuele ervaringen koesteren. Het is daarom ook niet onze bedoeling om deze groep een essentialistische identiteit toe te kennen. Wij verzetten ons juist tegen het feit dat algoritmes dit wél doen. In deze scriptie zal er meer inzicht worden gegeven in de rol van profilerende algoritmes in discriminatie jegens zwarte vrouwen. We richten ons op de oorzaken van dit fenomeen en stellen op basis van deze oorzaken oplossingsstrategieën voor. De concrete vraag die in deze scriptie centraal zal staan, luidt:

"Welk handelingsperspectief kan worden afgeleid uit de oorzaken van de discriminatie jegens zwarte vrouwen door profilerende algoritmes?"

Interdisciplinaire aanpak

Om een antwoord te bieden op deze vraag voeren wij een interdisciplinair onderzoek uit. Kenmerkend voor interdisciplinair onderzoek is de samenwerking tussen onderzoekers die afkomstig zijn uit verschillende disciplinaire vakgebieden. De inzichten uit deze verschillende vakgebieden worden niet alleen opgesomd, maar worden ook met elkaar geïntegreerd om zo tot nieuwe discipline-overstijgende kennis te komen. Deze stap maakt ons onderzoek geen multidisciplinair, maar een interdisciplinair onderzoek (Repko & Szostak, 2017). Discipline-overstijgende kennis is vaak noodzakelijk voor het oplossen van een maatschappelijk probleem, omdat individuele disciplines slechts een deel van het probleem bestuderen en omdat zij bovendien ieder op hun eigen manier *biased* zijn. Iedere discipline kijkt immers vanuit zijn eigen epistemologie naar de werkelijkheid. Wij hopen dat de uitkomst van deze scriptie het nut van interdisciplinair onderzoek zal onderschrijven.

In dit onderzoek zal het onderzoeksprobleem, de discriminatie van zwarte vrouwen door profilerende algoritmes, vanuit drie disciplines worden bestudeerd. De drie disciplines zijn de Geschiedenis, de Kunstmatige Intelligentie (KI) en de Filosofie. Hieronder zullen wij de disciplines één voor één toelichten en zullen wij het gebruik van deze disciplines voor een onderzoek naar discriminatie van zwarte vrouwen door algoritmes rechtvaardigen.

De Geschiedenis bestudeert zowel ‘datgene wat er in het verleden is gebeurd’ (*res gestae*) als ‘de verhalen over datgene wat er in het verleden is gebeurd’ (*historia rerum gestarum*). Met andere woorden: zij bestudeert wat mensen in verschillende tijden meemaakten en hoe zij daaraan betekenis gaven, maar ook hoe mensen jaren, decennia of eeuwen later omgaan met dat wat er in het verleden is gebeurd. In deze scriptie wordt vanuit Geschiedenis de deelvraag gesteld: Hoe kan historisch verklaard worden dat juist zwarte vrouwen door profilerende algoritmes worden gediscrimineerd? Deze vraag is relevant voor de beantwoording van de hoofdvraag van het onderzoek vanuit de veronderstelling dat grote maatschappelijke problemen altijd hun wortels hebben in het verleden. De discipline Geschiedenis kan op ten minste drie concrete manieren bijdragen aan ons onderzoek. Ten eerste is een belangrijk thema dat de geschiedwetenschap bestudeert ‘identiteitsvorming’. En laat nu net in ons onderzoek de identiteit van zwarte vrouwen centraal staan. Ten tweede bestudeert de geschiedwetenschap vaak

technologische uitvindingen, zoals de boekdrukkunst, atoombom of televisie. Historici weten hierdoor dat technologische uitvindingen altijd gepaard gaan met bedoelde en onbedoelde maatschappelijke veranderingen. De geschiedwetenschap is daarom bij uitstek geschikt voor het analyseren van de impact van algoritmes als technologische 'uitvinding'. Ten derde bestudeert de geschiedwetenschap de doorwerking van historische ideeën in het heden. In ons onderzoek kan vanuit de geschiedwetenschap daarom het fenomeen "*cybertyping*" bestudeerd worden: de doorwerking van historische ideeën over ras en etniciteit in digitale technologieën.

De discipline Kunstmatige Intelligentie heeft een interdisciplinair karakter waarbij onder ander wiskunde, filosofie, informatica en logica samenkomen. In deze discipline wordt onderzocht hoe menselijke intelligentie kan worden nagebootst door een computer en wat daarvan de gevolgen zijn, zowel technisch als maatschappelijk. Om deze reden is het nuttig om de Kunstmatige Intelligentie bij dit onderzoek te betrekken: er zijn de laatste jaren veel studies gedaan naar de werking van profilerende algoritmen, evenals hun gevolgen op de maatschappij. Vanuit dit technische oogpunt rijst de vraag waarom er nog geen goed eerlijk algoritme is ontworpen. Een verklaring hiervoor is dat er nog geen consensus bestaat over het de definitie van een eerlijk algoritme, waardoor er ook niet naar kan worden gestreefd. In dit hoofdstuk zal de onderliggende problematiek bij het definiëren van een eerlijk profileren algoritme centraal staan, om zo te verklaren waarom er nog geen technische mogelijkheden zijn. De verschillende definities voor eerlijke algoritmes zullen worden besproken, evenals hun beperkingen. Door deze technische inzichten is het mogelijk om de oorzaken van een discriminerend algoritme in kaart te brengen, wat invloed heeft op het handelingsperspectief.

De Filosofie is een breed georiënteerde discipline die zich bezighoudt met de vragen achter de wetenschappen en hun theorieën. Het gaat om theorieën over hoe objecten, individuen, groepen, systemen, organisaties en het bestaan zelf in elkaar steken. De tak van de filosofie die van groot belang is voor dit onderzoek, is de ethische, politieke en sociale filosofie. Vanuit deze tak van de filosofie zal onderzocht worden welk handelingsperspectief ontleend kan worden aan de morele bezwaren en morele verantwoordelijkheid ten aanzien van profilerende algoritmes. Zo wordt er niet alleen bekeken *wat* er gedaan moet worden, maar ook door *wie*.

In de volgende hoofdstukken zal de discriminatie van zwarte vrouwen door algoritmes vanuit de drie hierboven beschreven disciplines onderzocht worden. Hierna volgt een vierde hoofdstuk waarin de disciplinaire inzichten met elkaar worden geïntegreerd. Dit wordt gedaan door middel van de methode die geïntroduceerd is door Repko en Szostak (2017). In dit hoofdstuk zal eerst *common ground* worden gecreëerd door verschillende conflicten tussen de disciplines op te lossen. Vervolgens worden in de *more comprehensive understanding* alle inzichten met elkaar geïntegreerd, waardoor een interdisciplinair antwoord op de hoofdvraag ontstaat. Onze *more comprehensive understanding* zal bestaan uit een schematische weergave van een positieve terugkoppeling die in de maatschappij ontstaat door het gebruik van discriminerende algoritmes. Algoritmes veroorzaken zo'n positieve terugkoppeling doordat zij de discriminatie die al in de maatschappij aanwezig is versterkt. In de schematische weergave zullen ook verschillende oplossingen worden gesuggereerd. Tezamen vormen deze oplossingen het handelingsperspectief.

Hoofdstuk 1: Slavernij, surveillance en stereotypes

Waarom juist zwarte vrouwen worden gediscrimineerd door algoritmes

Margot Hoogerwerf, Geschiedenis

Het is waarschijnlijk niet de bedoeling van programmeurs van techbedrijven om discriminerend te zijn (Noble, 2018: 90). Hoe kan het dan toch dat algoritmes dat soms wel zijn, bijvoorbeeld tegenover zwarte vrouwen? Om deze vraag te beantwoorden zal ik verschillende inzichten van historici en andere wetenschappers over de geschiedenis van 'ras'[2] en technologie analyseren op basis van een literatuuronderzoek. De precieze vraag die ik in dit hoofdstuk van onze scriptie zal beantwoorden, luidt: "Hoe kan historisch verklaard worden dat juist zwarte vrouwen door profilerende algoritmes worden gediscrimineerd?"

Historici van technologie hebben nog niet veel aandacht geschonken aan het onderwerp 'ras', en andersom hebben historici van 'ras' nog weinig aandacht geschonken aan technologie. Eén van de weinige historici die dit verband wel uitgebreid heeft bestudeerd, Bruce Sinclair (2004: 156), stelt dat: "Even younger historians of technology have been surprisingly slow to consider the technological constitution of racial identity and difference." In mijn hoofdstuk zal ik inzichten over de geschiedenis van 'ras' en over de geschiedenis van technologie met elkaar verbinden. Hiertoe zal ik onder meer gebruikmaken van de ideeën van wetenschapper in de Afro-Amerikaanse geschiedenis Simone Browne over de relatie tussen slavernij en surveillance, van de ideeën van cultuurwetenschapper Safiya Noble over stereotypes jegens zwarte vrouwen

[2] Ik zet de term 'ras' met opzet tussen haakjes om te benadrukken dat dit geen biologische term is die toepasbaar is op mensen. Lange tijd deelden wetenschappers mensen in in verschillende categorieën op basis van uiterlijke en genetische kenmerken. Inmiddels hebben biologen aangetoond dat er binnen de soort *Homo Sapiens* geen rassen bestaan. 'Ras' heeft dus alleen betekenis als culturele constructie. 'Ras' moet als onderdeel worden gezien van lokale discoursen over etniciteit. 'Ras' wijst meer op de de zichtbare verschillen in het fysieke voorkomen van verschillende groepen, 'etniciteit' is een bredere term en omvat ook de culturele eigenschappen van groepen. In de praktijk worden de termen 'ras' en 'etniciteit' vaak door elkaar heen gebruikt, in een gemengde betekenis van kleur, land van persoonlijke of voorouderlijke afstamming en culturele eigenschappen (Yanow, 2013; JOHO, z.d.) Ik heb ervoor gekozen om in dit hoofdstuk de term 'ras' te hanteren in plaats van 'etniciteit', omdat ik voornamelijk zal schrijven over praktijken die gebaseerd zijn op het fysieke voorkomen van zwarte mensen. Bovendien wil ik anachronismen vermijden. Daartoe is het verstandiger om de term 'ras' te hanteren, die al werd gebruikt in de historische perioden waarover ik zal schrijven, dan de term 'etniciteit', die pas in 1972 voor het eerst in een Engels woordenboek verscheen. Ik zal 'ras' wel tussen haakjes blijven plaatsen, omdat ik wil erkennen dat het om een sociale constructie gaat.

en van de ideeën van socioloog David Lyon over de manier waarop algoritmes mensen classificeren en profileren.

Ik zal mij in dit disciplinaire deel van de scriptie richten op de Verenigde Staten, omdat daar de meeste algoritmen van grote techbedrijven zoals Google, Apple en Facebook worden ontwikkeld. Dit zijn de algoritmen die wereldwijd de grootste invloed hebben. Bovendien is rassenproblematiek diep verankerd in de Amerikaanse samenleving. Het is daarom interessant om de verbanden tussen de geschiedenis van technologie en de geschiedenis van 'ras' in de Verenigde Staten te onderzoeken (Sinclair, 2004: 1).

Allereerst zal ik betogen dat de achttiende-eeuwse slavernij onlosmakelijk verbonden was met de opkomst van de publieke sfeer en de daarmee samenhangende *racializing surveillance*. Met dit begrip doel ik op de surveillancepraktijken die normen produceren over 'ras', en die daardoor bepalen wat wel en wat niet geaccepteerd is voor individuen en groepen van een bepaald 'ras'. Ik zal vervolgens uiteenzetten hoe stereotypes over zwarte mensen, met name over zwarte vrouwen, werden gebruikt ter rechtvaardiging van deze *racializing surveillance*. Door stereotypes in hun historische context te plaatsen, beantwoord ik aan de behoefte die Noble (2018: 82) benoemt om te onderzoeken "why and how we get these stereotypes in the first place."

Daarna zal ik een grote sprong voorwaarts maken naar het huidige digitale tijdperk. Ik zal uiteenzetten hoe historische stereotypes en praktijken van *racializing surveillance* onbewust doorwerken in digitale technologieën. Hiermee beantwoord ik aan de behoefte die Noble (2018: 108) benoemt om het idee dat technologieën neutraal zijn onmiddellijk te ontkrachten. Om de doorwerking van historische discriminerende praktijken in digitale technologieën te verklaren, zal ik wijzen op de afwezigheid van zwarte, vrouwelijke programmeurs in de ICT. Ik zal betogen dat, om de discriminatie van zwarte vrouwen op te lossen, ten eerste meer zwarte vrouwen in de ICT zullen moeten gaan werken en ten tweede witte mannelijke programmeurs bewust gemaakt zullen moeten worden van hun stereotypes jegens zwarte vrouwen.

1. Transatlantische slavenhandel: opkomst van de publieke sfeer en surveillance

De opkomst van de achttiende-eeuwse 'publieke sfeer' werd door de bedenker van deze fameuze term, de filosoof Jürgen Habermas, bejubeld als hét begin van de democratie (Habermas, Lennox & Lennox, 1974: 50). Volgens Habermas kwamen burgers nu voor het eerst samen om op een rationele manier met elkaar te debatteren, bijvoorbeeld in koffiehuizen of in salons. Daar werd druk gediscussieerd over thema's als eigendomsrechten en literatuur.

Verschillende academici hebben er echter op gewezen dat de publieke sfeer in werkelijkheid niet zo inclusief was als Habermas suggereerde.[3] Houston Baker, wetenschapper in de Afro-Amerikaanse cultuur, betoogt bijvoorbeeld dat de Amerikaanse publieke sfeer niet toegankelijk was voor het merendeel van de zwarte mensen. Als gevolg van de trans-Atlantische slavenhandel werkten velen van hen tot ver in de negentiende eeuw als slaven[4] op plantages of in huishoudens. Zij waren eigendom van slavenhouders en werden verhinderd in het verwerven van geletterdheid. Hierdoor konden zij niet deelnemen aan de publieke sfeer (Baker, 1994: 9).

Slavenhouders hadden er natuurlijk baat bij de ondergeschikte positie van zwarte mensen in de publieke sfeer te handhaven. Zij konden hun economische winst alleen behalen door zwarte mensen uit te buiten. Het was daarom in hun belang om de macht te houden over zwarte mensen (Habermas, 1962: 124-125). Verschillende auteurs koppelen deze machtsuitoefening in de publieke sfeer aan het begrip *surveillance*.[5] Zij doelen met dit begrip op de manieren waarop lichamen in de publieke sfeer gecontroleerd en gedisciplineerd worden (Browne, 2015: 17 en 38). Volgens de filosoof Michel Foucault (1975) ontstond er tegelijk met de opkomst van de publieke sfeer een *surveillance*-samenleving waarin macht onzichtbaar werd en tegelijkertijd in iedere menselijke relatie een rol gingen spelen. Om te beschrijven hoe dit in de praktijk gebeurde, bijvoorbeeld in gevangenissen, scholen en ziekenhuizen, gebruikt Foucault de metafoor van het Panopticum. Het Panopticum is een gevangenismodel dat bestaat uit een

[3] Landes (1988) liet bijvoorbeeld zien dat vrouwen in de publieke sfeer doorgaans werden buitengesloten en Hardt (1996) toonde dat de publieke sfeer doorgaans enkel toegankelijk was voor de middenklasse.

[4] Er is een debat gaande over een alternatief voor de term 'slaaf', om aan te geven dat 'slaaf' niet een inherente conditie was maar een positie waarin mensen door dwang werden geplaatst en gehouden. Een alternatief dat vaak geopperd wordt is de term 'tot slaaf gemaakte'. Ik heb ervoor gekozen in mijn hoofdstuk toch de term 'slaaf' te gebruiken, omdat dit mijns inziens de negatieve lading al dekt. De term 'slaaf' verhult niets en we weten allemaal welke negatieve gebeurtenissen ermee worden aangeduid. De term 'tot slaaf gemaakte' zou bovendien kunnen leiden tot een minder leesbare tekst.

[5] Socioloog Patricia Hill Collins betoogt bijvoorbeeld dat "surveillance was a way of ensuring that Blacks would stay in their designated, subordinate places in white-controlled public and private spheres" (1988: 20).

centrale toren die omringd wordt door cellen. De opzichters in de centrale toren kunnen de gevangenen in de cellen zien, maar andersom niet. De gevangenen kunnen daarom nooit weten wanneer zij bekeken worden en hebben een ondergeschikte machtspositie (Foucault, 1975).

Browne (2015: 32) merkt op dat de bedenker van het Panopticum, de filosoof Jeremy Bentham, een poos meereisde met een schip waarop vrouwelijke slaven werden uitgebuit. Zij vraagt zich af hoe deze ervaring mogelijk Benthams ideeën over surveillance heeft geïnspireerd. Browne komt tot de conclusie dat de opkomst van surveillance onlosmakelijk verbonden is met slavernij. "Surveillance is the fact of antiblackness," stelt zij. Surveillance is ontwikkeld *om* zwarte mensen te controleren (Browne, 2015: 10 en 42).

2. Historische technieken van *racializing surveillance*

Omdat surveillance volgens Browne onlosmakelijk verbonden is met slavernij en daarmee met 'ras', introduceert zij het concept *racializing surveillance*. Zij definieert dit concept als volgt: "Racializing surveillance is a technology of social control where surveillance practices, policies, and performances concern the production of norms pertaining to race and exercise a power to define what is in or out of place" (Browne, 2015: 16). Het concept omvat dus de surveillancepraktijken die normen produceren over 'ras', en die daarmee bepalen wat wel en wat niet geaccepteerd is voor individuen en groepen van een bepaald 'ras'. De uitkomst van *racializing surveillance* is volgens Browne (2015: 16) vaak een discriminerende behandeling van degenen die negatief gekenschetst worden door de surveillancepraktijken.

Ik zal een paar historische surveillancepraktijken toelichten die het concept *racializing surveillance* illustreren. Te beginnen met enkele disciplinerende technieken die op slavenplantages werden gebruikt om zwarte mensen te controleren; op deze plantages trad *racializing surveillance* volgens Browne (2015: 52) voor het eerst in werking. Sleutelkenmerken van het surveillancesysteem op de slavenplantages waren bijvoorbeeld gezocht-posters en slavenpassen, die de mobiliteit van slaven beperkten. Gezocht-posters werden verspreid wanneer een slaaf was weggelopen. Ze beschreven uitgebreid de fysieke kenmerken en soms ook de persoonlijkheidskenmerken van de slaaf in kwestie. Met de posters hoopten slavenhouders hun weggelopen slaaf weer in bezit te krijgen. Zij impliceerden daarmee ook dat 'een vrije zwarte'

misplaatst was. Een zwart persoon behoorde niet vrij te zijn, maar moest gecontroleerd worden (Browne, 2015: 52-54).

CAUTION!!

COLORED PEOPLE

OF BOSTON, ONE & ALL,

You are hereby respectfully CAUTIONED and advised, to avoid conversing with the

Watchmen and Police Officers of Boston,

For since the recent ORDER OF THE MAYOR & ALDERMEN, they are empowered to act as

KIDNAPPERS

AND

Slave Catchers,

And they have already been actually employed in KIDNAPPING, CATCHING, AND KEEPING SLAVES. Therefore, if you value your LIBERTY, and the *Welfare of the Fugitives* among you, *Shun* them in every possible manner, as so many *HOUNDS* on the track of the most unfortunate of your race.

Keep a Sharp Look Out for KIDNAPPERS, and have TOP EYE open.

APRIL 24, 1851.

Afbeelding 1: "Caution! Colored People of Boston," poster (1851). Library of Congress, Printed Ephemera Collection; Portfolio 60, Folder 22. Geciteerd in Browne, 2015: 23.

Slavenpassen waren een soort identificatiedocumenten die slavenhouders moesten kunnen tonen wanneer de *overseer*, de slavenopzichter van de regio, het eigendom van de slavenhouders kwam monitoren. Deze praktijk droeg bij aan het behandelen van zwarte mensen als eigendom (Browne, 2015: 24). De behandeling van zwarte mensen als eigendom was ook zichtbaar in andere praktijken, zoals het brandmerken van de slaven met de initialen van de eigenaar. Cultuurwetenschapper Stuart Hall schaart dit brandmerken onder de term *epidermalization*: "the literal inscription of race on the skin" of, iets breder gedefinieerd, "the imposition of race on the body" (Browne, 2015: 97 en 113).

Ook buiten de slavenplantages werden zwarte mensen gecontroleerd. Dit blijkt bijvoorbeeld uit de introductie van zogenaamde 'lantaarnwetten' in het achttiende-eeuwse New York City. Vanaf 1713 werden hier verschillende wetten uitgevaardigd die zwarte mensen verboden om 's avonds zonder lantaarn de straat op te gaan. Het idee achter deze wetten was dat

zwarte mensen een risico voor de veiligheid vormden en daarom onder toezicht moesten worden gehouden. Door zichzelf te verlichten met een lantaarn konden zwarte mensen altijd bekeken worden door toezichthouders, maar zelf wisten zij nooit wanneer dit gebeurde (Browne, 2015: 78-79). De lantaarnwetten kunnen daarom beschouwd worden als een typische toepassing van het Panopticum.

3. Rechtvaardiging: stereotypes over de zwarte vrouw

Zwarte mensen werden dus door verschillende surveillancepraktijken onderdrukt, zowel op de slavenplantages als daarbuiten. Ook na de afschaffing van de slavernij in 1865 bleef de surveillance over zwarte mensen bestaan, met name door de invoering van de Jim Crow-wetgeving in het zuiden van de Verenigde Staten. Deze wetgeving was erop gericht zwarte mensen gescheiden te houden van witte mensen in publieke instellingen en onthield zwarte mensen van hun stemrecht (Browne, 2015: 57). Hoe kon deze surveillance over zwarte mensen gerechtvaardigd worden? Hadden slavenhouders behalve financiële belangen nog andere redenen voor de onderdrukking van zwarte mensen? Browne (2015: 57) wijst als antwoord op deze vraag naar de stereotypes over zwarte mensen, met name over zwarte vrouwen, die slavenhouders en andere witte mensen inzetten ter rechtvaardiging van hun gewelddadige en controlerende surveillance.

Een stereotype dient om eigenschappen van een groep toe te schrijven aan een individu dat beschouwd wordt tot deze groep te behoren (Gilman, 1985: 204). Cultuurwetenschapper Safiya Noble (2018: 94-99) beschrijft drie stereotypes over zwarte vrouwen die in de tijd van de slavernij in Amerika tot stand kwamen: de Jezebel, de Mammy en de Sapphire. Alledrie helpen ze te begrijpen waarom zwarte vrouwen vandaag de dag worden gediscrimineerd door profilerende algoritmes. Het Mammy-stereotype was gebaseerd op het slaaf-zijn en kan helpen te begrijpen waarom er vandaag de dag zo weinig zwarte vrouwen in de ICT werken. De Jezebel- en Sapphire-stereotypes waren gebaseerd op het zwarte vrouw-zijn en kunnen helpen te begrijpen waarom zwarte vrouwen ook vandaag de dag nog worden verhinderd in hun activisme. Ik zal alledrie de stereotypes toelichten.

Het stereotype van de Mammy portretteerde zwarte vrouwen als loyale slaven die blij waren met hun leven als slaaf. Ze waren loyaal naar de witte familie die zij verzorgden en 'kenden hun plaats' als gehoorzame slaaf (Browne, 2015: 57-58). Deze karikatuur impliceerde dat zwarte vrouwen te dom waren voor ander werk dan werk op de plantages of in huishoudens. Het werd daarmee een rationalisering voor economische discriminatie (Browne, 2015: 58; Pilgrim, 2012a). Het stereotype van de Mammy behelsde ook een deseksualisering van de zwarte vrouw: de Mammy werd afgeschilderd als een dikke, oude zwarte vrouw, zodat zij geen concurrentie zou vormen voor de witte vrouw van het huishouden waar zij werkte (Noble, 2018: 96-98; Pilgrim, 2012a).

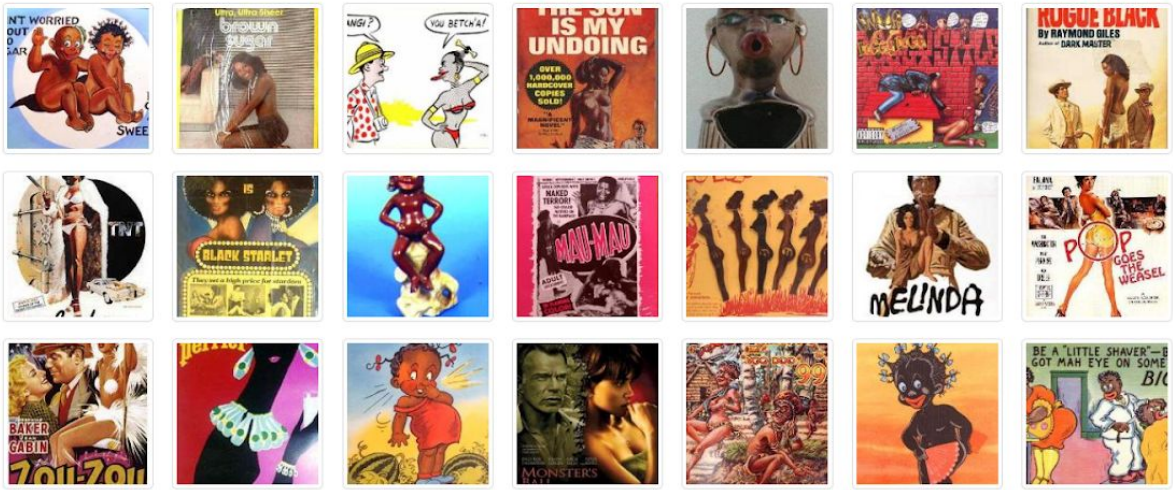
Het tweede stereotype, dat van de Jezebel, betrof het idee dat zwarte vrouwen seksueel 'makkelijk' en onverzadigbaar waren (Noble, 2018: 96). Dit stereotype baseerden witte Amerikanen op de uiterlijke kenmerken die sommige zwarte vrouwen bezaten, zoals dikke lippen, forse billen en grote genitaliën: "When they saw the female black, they saw her in terms of her buttocks and saw represented by the buttocks all the anomalies of her genitalia" (Gilman, 1985: 219). Vanwege hun veronderstelde seksuele ontembaarheid werden zwarte vrouwen ook wel gelijkgesteld aan apen. Zij werden beschouwd als minder menselijk. Nota bene de gerenommeerde bioloog Charles Darwin beschouwde de billen van zwarte vrouwen als een "somewhat comic sign of the primitive nature of the black female" (Darwin, 1871). De biologische fascinatie over zwarte vrouwen blijkt bijvoorbeeld uit de zogenaamde 'Hottentottenteententeententoonstellingen' waarin zwarte vrouwen geëtaleerd werden (Noble, 2018: 94).[6] De bekendste tentoongestelde zwarte vrouw was zonder twijfel Saartjie Baartman (1789-1815), een slavin die in Zuid-Afrika geboren was en die vanaf 1810 in Europa tentoongesteld werd als de *Hottentot Venus*. Na haar dood werden in haar autopsieverslag haar genitaliën uitgebreid beschreven en werd Saartjie vergeleken met een orang-oetan (Gilman, 1985: 213).

Het derde stereotype is dat van de zwarte vrouw als Sapphire: de 'boze zwarte vrouw'. Volgens dit stereotype zijn zwarte vrouwen altijd boos op zwarte mannen voor redenen variërend

[6] Het kunstmatige woord verwijst naar het Afrikaanse volk van de Hottentotten, ofwel de Khoikhoi. Het is onbekend wie de tongbreker 'Hottentottenteententeententoonstelling' heeft bedacht en wanneer. Zeker is dat het woord in Nederland in 1929 al bestond (Historiek, 2019).

van werkloos zijn tot seksuele interesse tonen in witte vrouwen. Hoewel zwarte mannen haar belangrijkste doelwit zijn, heeft de zwarte vrouw gif voor iedereen die haar beledigt of niet respecteert. Het Sapphire-stereotype is tegenwoordig vaak aanwezig in de populaire cultuur, zoals op televisie. Het gevaarlijke van dit stereotype is dat het zwarte vrouwen onderdrukt in hun activisme: het stereotype ontmoedigt zwarte vrouwen om actief te zijn. In plaats daarvan kunnen zij beter passief, gehoorzaam en onzichtbaar zijn (Pilgrim, 2012b).

Het Jezebel- en het Sapphire-stereotype werd volgens historicus Sander Gilman door witte mannen gebruikt als rechtvaardiging voor hun misbruik van zwarte vrouwen: deze vrouwen waren in navolging van dit stereotype gevaarlijk en moesten in toom worden gehouden. Hij stelt dat de zogenaamde angsten van mannen voor zwarte vrouwen werden omgezet in een zogenaamde behoefte om de seksualiteit van vrouwen te controleren: “The *white man's burden* thus becomes his sexuality and its control. The colonial mentality which sees 'natives' as needing control is easily transferred to 'women' needing control” (Gilman, 1985: 237). Gilman koppelt zo de stereotypes aan het streven naar surveillance dat kenmerkend was voor de achttiende-eeuwse publieke sfeer.



Afbeelding 2: Culturele artefacten die het Jezebel-stereotype weergeven (Ferris State University, Jim Crow Museum of Racist Memorabilia, z.d.)

Afbeelding 3: "The Hottentot Venus." Georges Cuvier, "Extraits d'observations faites sur le cadavre d'une femme connue à Paris et à Londres sous le nom de Vénus Hottentote," 1817 (Gilman, 1985: 214).

4. Doorwerking van *racializing surveillance* in digitale technologieën

Laten we nu even een flinke sprong in de tijd maken naar het digitale tijdperk, dat zo'n beetje begon in de jaren 70 en 80 van de twintigste eeuw. Dit digitale tijdperk ging gepaard met nieuwe technologieën en die brachten op hun beurt nieuwe vormen van surveillance met zich mee. Dit heeft sommige academici ertoe bewogen om de klassieke surveillance-metafoor van het Panopticum te herconceptualiseren. Thomas Mathiesen, bijvoorbeeld, veranderde het Panopticum in een 'Synopticum' waarin alle mensen, door de komst van massamedia, veranderd zijn in kijkers: er is nu geen sprake meer van 'the few watch the many', maar van 'the many watch the few' (Browne, 2015: 38; Siapera, 2010: 109). Met de komst van digitale media liet mediawetenschapper Mark Andrejevic (2004) de concepten van het Pan- en Synopticum zelfs achterwege. Hij betoogde dat de digitale surveillance gekenmerkt wordt door een 'the many watch the many': mensen kunnen nu zelf participeren aan de media en handelen daardoor als constante kijkers van elkáár. Dit is in feite een nog sterkere vorm van surveillance dan het Pan- en Synopticum, omdat we nu niet weten wanneer, hoe en door wie precies ons online gedrag wordt bekeken. We weten niet waar het kantoortje van de bewaker zich bevindt.

Dat nieuwe digitale technologieën gepaard gaan met nieuwe vormen van surveillance, betekent echter niet dat oude vormen van surveillance compleet afwezig zijn in digitale technologieën. Volgens David Lyon (2003), bijvoorbeeld, werken het Panopticum en het

Synopticum in het digitale tijdperk nog steeds door, en wel in tandem: ze complementeren elkaar als controlemechanismen. Deze gedachtegang volgend betoogt Browne dat oude raciale surveillancepraktijken uit de tijd van het Panopticum onbewust doorwerken in nieuwe digitale technologieën. Zij sluit zich hiermee aan bij andere wetenschappers, zoals Brock (2011) en Nakamura (2002: 35) die betogen dat het narratief van 'whiteness as normality' bevestigd en gereproduceerd wordt in digitale media en technologieën.

De doorwerking van raciale surveillancepraktijken in technologieën is niet nieuw. De eerste fotocamera's waren bijvoorbeeld al erg slecht in het detecteren van zwarte mensen. De technologie in de fotocamera's was niet geschikt voor het fotograferen van donkerdere huidskleuren, omdat de ontwerpers ervan uit waren gegaan dat hun consumenten – degenen die gefotografeerd wilden worden – zouden bestaan uit mensen met een lichte huidskleur (Van Schie, 2020). Filmwetenschapper Richard Dyer (2000: 103) betoogt daarom dat er een algemene cultuur van licht bestond in de fotografie, film en kunst waarin witte mensen centraal stonden in die zin dat zij verondersteld werden 'het licht waard te zijn.'

Net zoals in de fotografie, werken raciale surveillancepraktijken nu door in nieuwe digitale technologieën. Dit is bijvoorbeeld zichtbaar in biometrische technologieën, zoals technologieën voor gezichtsherkenning, iris-scans, vingerafdrukken of het waarnemen van loopbewegingen. Biometrische technologieën kunnen lichamen digitaliseren, doordat algoritmes het lichaam – of delen van het lichaam en de performances van een lichaam – coderen als data. Op basis van die data kunnen computers de lichamen verzamelen en sorteren (Browne, 2015: 109).

Dit sorteren is volgens Lyon (2003: 2) problematisch, ten eerste omdat de categorieën waarin mensen gesorteerd worden, zijn gebaseerd op stereotypes, bijvoorbeeld op raciale stereotypes. Er is dus sprake van een soort digitale *epidermalization*, waarbij biometrische technologieën raciale stereotypes benadrukken door lichamen van het ene 'ras' anders te coderen dan lichamen van het andere 'ras' (Browne, 2015: 109-113). Het sorteren is bovendien problematisch, omdat mensen op basis van 'hun categorie' risico's of juist waarden toegekend krijgen die een effect hebben op hun levenskansen. "Surveillance as social sorting" kan zo leiden tot discriminatie, bijvoorbeeld wanneer iemand op basis van zijn huidskleur de toegang tot een

stadion, vliegtuig of baan ontzegd wordt (Lyon, 2003). Wat de doorwerking van raciale surveillancepraktijken in biometrische technologieën nog eens extra problematisch maakt, is dat raciale surveillance niet alleen doorwerkt in de biometrische technologieën die bedoeld zijn voor surveillance, maar ook in degenen die daar niet voor bedoeld zijn. Sommige biometrische technologieën die gebruikt werden voor handenvrije toiletten, kranen en handdrogers, blijken bijvoorbeeld niet in staat om de handen van zwarte mensen te detecteren (Browne, 2015: 109).

5. Verklaring voor de doorwerking van *racializing surveillance*

De vraag rest nu hoe de doorwerking van *racializing surveillance* in digitale technologieën kan worden verklaard. Volgens Browne (2015: 116) ligt het antwoord op deze vraag in het gegeven dat technologieën bedacht en ontworpen worden door echte mensen: "Machines are designed and operated by real people to sort real people." De echte mensen die aan de basis staan van vele digitale technologieën zijn voornamelijk witte mannen: "Whereas men shape technology, they shape it to the exclusion of women, especially of Black women," aldus Noble (2018: 107). Uit het meest recente *Diversity Report* van Google (2020: 60) bleek bijvoorbeeld dat zwarte vrouwen in dat jaar slechts 1,6 procent van het totale aantal werknemers uitmaakte en slechts 0,7 procent van het aantal werknemers met een technische functie. Het gebrek aan diversiteit onder programmeurs is problematisch, omdat hierdoor de perspectieven, denkbeelden en vooroordelen van witte mannen van grotere invloed zullen zijn op de algoritmes, dan die van zwarte vrouwen. In de volgende paragraaf zal ik ingaan op het oplossen van dit probleem. Eerst rest er nog een andere vraag die we moeten stellen: waarom werken er zo weinig zwarte vrouwen in de ICT?

Verschillende auteurs, zoals Nakamura (2002) en Sinclair (2004) wijzen als deel van het antwoord op deze vraag naar het stereotype van 'technoprimitiviteit'. In de achttiende eeuw verscheen er veel literatuur waarin vermeld stond dat Afrikaanse beschavingen achterliepen op westerse in hun ontwikkeling van wetenschap en technologie. Hierdoor was het gemakkelijk voor Amerikanen en Europeanen om te geloven dat zwarte mensen met technische competenties niet bestonden. Het stereotype van technoprimitiviteit was te verenigen met het stereotype van zwarte vrouwen als domme Mammy's. Slavenhouders benadrukten bijvoorbeeld dat sommige

slaven hun gereedschap kapot hadden gemaakt, niet uit opstandigheid, maar omdat zij niet wisten hoe ze het moesten gebruiken (Sinclair, 2004: 4-5).

Tegen het einde van de negentiende eeuw was het idee dat witte mannen technisch capabel en inventief waren en dat zwarte mensen, in het bijzonder zwarte vrouwen, niet technisch onderlegd waren wijdverbreid geaccepteerd (Sinclair, 2004: 3-4). Vanwege dit stereotype van technoprimiviteit, kregen Afro-Amerikanen slechts beperkt toegang tot onderwijs, kregen zij geen controle over complexe machines en geen toegang tot het patentrecht (Sinclair, 2004: 2).

Lisa Nakamura, een vooraanstaande wetenschapper in het onderzoek naar 'ras' in digitale media, betoogt dat het internet oude raciale stereotypes, waaronder dit stereotype van technoprimitiviteit, verbreid, versterkt en 'hercodeert'. Dit noemt zij *cybertyping* (Nakamura, 2009: 3 en 18). Ze schrijft dat het internet Aziaten representeert als superieure werkkrachten, terwijl zwarte mensen worden *gecybertyped* als technoprimitieve mensen, die daarom niet geschikt zijn om te werken in de ICT. De aanwas van Aziatische immigranten naar de Amerikaanse techbedrijven creëert een vals gevoel van raciale gelijkheid, maar ondertussen worden de zwarte Amerikanen niet verwelkomt. De nationale rassenproblematiek wordt zo onder het vloerkleed geschoven. Zwarte Amerikanen kunnen zich nog steeds niet ontdoen van het stereotype van technoprimitiviteit (Nakamura, 2009: 20-22).

6. Naar een handelingsperspectief

Wat kunnen we nu doen om discriminatie van zwarte vrouwen door profilerende algoritmes tegen te gaan? Met dit hoofdstuk heb ik laten zien dat discriminatie en racisme diep verankerde maatschappelijke problemen zijn. Technologische oplossingen voor discriminatie en racisme zullen daarom alleen toereikend zijn, als aan deze oplossingen een mentaliteitsverandering vooraf gaat, een bewustwording van de bezwaarlijke stereotypes over zwarte vrouwen. Zonder zo'n mentaliteitsverandering, blijft de kans altijd bestaan dat degenen die nieuwe technologieën ontwerpen of degenen die technologieën toepassen zich nog steeds schuldig maken aan discriminatie en racisme (Hooyman, 2019: 17).

Concreet wijzen inzichten uit dit hoofdstuk richting twee oplossingen: ten eerste moeten witte mannelijke programmeurs bewust worden gemaakt van hun stereotypes ten aanzien van zwarte vrouwen. Ten tweede moet het zwarte vrouwen gemakkelijker worden gemaakt om in de techwereld te gaan werken. De eerste oplossing leidt in de praktijk gelijk tot moeilijkheden. Natuurlijk is het de taak van witte mannelijke programmeurs om zich bewust te worden van hun eigen vooroordelen, maar sommige programmeurs betogen dat zij zich niet bewust kunnen worden van de ervaring om als zwarte vrouw gediscrimineerd te worden, omdat zij zelf niet weten hoe het is om een zwarte vrouw te zijn. Zij vertelden bijvoorbeeld dat "My brain database doesn't see skin color" (Daniels, 2015: 1328) en dat "It only seems commonsensical, that if you go ten years in the high-tech corporate world and extremely rarely in your daily life ever encounter an American black person, they disappear from your awareness" (Nakamura, 2009: 22).

Ik denk daarom dat we vooral moeten inzetten op de tweede oplossing: meer zwarte vrouwen in de techwereld. "Wat zwarte mensen nodig hebben,"zo oreerde de bekende Afro-Amerikaanse activist Du Bois al, "zijn mogelijkheden en middelen voor waardige arbeid" (Baker, 1994: 12). Omdat zwarte vrouwen al eeuwenlang het doelwit zijn van tal van stereotyperingen, hebben zij een 'dubbel bewustzijn' ontwikkeld: zij bekijken zichzelf niet alleen vanuit hun eigen ogen, maar internaliseren ook altijd het beeld dat witte mannen van hen hebben. Juist vanwege dit dubbele bewustzijn hebben zwarte vrouwen een uniek standpunt van waaruit zij kunnen bijdragen aan het verbeteren van hun eigen positie (Collins, 1990).

Om mijn argument dat er meer zwarte vrouwen in de techwereld zouden moeten werken kracht bij te zetten, wil ik tenslotte een vergelijking maken met de fotografie. De feministe bell hooks[7] betoogt dat de fotografie cruciaal is geweest in de strijd van zwarte mensen om *zichzelf* te representeren in plaats van zich te laten representeren door witte mensen. "It gave them a means to participate fully in the production of images." Bovendien stelde de camera hen in staat om zichzelf te representeren als technisch competente mensen. Het werkte dus op twee manieren: "Cameras in black hands allow for the creation of an alternative image, and that image also

[7] De schrijfster bell hooks is geboren als Gloria Jean Watkins. Ze koos als pseudoniem de naam van haar overgrootmoeder, Bell Blair Hooks. De schrijfster spelt de naam zelf zonder hoofdletters, omdat ze vindt dat het de inhoud van haar werk is die telt en niet haarzelf als persoon (RoSa, z.d.).

enables African-Americans to represent themselves as skillful in the management of those technologies" (Sinclair, 2004: 10). Ik durf te voorspellen dat algoritmes in handen van zwarte vrouwen dezelfde effecten zullen hebben.

Conclusie

In dit hoofdstuk heb ik onderzocht hoe historisch verklaard kan worden dat juist zwarte vrouwen door profilerende algoritmes worden gediscrimineerd. Ik heb betoogd dat de oorsprong van deze discriminatie gevonden kan worden in de achttiende eeuw, de eeuw waarin slavernij op haar hoogtepunt was en waarin tegelijkertijd een surveillancesamenleving ontstond. Deze ontwikkelingen waren onlosmakelijk met elkaar verbonden en bleven dat ook na de achttiende eeuw. Dit kwam onder meer door het creëren en het voortbestaan van stereotypes over zwarte mensen, in het bijzonder over zwarte vrouwen. Deze stereotypes – zoals de Jezebel, de Mammy en de Sapphire – werden door sommige witte mensen ingezet ter rechtvaardiging van hun gewelddadige en controlerende behandeling van zwarte mensen.

Vanwege de onlosmakelijke verbondenheid tussen 'ras' en surveillance spreekt Browne, over *racializing surveillance*. Zij dateert dit begrip niet alleen in het verleden, maar beschrijft ook hoe historische praktijken van *racializing surveillance* doorwerken in hedendaagse digitale technologieën. Ik heb dit geïllustreerd met voorbeelden zoals de fotografie en biometrische technologieën. Een verklaring voor de doorwerking van *racializing surveillance* in digitale technologieën heb ik gezocht in de mensen achter deze technologieën. Digitale technologieën worden voornamelijk ontworpen door witte mannen die zich onbewust zijn van hun stereotypes over zwarte vrouwen. Volgens Nakamura werken er weinig zwarte vrouwen in de techwereld, omdat het internet oude raciale stereotypes zoals het stereotype van technoprimitiviteit 'hercodeert'. Zwarte vrouwen worden daarom niet geschikt geacht om te werken in de techwereld. Ik heb betoogd dat, om de discriminatie van zwarte vrouwen op te lossen, het zwarte vrouwen ten eerste makkelijker gemaakt moet worden om in de ICT te gaan werken en dat ten tweede witte mannelijke programmeurs bewust moeten worden gemaakt van hun stereotypes jegens zwarte vrouwen.

Uiteraard heb ik in dit hoofdstuk slechts een fractie van de geschiedenis over ‘ras’ en technologie kunnen incorporeren. Meer onderzoek naar de relatie tussen deze geschiedenissen is zeker nodig. Het is bijvoorbeeld opvallend dat er tijdens de opkomst van de eerste computers aan het einde van de jaren ‘30 veel zwarte vrouwen werkten als programmeur. Het stereotype van technoprimitiviteit speelde toen blijkbaar een ondergeschikte rol in het aanstellen van programmeurs. Ergens in de tweede helft van de twintigste eeuw is de representatie van zwarte vrouwen in de ICT flink achteruit gegaan. Sommige artikelen geven hiervoor als verklaring dat computers steeds vaker op de markt gebracht werden als iets voor mannen (NPR, 2014; Gold, 2018). Het blijft echter onbekend wanneer de representatie van zwarte vrouwelijke programmeurs precies achteruit is gegaan en welke rol het stereotype van technoprimitiviteit hier mogelijk in heeft gespeeld.

Dit hoofdstuk heeft geen concrete voorstellen gegeven voor hoe het zwarte vrouwen gemakkelijker kan worden gemaakt om in de ICT te gaan werken en hoe witte mannelijke programmeurs bewust kunnen worden gemaakt van hun stereotypes. Dit valt buiten het disciplinaire domein van dit hoofdstuk. Toch hoop ik dat interdisciplinaire wetenschappers wel kunnen voortbouwen op mijn hoofdstuk in het ontwikkelen van concrete voorstellen.

Hoofdstuk 2: Is het wel zo eerlijk?

Over de onderliggende problematiek bij het definiëren van een eerlijk profilerend algoritme

Eline Kempkes, Kunstmatige Intelligentie

In profilerende algoritmen is vaak niet expliciet geprogrammeerd dat zij keuzes moeten maken op basis van gender of etniciteit, maar toch kunnen deze algoritmen discriminerend zijn tegen bijvoorbeeld zwarte vrouwen (Barocas & Selbst, 2018; Buolamwini & Gebru, 2018; Bolukbasi, Chang, Zou, Saligrama & Kalai, 2016). Aangezien deze algoritmes schadelijk en ongewenst zijn, rijst nu de vraag vanuit de Kunstmatige Intelligentie waarom er geen goede technische mogelijkheden zijn om een eerlijk (*fair)* algoritme te ontwerpen (Angwin, Larson, Mattu & Kirchner, 2016; Bolukbasi, Chang, Zou, Saligrama & Kalai , 2016). Een verklaring hiervoor is dat er nog geen consensus is bereikt over de definitie van eerlijkheid (*fairness*). Om een algoritme te ontwerpen dat in alle opzichten echt eerlijk is, moet er eerst een praktische definitie voor eerlijkheid zijn. Dan kan het algoritme daar naar streven. Op basis van literatuuronderzoek zullen er in dit hoofdstuk verschillende problemen worden geanalyseerd die zich voordoen bij het definiëren van een eerlijk profilerend algoritme. Centraal zal staan welke verschillende definities er zijn voor eerlijkheid en wat hier de argumenten voor en tegen zijn, om zo inzicht te verschaffen in de complexiteit van het probleem. Omdat een definitie van eerlijk voor alle gemarginaliseerde groepen tegelijk moet gelden, wordt er in dit hoofdstuk niet specifiek op zwarte vrouwen gericht maar op gemarginaliseerde groepen in het algemeen. De concrete vraag die wordt onderzocht luidt: "*Wat is de onderliggende problematiek bij het definiëren van een profilerend algoritme dat gemarginaliseerde groepen eerlijk behandeld*?".

Ten eerste zal er een korte inleiding gegeven worden in profilerende algoritmen, om zo inzicht te krijgen in de termen die worden gebruikt. Ook zal hiermee worden verklaard hoe het kan dat sommige profilerende algoritmes discriminerend zijn, opdat er inzicht komt in de oorzaak van het probleem. De termen en mechanismen die in deze paragraaf worden besproken, zijn nodig in paragraaf 2. Hier worden de drie meest gangbare definities van eerlijkheid gegeven

en daarbij worden hun onderliggende individuele problemen besproken. Ten slotte worden er drie overkoepelende problemen aangekaart, die voor alle drie de definities van eerlijkheid gelden. Hierbij worden ook studies aangehaald die mogelijke oplossingen bieden. Het eerste probleem is dat er een bepaalde compromis bestaat tussen het behalen van eerlijkheid en accuratesse. Het volgende probleem betreft de onverenigbaarheid van de drie definities. Het laatste probleem is het gebrek aan intersectionaliteit van alle drie de definities.

1. Werking profilerende algoritmen

Het is eerst nodig om de werking profilerende algoritmen te bespreken en te bekijken hoe het kan dat sommige algoritmes discriminerende beslissingen maken jegens gemarginaliseerde groepen. Daarna kan met behulp van de begrippen die hier worden gedefinieerd het begrip eerlijkheid nader worden toegelicht, om vervolgens de onderliggende problematiek te bestuderen.

Bij machine learning wordt er gebruik gemaakt van een set trainingsdata om bepaalde patronen te ontdekken. Profilerende algoritmen maken gebruik van machine learning om bepaalde individuen te classificeren, om daarmee een voorspelling te doen (Alpaydin, 2020**).** Een voorbeeld hiervan is een gezichtsherkennings-applicatie die voorspelt hoe oud iemand is, het classificeert individuen in een bepaalde leeftijdscategorie. Het doel van classificatie is om uiteindelijk op basis van een aantal gegevens X van een individu, te voorspellen wat de waarde van de onbekende variabele Y is.

Een voorbeeld van een toepassing hiervan is of op basis van alle gegevens X een bedrijf een bepaalde sollicitant zal aannemen. Hierbij kunnen de gegevens X de werkervaring en opleidingsniveau van de sollicitant zijn. De onbekende variabele Y is in dit geval het wel of niet aannemen van de sollicitant: het doel is dat het algoritme dit gaat voorspellen. Het algoritme is hiervoor eerder getraind op een grote dataset. In deze trainingsset zit data van eerdere sollicitanten die het bedrijf heeft aangenomen of afgewezen. Bij al deze sollicitanten zijn de gegevens X (de werkervaring en het opleidingsniveau) gekoppeld aan Y (het wel of niet aangenomen zijn). Het doel van het algoritme om deze trainingsdata te gebruiken en op basis daarvan te voorspellen of het bedrijf een sollicitant aanneemt: dus om de waarde van Y te voorspellen (Barocas, Hardt & Narayanan, 2017).

Dit bovenstaande is een voorbeeld van een classificatie-algoritme: een individu wordt aan een klasse gekoppeld (wel of niet aangenomen) (Alpaydin, 2020). Om te kunnen classificeren vat het algoritme alle gegevens X samen in één variabele, de score R (Barocas, Hardt & Narayanan, 2017). De score R is de voorspelling die de machine maakt voor de variabele Y, het is dus gebaseerd op alle gegevens X. De functie die de score berekent, die dus alle gegevens X samenvat, wordt gedefinieerd met behulp van de trainingsdata. Deze uitgerekende score R is dus de voorspelling voor de doelvariabele Y gegeven de observaties X (Barocas, Hardt & Narayanan, 2017). Het algoritme rekent een bepaalde score R uit voor de sollicitant en als deze hoog genoeg is, zal het algoritme voorspellen dat de sollicitant wordt aangenomen.

Een belangrijk begrip bij profilerende algoritmen is de accuratesse. De accuratesse een maat voor hoe goed het classificatie-algoritme werkt. De accuratesse is het percentage dat de voorspelling van de waarde van de doelvariabele Y correct was (Barocas, Hardt & Narayanan, 2017).

Nu de werking van de profilerende algoritmes is belicht, rijst de vraag wat de oorzaak is van de discriminatie van algoritmes. De belangrijkste verklaring is dat een algoritme net zo subjectief is als de data waarop deze is gebaseerd (Barocas & Selbst, 2016). Wanneer in data bepaalde vooroordelen zijn verborgen, zal het algoritme deze reproduceren. Een resultaat hiervan is dat mannen en vrouwen door hetzelfde algoritme anders kunnen worden beoordeeld. Een voorbeeld hiervan wordt gegeven door een onderzoek van Lahoti, Preethi, Gummadi en Weikum (2019) naar het sociale medium XING, een platform om beroepsmatige contacten te onderhouden. Het algoritme gaf de profielen van vrouwen een lager cijfer dan mannen, ook al waren de vrouwen beter gekwalificeerd. Doordat er in de data bepaalde vooroordelen tegenover vrouwen zaten, maakte het algoritme van XING discriminerende keuzes. In deze reproductie komen vooroordelen uit de oude data terug in de nieuwe data, waardoor het bepaalde stereotypen kan versterken (Celis, Straszak & Vishnoi, 2017).

De begrippen en mechanismen achter profilerende algoritmen zijn belicht en het is duidelijk dat subjectieve data belangrijke oorzaken van discriminerende algoritmen zijn. Nu kunnen in de

volgende paragraaf deze begrippen en termen worden gebruikt om eerlijkheid te definiëren, om vervolgens hun problemen omtrent discriminatie te analyseren.

2. De verschillende definities van eerlijkheid en hun beperkingen

Om eerlijke algoritmes te ontwerpen is het van belang om het begrip 'eerlijkheid' een duidelijke definitie te geven. Dan kan er worden gestreefd naar eerlijkheid en kan er worden gecontroleerd of het algoritme daadwerkelijk daar aan voldoet. Er zijn de laatste jaren meer dan twintig verschillende definities voorgesteld, maar er is nog geen consensus bereikt (Verma & Rubin, 2018; Binss, 2017). Het begrip 'eerlijkheid' wordt vaak omschreven als 'niet-discriminerend'. Er wordt gekeken of het algoritme voorspellingen maakt die onafhankelijk zijn van bijvoorbeeld de etniciteit of gender van het individu. Hoe deze onafhankelijkheid exact wordt ingevuld hangt van de precieze definitie van eerlijkheid af.

De meeste definities van eerlijkheid kunnen drie verschillende categorieën worden ingedeeld (Verma & Rubin, 2018; Corbett-Davies & Goel 2018). De overige definities van eerlijkheid zijn ofwel equivalent met een van deze drie definities, ofwel een versoepeling ervan (Barocas, Hardt & Narayanan, 2017). Om inzicht te krijgen in dit begrip zullen de drie noties van eerlijkheid in deze paragraaf worden besproken, evenals kritiek op deze definities. In paragraaf 3 zullen er drie overkoepelende problemen van deze definities worden besproken.

Bij deze definities worden de variabelen A, Y en R gebruikt. A is de *sensitive attribute*, een variabele uit de gegevens X van het individu. A is de eigenschap van het individu waarop het algoritme niet moet discrimineren, bijvoorbeeld gender of etniciteit. Y is de doelvariabele die wordt voorspeld en R de score die het algoritme geeft.

2.1 Onafhankelijkheid (*Independence)*

Deze invulling van eerlijkheid eist het volgende: het percentage van de mensen dat als 'positief' is geclassificeerd moet gelijk zijn bij beide groepen (Binns, 2017). Stel dat een algoritme voorspelt of iemand een jurk gaat kopen. Het algoritme voorspelt bij 60% van de vrouwen dat zij een jurk zal kopen. Het algoritme is eerlijk volgens onafhankelijkheid als er ook bij alle andere genderidentiteiten wordt voorspeld dat 60% van deze groep een jurk zal kopen.

Dat wil zeggen dat dus ook bij de groep mannen moet worden voorspeld dat 60% een jurk gaat kopen.

In andere woorden: bij onafhankelijkheid moet gelden dat de score R onafhankelijk is van de waarde van A, de sensitive attribute (Barocas, Hardt & Narayanan, 2017). In dit voorbeeld was dat genderidentiteit: wat die waarde ook is, het zou niet uit moeten maken voor de score die het algoritme geeft aan het individu. Een individu met als waarde voor A "vrouw" krijgt dezelfde score als de waarde voor A = "man". De formele definitie luidt:

$$P(R = r \mid A = a) = P(R = r \mid A = b) \qquad \forall r \in R,\ \forall a, b \in A$$

Onafhankelijkheid kijkt dus alleen naar de voorspelde uitkomst van het algoritme (Verma & Rubin, 2018). Wat de daadwerkelijke waarde is van de doelvariabele Y (of iemand wel of geen jurk koopt), doet er in deze definitie niet toe: onafhankelijkheid kijkt alleen naar wat het algoritme heeft voorspeld en of dat eerlijk is.

De kritiek op deze definitie van eerlijkheid is dat het te kort door de bocht is en geen rekening houdt met verschillen die tussen groepen bestaan (Binss, 2017). Een bekend voorbeeld hierbij is het COMPAS-algoritme, dat de kans op recidive van gevangen voorspelt. Dit algoritme zal later nog uitgebreider aan bod komen. Het COMPAS-algoritme voldoet aan onafhankelijkheid en is dus in een bepaalde zin eerlijk. Echter, COMPAS negeert het feit dat statistisch gezien vrouwen minder vaak recidive plegen dan mannen en heeft daarom de neiging om het recidive risico van vrouwen te overschatten. Doordat het algoritme genderneutraliteit nastreeft, wordt de recidivekans van vrouwen niet goed voorspeld en wordt juist daardoor het algoritme op een bepaalde manier oneerlijk (Corbett-Davies & Goel 2018).

2.2 Scheiding (*Separation*)

In tegenstelling tot onafhankelijkheid, houdt de definitie van scheiding wel rekening met de echte waarde van de doelvariabele Y (Verma & Rubin, 2018). Hierbij wordt er dus ook gekeken of het individu de jurk daadwerkelijk heeft gekocht. Daardoor wordt er rekening gehouden met het feit dat A een correlatie kan hebben met de doelvariabele Y: dus dat blijkt dat vrouwen

gemiddeld gezien vaker jurken kopen dan mannen. Om die reden is het in dat geval niet problematisch dat het algoritme dit zo voorspelt.

Scheiding houdt onder andere in dat de sensitiviteit dezelfde waarde heeft bij elke groep. (Barocas, Hardt & Narayanan, 2017). De sensitiviteit is het percentage dat het algoritme Y als positief voorspelt en het daadwerkelijk positief is. In het voorbeeld is de sensitiviteit het aantal keer dat het algoritme voorspeld heeft dat iemand een jurk zou kopen, gedeeld door het aantal individuen dat de jurk daadwerkelijk koopt.

Naast de eis voor sensitiviteit, moeten alle groepen ook dezelfde waarde hebben voor de vals positieve ratio om aan scheiding te voldoen. Wanneer het algoritme ten onrechte een individu als positief heeft geclassificeerd, is het een vals positieve. In het voorbeeld heeft het algoritme dan voorspeld dat het individu de jurk zal kopen, maar heeft het individu dat niet gedaan. De vals positieve ratio wordt berekend door het aantal vals positieven te delen door het totaal aantal negatieven. Wanneer een algoritme eerlijk is volgens de definitie van scheiding, zijn de sensitiviteit en de vals positieve ratio hetzelfde bij alle groepen (Barocas, Hardt & Narayanan, 2017).

Hieronder zijn de formele definities weergegeven van seperation, dus dat de kans op de score R = 1, bij beide groepen hetzelfde moet zijn, als Y het geval is (sensitiviteit). De kans op de score R = 1 moet ook gelijk zijn als Y niet het geval is (vals positieve ratio).

$$P(R = 1 \mid Y = 1, A = a) = P(R = 1 \mid Y = 1, A = b) \qquad \forall a, b \in A$$

$$P(R = 0 \mid Y = 1, A = a) = P(R = 1 \mid Y = 0, A = b) \qquad \forall a, b \in A$$

Corbet-Davies en Goel (2018) bespreken de beperkingen van scheiding. De kern van deze kritiek is dat er bij verschillende groepen de waarden van Y anders verdeeld zijn. Hiermee wordt niet alleen bedoeld dat het gemiddelde van deze groepen anders is, maar ook de standaardafwijking. Wanneer de standaardafwijking van de ene groep veel groter is, zal het algoritme deze groep minder vaak accuraat voorspellen en zal de sensitiviteit ook lager zijn. Ook wanneer er geen sprake is van discriminatie door het algoritme, kunnen de cijfers van

sensitiviteit anders zijn tussen de twee groepen: omdat de verdelingen er anders uitzien. Dit fenomeen heet het probleem van de infra-marginality (Simoiu, Corbett-Davies & Goel, 2017).

Echter, deze kritiek van Corbet-Davies en Goel neemt aan dat de maatschappij individuen gelijk behandelt, ongeacht hun etniciteit of gender, waardoor de verschillen in data en voorspellende algoritmen legitiem zijn. Een reactie hierop is te vinden vanuit de intersectionaliteit. Dit is een theoretische benadering die stelt dat deze verdelingen verschillen door systematische en structurele onderdrukking zoals racisme en seksisme. Dat er verschillen zijn tussen deze twee groepen komt door oneerlijke maatschappelijke processen en daarom is het niet gerechtvaardigd om ze om die reden anders te behandelen (Foulds, Islam, Keya & Pan, 2018). Deze kritiek vanuit intersectionaliteit zal in paragraaf 3 uitgebreider worden besproken..

2.3 Toereikendheid (*Sufficiency)*

Toereikendheid hangt van zowel de echte uitkomst Y af, als de gegeven score R (Verma & Rubin, 2018). Deze definitie houdt in dat alle scores R dezelfde betekenis hebben voor alle groepen. Een voorbeeld hierbij is dat wanneer een recidive-algoritme dezelfde score toekent aan een wit iemand en aan iemand van kleur, dat deze mensen in de praktijk inderdaad dezelfde kans op recidive hebben (Corbett-Davies & Goel 2018).

Wiskundig gezien betekent het dat gegeven de waarde voor de score R, Y en A onafhankelijk van elkaar moeten zijn. Het verschil met onafhankelijkheid is nu dus dat de score R bij deze definitie is betrokken: wanneer de scores van twee individuen uit verschillende groepen gelijk aan elkaar zijn, moet de echte waarde voor de doelvariabele Y ook gelijk aan elkaar zijn (Barocas, Hardt & Narayanan, 2017). De formele definitie luidt:

$$P(Y = q \mid R = r, A = a) = P(Y = q \mid R = r, A = b) \qquad \forall q \in Y,\ r \in R,\ \forall a, b \in A$$

De kritiek op toereikendheid luidt dat het een belangrijke eis is, maar niet genoeg om ervoor te zorgen dat het algoritme accuraat en eerlijk is. Hiervoor wordt de volgende hypothetische situatie geschetst: er is een bank die wil voorspellen of mensen hoge schulden hebben. In deze hypothetische situatie zijn er twee voorwaarden: 1) witte mensen en mensen van

kleur die in dezelfde buurt wonen hebben ongeveer even hoge schulden, 2) mensen van kleur wonen in buurten waar relatief veel mensen met schulden wonen. Wanneer de bank alleen kijkt naar de buurt waarin iemand woont, dan kan de bank zowel discrimineren op huidskleur als aan toereikendheid voldoen. Omdat witte mensen en mensen van kleur in dezelfde buurt, dezelfde score krijgen en daardoor hetzelfde behandeld worden, wordt er voldaan aan toereikendheid. Echter, omdat mensen van kleur vaak in andere buurten wonen en het onderscheid wordt gemaakt op woonbuurt, wordt er alsnog gediscrimineerd op huidskleur (Corbett-Davies & Goel 2018). Deze hypothetische situatie laat zien dat toereikendheid niet voldoende is om discriminatie uit te sluiten.

De notie eerlijkheid draait dus vooral om of de voorspellingen van het algoritme niet afhangen van bijvoorbeeld de etnische groep waarin iemand zich bevindt. Elke definitie van eerlijkheid heeft zijn eigen beperkingen, die het ontwerpen van een echt eerlijk algoritme belemmeren. In de volgende paragraaf komen drie overkoepelende problemen van de drie definities van eerlijkheid, waardoor de onderliggende problematiek nog meer zichtbaar wordt.

3. Overkoepelende problemen

3.1 Compromis tussen eerlijkheid en accuratesse

Een van de moeilijkheden is dat wanneer het algoritme eerlijker wordt, de accuratesse vaak omlaag gaat (Menon & Williamson, 2018). Dit komt omdat het algoritme nu twee doelen heeft: zo accuraat mogelijk zijn en niet discrimineren. Omdat beide doelen geoptimaliseerd moeten worden, moet er een bepaalde compromis hiertussen worden gemaakt (Zafar, Valera, Rodriguez, Gummadi, 2015). Er zijn inmiddels wel studies gedaan om te onderzoeken of het mogelijk is om eerlijkheid te behalen en ook nog accuraat te zijn, bijvoorbeeld door Fish, Kun en Lelkes (2016). Echter, vaak zal de accuratesse omlaag gaan wanneer er naar eerlijkheid wordt gestreefd (Zafar, Valera, Rodriguez, Gummadi, 2015).

3.2 Definities zijn onverenigbaar

Het tweede probleem houdt in dat de drie definities van eerlijkheid in de praktijk tegenstrijdig zijn met elkaar, waardoor er nooit aan alledrie tegelijkertijd kan worden voldaan. Hierbij zal het voorbeeld van COMPAS worden besproken, een bekend algoritme waarmee de discussie duidelijk in beeld wordt gebracht.

De drie verschillende definities van eerlijkheid zijn onder bepaalde voorwaarden tegenstrijdig met elkaar en kunnen om die reden niet tegelijkertijd worden behaald. Wanneer A en Y statistisch gezien afhankelijk van elkaar zijn, dan kan er nooit tegelijkertijd aan onafhankelijkheid en toereikendheid worden voldaan (Barocas, Hardt & Narayanan, 2017). Wanneer bijvoorbeeld vrouwen gemiddeld vaker jurken kopen, dan kan een algoritme niet onafhankelijk en toereikend tegelijkertijd zijn. Ook andere combinaties van de drie definities kunnen onder bepaalde restricties niet samengaan (Barocas, Hardt & Narayanan, 2017).

Een algoritme dat dit fenomeen in de praktijk bracht is het eerder besproken COMPAS-algoritme. COMPAS wordt veel gebruikt in de Verenigde Staten om de kans op recidive van een veroordeelde te voorspellen. Dit algoritme heeft veel discussie gewekt, omdat het enerzijds wel eerlijk was, maar anderzijds niet: het laat zien hoe moeilijk het in de praktijk is om een echt eerlijk algoritme te maken.

In de trainingsdata van COMPAS wordt niet expliciet etniciteit gebruikt, maar wel andere aspecten uit de data die hiermee een correlatie hebben. Uit een analyse van Angwin, Larson, Mattu en Kirchner (2016) blijkt dat het algoritme een vals positieve ratio had van 44.9% bij zwarte mensen, en 23.5% bij witte mensen. Dit houdt in dat bij zwarte mensen die geen recidive pleegden, het 44.9% van de keren verkeerd voorspeld was. Dit percentage ligt bijna twee keer zo hoog als bij witte mensen. Volgens de notie van scheiding is dit algoritme dus niet eerlijk, want de vals positieve ratio van beide groepen zijn niet gelijk aan elkaar.

Echter, er kwam veel kritiek op deze analyse (Dieterich, Mendoza & Brennan 2016; Barenstein, 2019). Een van deze kritiekpunten is dat er bij het algoritme wel voldaan werd aan toereikendheid. Wanneer de scores hetzelfde zijn, zijn etniciteit en de kans op recidive onafhankelijk van elkaar. Bij veroordeelden die een score van 7 kregen op de compas-schaal, is in het echt 60% van de witte mensen in de fout gegaan, en 61 % van de zwarte mensen. De score

heeft dus dezelfde betekenis voor witte mensen als voor zwarte mensen. Omdat deze percentages nagenoeg gelijk zijn, is in dit voorbeeld voldaan aan toereikendheid.

Het conflict zit in het gebruik van de verschillende definities van eerlijkheid: op de ene manier is het algoritme wel eerlijk, op de andere manier niet. Het blijkt in dit geval onmogelijk te zijn om aan beide definities van eerlijkheid volledig te voldoen, omdat in de data die het algoritme gebruikt het statistisch gezien zo is dat zwarte mensen een hoger recidivepercentage hebben (Dressel & Farid, 2018). Toereikendheid en scheiding sluiten elkaar in deze situatie uit.

Dit voorbeeld toont aan dat het in de praktijk lastig is om aan eerlijkheid te voldoen: de verschillende definities zijn onverenigbaar met elkaar. Volgens Kleinberg, Mullainathan en Raghavan (2016) kan alleen in hele speciale gevallen voldaan worden aan alledrie de voorwaarden tegelijkertijd. Er is dus geen technische oplossing mogelijk om in alle andere gevallen aan alle drie de definities tegelijkertijd te voldoen en er zal in de praktijk gekozen moeten worden voor een van de definities.

3.3 De definities zijn niet intersectioneel

Een andere kritiek op deze huidige definities van eerlijkheid komt vanuit de theorie over intersectionaliteit (Foulds, Islam, Keya & Pan, 2018; Hoffmann, 2019). Het centrale idee is dat eerlijkheid en vooroordelen in algoritmen worden gekoppeld naar de bredere context van eerlijkheid en vooroordelen in de maatschappij (Foulds, Islam, Keya & Pan, 2018). Het algoritme moet er rekening mee houden dat de maatschappij individuen niet gelijk behandelt. Daarnaast is het samenspel tussen verschillende tussen verschillende vormen van onderdrukking belangrijk, zoals de interferentie van racisme en seksisme. Intersectionele kritiek is al kort aan bod gekomen in paragraaf 2 en zal nu uitgebreider worden besproken.

Intersectionaliteit is een analytische manier om maatschappelijke ongelijkheid te bekijken, waarbij benadrukt wordt dat verschillende vormen van onderdrukking met elkaar kunnen interfereren (Foulds, Islam, Keya & Pan, 2018). Het resultaat van deze interferentie is dan groter dan de som der delen. Zo kan een vrouw van kleur last hebben van seksisme en racisme, maar juist ook van de interactie tussen deze twee factoren.

In het artikel van Kearns, Neel, Roth en Wu (2017) wordt er vanuit een intersectioneel standpunt kritiek gegeven op de huidige definities van eerlijkheid door de term *fairness gerrymandering*. De kern van zijn kritiek is dat er bij de doorgaande definities van eerlijkheid van tevoren een klein aantal *sensitive attributes* worden gedefinieerd, zoals etniciteit en gender. Vervolgens wordt er gekeken of er bij deze sensitive attributes de groepen niet verschillend behandeld worden. Echter, door deze techniek kan er fairness gerrymandering ontstaan. Dit concept wordt door Kearns et al. (2017) uitgelegd door de volgende situatie: stel er is een classificerend algoritme dat een individu labelt als positief of als negatief. Een individu wordt positief gelabeld als en alleen als het gaat om een man van kleur of een witte vrouw. Wanneer de populatie uniform is verdeeld, zal het algoritme 50% van de mannen labelen als positief en 50% van de vrouwen. Ook wordt 50% van de mensen van kleur beoordeeld als positief, en 50% van de witte mensen. Dit algoritme is eerlijk volgens het criterium van onafhankelijkheid en dus niet discriminerend. Echter, geen van de vrouwen van kleur wordt gelabeld als positief, omdat het alleen mannen van kleur en witte vrouwen als positief labelt. Het algoritme lijkt op het eerste gezicht niet seksistisch of racistisch, maar door de wisselwerking tussen deze twee factoren worden vrouwen van kleur toch niet eerlijk behandeld (Kearn et al., 2017).

Volgens Kearns et al. (2017) is dit voorbeeld niet volledig hypothetisch en gebeurt ook in de praktijk dat bij subgroepen geen eerlijkheid wordt behaald. Zo laten Buolamwini en Gebru (2018) zien dat bij gezichtsherkennings-applicatie vrouwen van kleur het minst vaak herkend worden, nog minder vaak dan vrouwen in het algemeen of mensen van kleur in het algemeen. Deze subgroep heeft de laagste accuratesse. Wanneer een algoritme ook op een intersectionele manier eerlijk moet zijn, zal er dus ook moeten worden gekeken naar subgroepen.

Er worden nog veel studies gedaan naar de intersectionele aanpak van eerlijk algoritmes. Zo hebben Foulds, Islam, Keya & Pan (2018) een aantal criteria opgesteld voor een intersectioneel en eerlijk algoritme. Ook Kearn et al. (2017) hebben een methode bedacht waarbij er expliciet werd gekeken naar subgroepen. Buolamwini en Gebru (2018) bedachten een methode om de vooroordelen in gezichtsherkennings-algoritmen te onderzoeken, met oog voor de subgroepen.

Kritiek op de studies van Kearn et al. (2017) en van Buolamwini en Gebru (2018) wordt gegeven door Hoffmann (2018). Zij schrijft dat er niet alleen gekeken moet worden of subgroepen niet nog harder worden geraakt, maar dat intersectionaliteit ook draait om het aankaarten van de sociale hiërarchie die deze ongelijkheid heeft veroorzaakt. Hierin valt ook de kritiek die eerder werd gegeven op de definitie van toereikendheid: er wordt geen rekening gehouden dat bepaalde verschillen tussen groepen wordt veroorzaakt door onderdrukking van de maatschappij (Foulds, Islam, Keya & Pan, 2018). Het feit dat vrouwen minder snel recidive plegen dan mannen, is niet inherent aan vrouwen, maar komt door de vorming in de maatschappij, stelt deze kritiek. Een andere eis voor een intersectioneel eerlijk algoritme is dat er dus wordt aangenomen dat systematische verschillen tussen groepen ontstaan zijn door de maatschappij en dat deze verschillen rechtgezet moeten worden (Foulds, Islam, Keya & Pan, 2018). Een echte intersectioneel en eerlijk algoritme is tot dusver nog niet gevonden.

Conclusie

Profilerende algoritmen zijn belangrijk in de huidige maatschappij en steeds vaker blijkt dat deze algoritmen discriminerende keuzes kunnen maken. Dit kan worden verklaard doordat het algoritme zich baseert op data, waarin bepaalde menselijke vooroordelen zijn verborgen. Hierdoor wordt het profilerende algoritme ook bevooroordeeld en zal het algoritme deze vooroordelen of stereotypen reproduceren. Hierbij rijst de vraag waarom er nog geen technologische oplossingen voor dit probleem zijn. Een van de verklaringen hiervoor is dat er veel onderliggende problematiek speelt om een eerlijk algoritme te definiëren. Het doel van dit hoofdstuk was om deze problematiek te analyseren om zo inzicht te verschaffen waarom er nog geen echt eerlijk algoritme bestaat. Hiertoe is gekeken naar de problemen die ontstaan bij het definiëren van het begrip eerlijkheid, aangezien dat de eerste stap is om een eerlijk algoritme te ontwerpen.

Er zijn drie verschillende definities voor eerlijkheid, waarbij elke definitie zijn eigen problemen kent. De notie van onafhankelijkheid houdt geen rekening met verschillen in de doelvariabele Y die tussen groepen ontstaan, waardoor er juist discriminatie kan optreden. De tweede definitie, scheiding houdt geen rekening met verschillen in de standaardafwijking tussen

groepen, waardoor de sensitiviteit tussen twee groepen anders kan zijn. Hierdoor kan een algoritme wel eerlijk zijn, maar niet aan scheiding voldoen. Ten derde is toereikendheid niet genoeg om uit te sluiten dat het algoritme discriminerend is.

Naast de problemen bij de verschillende definities van eerlijkheid, zijn er ook drie overkoepelende problemen. Het eerste probleem is dat er een compromis tussen eerlijkheid en accuratesse bestaat. Verschillende studies hebben dit proberen om te lossen, maar het compromis blijft bestaan. Het tweede probleem is dat het in de praktijk onmogelijk is om aan alle drie de criteria van eerlijkheid tegelijkertijd te voldoen: ze sluiten elkaar uit. Dit wordt geïllustreerd aan de hand van het COMPAS-algoritme, die enerzijds wel eerlijk is, maar anderzijds niet. Om deze reden zal er voor een van de drie definities moeten worden gekozen. Het laatste punt van kritiek is dat de huidige definities niet intersectioneel zijn. Het houdt niet rekening met de mogelijkheid dat subgroepen extra hard kunnen worden geraakt. Daarnaast moet een intersectioneel eerlijk algoritme aannemen dat systematische verschillen tussen groepen ontstaan zijn door de maatschappij en deze proberen recht te zetten. Er zijn al verschillende studies gedaan naar het implementeren van intersectionaliteit in een algoritme, maar deze blijken in de praktijk nog niet voldoende te zijn.

Er zijn dus nog veel problemen rondom het definiëren en ontwerpen van een eerlijk algoritme. Het is nodig dat deze problemen worden overbrugd, zodat er een goede werkbare definitie voor eerlijkheid ontstaat. Wanneer er een werkbare definitie is, kan dit worden geïmplementeerd in een algoritme en kan discriminatie jegens gemarginaliseerde groepen worden voorkomen. Hiervoor moet nog veel onderzoek worden gedaan, waardoor er in de toekomst een echt eerlijk algoritme kan worden ontworpen.

Hoofdstuk 3: Profilerende algoritmes en discriminatie

Een handelingsperspectief vanuit morele bezwaren en verantwoordelijkheid

Emma Sofie Achterberg, Filosofie

In 2015 kwam Amazon erachter dat hun recruitmentsysteem waarin een algoritme gebruikt werd, geleerd had om geen vrouwen te selecteren. Dit kwam doordat het algoritme zich baseerde op data waarin weinig vrouwen functies in het bedrijf vervulde (Dastin, 2018). Dit voorbeeld laat zien dat het gebruik van profilerende algoritmes grote morele consequenties kan hebben. Uit een enquête door CareerBuilder (2017) blijkt dat meer dan de helft van de HR-managers uit de VS verwacht dat kunstmatige intelligentie in de komende vijf jaar een groot onderdeel van Human Resource Management wordt. Het is daarom belangrijk om te onderzoeken hoe we daar mee om moeten gaan. De onderzoeksvraag van dit hoofdstuk is dan ook: Welk handelingsperspectief kan uit de ethische problematiek rondom profilerende algoritmes ten aanzien van discriminatie worden opgemaakt?

In dit hoofdstuk zal allereerst onderzocht worden hoe discriminatie door profilerende algoritmes veroorzaakt kan worden om zo de problematiek te identificeren op basis waarvan oplossingen geformuleerd kunnen worden. Vervolgens wordt bekeken bij wie of wat de verantwoordelijkheid hiervoor ligt. Aan de hand daarvan worden een aantal concrete oplossingen geformuleerd. Met dit hoofdstuk wordt zo een handelingsperspectief geboden om ethischer te kunnen werken met profilerende algoritmes.

1. Hoe profilerende algoritmes kunnen leiden tot discriminatie

De in data-ethiek gespecialiseerde filosofen Mittelstadt et al. (2016) hebben in het artikel ‘’The Ethics of Algorithms: Mapping the Debate’’ een leidraad geschreven voor verder onderzoek naar de ethiek van algoritmes. Hierin hebben zij zes ethische probleempunten geïdentificeerd die optreden bij het gebruik van algoritmes. Aan de hand van de eerste vijf problemen en mijn

kritische reflectie daarop zal in deze paragraaf onderzocht worden hoe profilerende algoritmes tot discriminatie kunnen leiden. Eventuele oplossingen die hiertegen geboden zijn of hieruit af te leiden zijn zullen ook worden besproken.
Met regelmaat zullen de termen ‘’ontwikkelaars’’ en ‘’gebruikers’’ genoemd worden. Met ontwikkelaars worden de ontwerpers en programmeurs van het algoritme bedoeld. De gebruikers zijn de personen binnen een bedrijf of organisatie die ontwikkelaars inhuren en het algoritme inzetten.

1.1 Onvoldoende bewijs

Mittelstadt et al. (2016: 5) voeren als eerste probleem aan dat profilerende algoritmes tot incorrecte conclusies kunnen komen doordat ze onvoldoende bewijs hebben voor de conclusies die ze trekken. Dit komt doordat ze tot conclusies komen op grond van correlaties die ze gevonden hebben in data, maar hierbij de causaliteit voor de correlaties niet mee kunnen nemen (Mittelstadt et al., 2016: 5). Als we even teruggaan naar het voorbeeld van het recruitmentsysteem van Amazon, zien we dat het algoritme geleerd heeft om vrouwen niet voor banen te selecteren omdat zij in mindere mate in de data aanwezig zijn als werknemers dan mannen. Een algoritme is niet in staat om te evalueren op de causaliteit van deze data en de voorspelling daarop aan te passen. Het zoekt immers alleen naar de correlaties in de data. Een geschiedenis van genderongelijkheid is echter aan deze data voorafgegaan en *dit* is de causaliteit voor de correlatie, niet een incompetentie van vrouwen om te werken. Zo kan het algoritme onterecht tot de conclusie komen dat vrouwen niet geselecteerd moeten worden voor een baan. Dit is feitelijk incorrect omdat de verkeerde conclusie getrokken wordt door het ontbreken van causaliteit.

We zien echter ook dat deze conclusie moreel incorrect is. Dit is niet zozeer een gevolg van het ontbreken van causaliteit (hoewel dat er in het voorbeeld wel aan bijdraagt), maar van het verbreken van Hume’s Law. Volgens Hume’s Law is het niet mogelijk om uit hoe iets is, ook af te leiden hoe het zou moeten zijn (Hume, 1739: T3.1.1: 520). In feite is dit wel wat een algoritme

doet: op basis van data uit het verleden vindt het correlaties die vervolgens toegepast worden op een toekomstige situatie. In het geval van dit recruitmentsysteem zien we dat als we informatie uit het verleden toepassen op de toekomst, we een moreel verwerpelijke ongelijkheid in stand kunnen houden. Een algoritme komt dan ook niet tot een actief moreel besluit zoals de conclusie dat een vrouwenquotum nodig is om gelijke kansen in de toekomst te realiseren. Als uitkomsten van algoritmes zonder ethische controle toegepast worden is er dus ook geen ruimte voor verbetering van bijvoorbeeld ongelijkheden uit de samenleving. Een profilerend algoritme levert dus ook vanuit moreel perspectief onvoldoende bewijs voor een uitkomst.

1.2 Misleidend bewijs

Het gebruik van algoritmes met weinig tot geen controle wordt meestal verdedigd op grond van het idee dat een algoritme geen bias[8] mee zou nemen in het besluitvormingsproces, in tegenstelling tot mensen (Bozdag, 2013: 210). Als er een algoritme ingezet wordt, wordt er daarom al snel op vertrouwd dat de uitkomst en toepassing objectief is. Deze gedachtegang is echter onjuist.

Zoals we in paragraaf 1.1 hebben gezien kunnen data waarop algoritmes zich baseren al *biased* zijn omdat het ongelijkheden uit de samenleving reflecteert.

Daar komt bij dat volgens Johnson (2006) algoritmes altijd de waarden van de ontwerper reflecteren: ‘’Development [of an algorithm] is not a neutral, linear path; there is no objectively correct choice at any given stage of development, but many possible choices.’’ En elke keer dat dit soort keuzes gemaakt worden door de ontwikkelaar, wordt het denkbeeld van de ontwikkelaar deel van het algoritme gemaakt (Johnson, 2006, in: Mittelstadt et al. 2016: 7). Ook de functionering van het algoritme zelf is dus niet vrij van bias.

Tenslotte wordt een cruciale stap bij het gebruik van algoritmes vaak vergeten. Er is ook nog een gebruiker die besluit wat er met de uitkomst van het algoritme gedaan wordt. Hildebrandt (2011: 376) wijst erop dat het besluit of de uitkomst van een algoritme toegepast

[8] Met bias worden vooroordelen bedoeld gebaseerd op de waarden van een individu of groep.

wordt, de motivaties en waarden van de gebruiker reflecteren en niet een 'objectieve' beslissing vanuit het algoritme is.
Zo vormt het gebruik van een algoritme een misleidend bewijs voor een objectieve besluitvorming omdat het niet vrij van bias is zoals veelal gedacht wordt.

1.3 Ondoorzichtig bewijs

Het volgende probleem dat wordt aangekaart is de ondoorzichtigheid van de werking van algoritmes. Dit valt te wijten aan de complexiteit van algoritmes. Deze ondoorzichtigheid vormt een probleem voor zowel de gebruikers en ontwikkelaars, als de datasubjecten (Mittelstadt, et al., 2016: 6).

De datasubjecten zijn de personen van wie de data gebruikt wordt door een algoritme. Wanneer het besluitvormingsproces niet transparant is, wordt hiermee volgens Rubel & Jones (2014: 155) het agentschap[9] van datasubjecten ondermijnd. Datasubjecten kunnen namelijk geen betekenisvolle toestemming geven als ze de informatie niet voorhanden hebben die nodig is om een goede inschatting van risico's zoals discriminatie te maken (Schermer, 2011: 49). Het wordt zo erg moeilijk voor datasubjecten om controle uit te oefenen op hoe hun data verwerkt wordt, waardoor er weinig controle is op discriminerende gevolgen. Maar zelfs als de werking van het algoritme inzichtelijk gemaakt kan worden, is het de vraag of dit een verschil maakt. Datasubjecten zijn natuurlijk niet gespecialiseerd in data en algoritmes. Het lijkt dan ook onrealistisch om er vanuit te gaan dat zij genoeg tijd hebben om zich hierin te verdiepen en het aan te vechten indien nodig. Toch is dit een stap vooruit, omdat op deze manier in ieder geval aan het recht van datasubjecten voldaan wordt om een geïnformeerde beslissing te nemen. Het wordt ook makkelijker om bezwaar aan te tekenen tegen het gebruik van een profilerend algoritme als men zich überhaupt bewust is van de risico's.

[9] Agentschap impliceert de mogelijkheid om zelfstandig en bewust te handelen.

Ook de ontwikkelaars en gebruikers zelf hebben maar beperkt inzicht in het besluitvormingssysteem van het algoritme. Dit is vooral het geval bij *machine-learning* algoritmes die zelf nieuwe besluitvormingsstructuren ontwikkelen (Mittelstadt et al., 2016: 6). Ook handgeschreven algoritmes kunnen echter onoverzichtelijk worden doordat grote teams er over lange tijdsperioden aan werken, waardoor niemand een volledig overzicht ervan heeft (Sandvig et al., 2014: 10). Daarbij is het merendeel van computerprogrammeurs wit en man (The Census Bureau, 2017). Wanneer er onder de ontwikkelaars en gebruikers van de algoritmes weinig diversiteit is en dit niet gecompenseerd wordt door controle door een diverse groep datasubjecten, kan bias ten aanzien van bepaalde groepen gemakkelijker een weg vinden in de besluitvorming van algoritmes. Het wordt dan minder snel opgemerkt of voor gecompenseerd. Ontwikkelaars zijn tenslotte ook geen ethici, dus zullen ze mogelijk niet doorzien dat een stereotypering die uit data voortkomt, op onrechtvaardige wijze tot stand gekomen kan zijn en morele consequenties met zich mee brengt.

Zarsky (2013: 1538) suggereert dan ook dat het nodig is dat getrainde derde partijen controle op algoritmes uitvoeren, omdat dit niet van datasubjecten zelf verwacht kan worden. Het probleem blijft overeind staan dat deze derde partijen, ook al zijn ze getraind op algoritmes, nog steeds beperkt inzicht kunnen krijgen. In het beste geval zullen ze hetzelfde inzicht krijgen als de ontwikkelaars. Echter, als deze derde partij sterk divers is zullen ze in dat opzicht meer inzicht krijgen in de bias die onderhevig kan zijn aan het algoritme dan de ontwikkelaars alleen.
Een extra voordeel van zo'n derde partij is dat het een oplossing kan bieden tegen eventueel zakelijk wangedrag. Denk hierbij aan een bedrijf dat een discriminerend algoritme gebruikt voor advertenties en hier baat bij heeft in verkoopcijfers. Onafhankelijk controle kan ervoor zorgen dat dit minder snel gebeurt en het gestopt wordt als het toch gebeurt.

1.4 Onrechtvaardige uitkomsten

De drie problemen van onvoldoende aantoonbaar bewijs, misleidend bewijs en ondoorzichtig bewijs, kunnen allemaal leiden tot of bijdragen aan onrechtvaardige uitkomsten. Mittelstadt et al. (2016: 8) laten zien hoe onrechtvaardige uitkomsten vervolgens leiden tot discriminatie. Profilerende algoritmes behandelen mensen namelijk niet als individuen, maar beoordelen ze allereerst op grond van correlaties met groepen waartoe ze behoren (Newell & Marabelli, 2015: 5). Dat betekent dat bijvoorbeeld een vrouw van kleur door een algoritme allereerst beoordeeld kan worden als vrouw van kleur in plaats van als het individu dat zij is. Dit kan leiden tot discriminatie, bijvoorbeeld doordat individuen die onder bepaalde groepen vallen niet geselecteerd worden voor een baan, of bepaalde informatie niet tot zich krijgen door algoritmes in het internet. Dit kan vervolgens invloed hebben op keuzevrijheid omdat bijvoorbeeld bepaalde advertenties voor producten of zelfs opleidingen en vacatures alleen bij individuen uit bepaalde groepen terechtkomen.

Er zijn manieren waarop algoritmes geprogrammeerd kunnen worden zodat ze niet expliciet onderscheid maken op het gebied van gender of etniciteit. Dit kan echter op meer indirecte wijze nog steeds plaatsvinden, door bijvoorbeeld postcodes of andere data die toch overlapt met kenmerken zoals gender en etniciteit, wat tot gevolg heeft dat discriminatie alleen maar moeilijker te detecteren wordt (Schermer, 2011: 52).

1.5 Transformatieve effecten

‘’Algorithmic activities like profiling reontologise the world by understanding and conceptualizing it in new, unexpected ways, and triggering and motivating actions based on the insights it generates.’’ (Mittelstadt, et al., 2016: 5). Algoritmes kunnen zo transformatieve effecten[10] hebben in onze samenleving, met grote ethische consequenties van dien.[11]

[10] M.a.w. veranderende effecten. Het kan ideeën en structuren in onze samenleving veranderen.
[11] Mittelstadt et al. (2016: 9-12) identificeren een aantal andere problemen die ontstaan uit transformatieve effecten die niet van toepassing zijn op discriminatie. Daar zal ik dus niet op ingaan in dit hoofdstuk. Vanaf hier zal ik beargumenteren hoe transformatieve effecten de eerdere problemen die leiden tot discriminatie versterken.

Discriminatie kan plaatsvinden doordat individuen door algoritmes beoordeeld worden op basis van de groep waar zij onder vallen, wat tot gevolg heeft dat keuzevrijheid en toegang tot bepaalde informatie beperkt worden. Dit is allereerst natuurlijk een directe vorm van discriminatie (zie paragraaf 1.4), maar kan ook een transformatief effect hebben. Op het moment dat individuen door algoritmes gestimuleerd worden om keuzes te maken of informatie te vergaren die conform aan de groep is waarin het algoritme ze heeft ingedeeld (Mittelstadt et al., 2016: 9), worden stereotypes bevestigd omdat individuen uit een groep steeds meer naar de stereotypes die over hen bestaan handelen. Er treedt dan een *self-fulfilling prophecy* op (Barocas, 2014 in Mittelstadt et al., 2016: 9). Wanneer deze nieuwe data waarin stereotype eigenschappen versterkt aanwezig zijn opnieuw worden verwerkt door een profilerend algoritme, worden de correlaties tussen bepaalde groepen en eigenschappen nog sterker teruggevonden waardoor de uitkomsten van een algoritme nog meer stereotyperend zijn.
Daarnaast kunnen profilerende algoritmes in zoekmachines ook stereotypes in de samenleving versterken door menselijke bias te vormen of bevestigen: De zoekterm ‘’*black women*’’ leverde op Google tot voor kort nog racistische resultaten op (Noble, 2018: 98). Dit bevestigt of vormt stereotyperingen voor mensen die gebruikmaken van een zoekmachine.

In paragraaf 1.1 stelde ik dat algoritmes structurele ongelijkheden in stand kunnen houden en geen ruimte voor verbetering overlaten. Wanneer we dat probleem combineren met het probleem van transformatieve effecten, zien we in dat het niet alleen structurele ongelijkheden in stand kan houden, maar zelfs kan versterken.

2. Traceerbaarheid en morele verantwoordelijkheid

Het zesde en overkoepelende probleem dat Mittelstadt et al. (2016: 10) identificeren is het probleem van traceerbaarheid en morele verantwoordelijkheid: Als het moeilijk te identificeren is waardoor bepaalde uitkomsten veroorzaakt worden (zie paragraaf 1), is het ook lastig te bepalen wie of wat verantwoordelijk is om er iets aan te doen. Dit vormt een probleem voor het

formuleren van een handelingsperspectief. In deze paragraaf zal ik daarom trachten dit probleem op te lossen om zo een basis voor een handelingsperspectief te leggen.

2.1 Wie of wat is verantwoordelijk?

Volgens Matthias (2004: 175) kunnen we alleen de verantwoordelijkheid voor een actie toewijzen aan een actor als deze er een zekere mate van controle over heeft. Mittelstadt et al. (2016: 10) voegen hier nog een eis van intentionaliteit aan toe.

Ontwikkelaars of algoritmes

We hebben in paragraaf 1 gezien dat zelfs de ontwikkelaars beperkt inzicht hebben in de onrechtvaardige werkingen en discriminerende effecten van een profilerend algoritme. Hierdoor wordt hun controle over de uitkomsten van het algoritme verminderd. Daarbij denken *machine-learning* algoritmes tot op zekere hoogte zelfstandig en hebben ontwikkelaars hier dus geen volledige controle meer over de besluitvorming (Matthias, 2004: 182). Verder zijn discriminerende gevolgen in principe niet de bedoeling van ontwikkelaars, dus lijkt er geen sprake te zijn van intentionaliteit. Om die redenen stellen filosofen zoals Matthias (2004: 177) dat we de verantwoordelijkheid niet volledig bij de ontwikkelaars kunnen leggen en we op zoek moeten naar een nieuwe verdeling van verantwoordelijkheid (Matthias, 2004: 183).

Volgens een aantal filosofen dragen de algoritmes zelf een zekere vorm van verantwoordelijkheid vanwege het zelfstandige denken (Mittelstadt et al., 2016: 11). Sullins (2006: 159) stelt dat hier geen sprake van is omdat het agentschap van een machine anders is dan moreel agentschap. Voor moreel agentschap is ook intentionaliteit en een gevoel van verantwoordelijkheid tegenover anderen nodig, en dat is (nog) niet het geval bij algoritmes (Sullins, 2006: 169)[12]. Als het algoritme machine-agentschap heeft geeft dit echter nog steeds ruimte aan ontwikkelaars om de schuld naar algoritmes te schuiven, omdat ze geen controle en

[12] Sinds 2006 zijn er wel meer theorieën en strategieën ontwikkeld om normatieve ethiek in algoritmes op te nemen, maar tot nu toe is dit nog niet zodanig ontwikkeld dat een algoritme volledig moreel agentschap heeft volgens deze definitie van Sullins. In paragraaf 2.2 worden ''ethische algoritmes'' verder besproken.

intentie hebben over die autonome besluitvorming (Mittelstadt, et al., 2016: 11). Dit is gevaarlijk omdat dit ertoe kan leiden dat ontwikkelaars onvoldoende handelen om discriminatie te voorkomen. Volgens Mittelstadt et al. (2016: 11) is het echter ook onwenselijk om de verantwoordelijkheid volledig bij de ontwikkelaars te leggen omdat dat onterecht is. Hier ben ik het niet volledig mee eens.

Zoals ik beargumenteerd heb aan de hand van Hume's Law kunnen de uitkomsten van algoritmes niet prescriptief gebruikt worden, omdat uit een feitelijke situatie uit het verleden niet afgeleid kan worden hoe iets moreel gezien zou moeten zijn. Een algoritme kan daarom bijvoorbeeld niet de beslissing nemen om een vrouwenquotum in te stellen bij het selecteren van sollicitanten. Het is hier daarom allereerst de verantwoordelijkheid van de ontwikkelaars en gebruikers om überhaupt een morele afweging te maken over de uitkomst.

Omdat ontwikkelaars en gebruikers zo weinig inzicht en controle hebben over het algoritme, kunnen ze door gebrek aan bepaalde informatie een verkeerde afweging maken. Voor zover ze hier geen controle over hebben door ondoorzichtigheid die ze niet hadden kunnen voorkomen, zijn ze voor die fouten *zelf* niet verantwoordelijk. Ze blijven dan echter verantwoordelijk voor het feit dat ze een algoritme ontwikkelen en inzetten waar ze geen volledige controle over hebben en waarvan ze weten dat het tot discriminatie kan leiden. Over het ontwikkelen en inzetten zelf hebben ze namelijk wel controle. Hier is wel degelijk sprake van intentie omdat ze op de hoogte zijn van de risico's. Dit is conform de definitie van verantwoordelijk die Matthias (2004: 175) zelf geeft.

Ontwikkelaars of gebruikers

Een belangrijkere vraag over de verdeling van verantwoordelijkheid gaat nu om de verdeling tussen ontwikkelaars en gebruikers. Zoals gezegd zijn de ontwikkelaars de ontwerpers van het algoritme en zetten de gebruikers het algoritme in. Logischerwijs lijkt hieruit te volgen dat de gebruikers verantwoordelijk zijn voor het besluit of ze de uitkomst van het algoritme wel of niet inzetten, en daarom voor de gevolgen van discriminatie. Een probleem bij de verantwoordelijkheid van de gebruiker is echter dat ze zich misschien niet bewust zijn van het feit dat algoritmes tot discriminatie kunnen leiden. Ze zijn immers geen programmeurs of ethici.

Ook hebben ze weinig inzicht in hoe de ontwikkelaars te werk gegaan zijn en hoe betrouwbaar de uitkomst van het algoritme daarom is. Hieruit wordt duidelijk dat de verantwoordelijkheid dan ook niet zomaar van de ontwikkelaar naar de gebruiker verplaatst.

De ontwikkelaar is zich wel bewust van de risico's op discriminatie, of zou zich daar in ieder geval bewust van moeten zijn om op gerechtvaardigde manier als programmeur te werken. Daarnaast heeft de ontwikkelaar de mogelijkheid (en dus controle) om de gebruiker van deze informatie te voorzien, dus is hij/zij daar ook verantwoordelijk voor. Daarnaast moeten ontwikkelaars natuurlijk in het ontwikkelingsproces gebruikmaken van alle manieren die geïdentificeerd zijn om discriminatie te voorkomen die onder hun controle vallen. Als de ontwikkelaar deze methodes toegepast heeft en de gebruiker ingelicht heeft, is de verantwoordelijkheid echter nog steeds niet volledig naar de gebruiker verplaatst. Ontwikkelaars kunnen discriminatie namelijk ook voorkomen door het algoritme überhaupt niet aan de gebruiker aan te leveren. Daarnaast blijft de vraag overeind staan wanneer iemand alles gedaan heeft dat mogelijk is. Zelfs als alle bekende manieren om discriminatie te voorkomen ingezet zijn, had er misschien nog een nieuwe manier gevonden kunnen worden. Het feit dat er niet één ontwikkelaar aan een algoritme werkt maar een heel team over een langere tijdsperiode draagt bij aan dit probleem: wanneer heeft de ene ontwikkelaar genoeg gedaan en mag hij/zij erop vertrouwen dat de andere ontwikkelaar ook zijn/haar aandeel levert? Is het dan de verantwoordelijkheid van die ene ontwikkelaar om de ander te controleren? Dit ligt wel binnen zijn/haar controle, dus volgens de definitie die Matthias (2004) geeft, is de ontwikkelaar daar inderdaad verantwoordelijk voor en dus ook voor de gevolgen als hij/zij dit niet doet.

Dit laat zien dat hoewel we de verantwoordelijkheid van ontwikkelaars wel kunnen minimaliseren door te stellen dat de verantwoordelijkheid verkleint als ze alle beschikbare methodes om discriminatie te voorkomen toepassen en de gebruiker informeren over de onzekerheden, de verantwoordelijkheid nooit helemaal verdwijnt. Het komt er dan op neer dat dit een veeleisende formulering van verantwoordelijkheid is. Veeleisendheid lijkt me echter geen reden om deze verantwoordelijkheid te ontkennen, het lijkt eerder een onontkoombaar gevolg te zijn van technologische ontwikkelingen waarin besluitvorming wordt overgedragen op een machine.

Matthias (2004: 183) levert een bezwaar van veeleisendheid door te stellen dat controle op de uitkomst van een algoritme in de praktijk niet altijd mogelijk is, omdat algoritmes juist vaak gebruikt worden om snellere beslissingen te kunnen nemen. Dit lijkt mij geen argument om de verantwoordelijkheid weg te kunnen nemen, maar vooral een probleem dat vraagt om oplossingen. Het incorporeren van een normatieve ethiek in het algoritme is zo'n mogelijke oplossing.

2.2 Een ethische algoritme

Een aantal filosofen stellen voor om normatieve ethiek in een algoritme te incorporeren en onderzoeken hoe dit eruit zou moeten zien. Hier zijn verschillende mogelijkheden voor.

Ethisch protocol van het bedrijf

Turilli & Floridi (2009: 111) stellen voor dat er binnen een organisatie dat een algoritme in werking wil stellen, een ethisch protocol vastgesteld wordt. Dit protocol wordt bepaald aan de hand van alle ethische regels waaraan alle actoren binnen een organisatie onderhevig zijn (denk aan zowel wettelijke regels als de vastgestelde ethiek van een bedrijf). Dit protocol wordt dan in het algoritme geprogrammeerd als regels. Mittelstadt et al. (2016: 11) werpen hier tegen op dat het moeilijk kan zijn om menselijke ethische principes in de taal van algoritmes om te zetten. Een groter bezwaar hiertegen lijkt mij dat de ethiek van een organisatie of bedrijf niet noodzakelijk juist hoeft te zijn en dit ruimte overlaat voor zakelijk wangedrag. Ook kunnen er fouten gemaakt worden in het achterhalen van de ethiek van het bedrijf, als een bedrijf geen ethici heeft die hierbij kunnen helpen. Dit lijkt echter een te hoge eis te worden om een algoritme te mogen gebruiken. Bedrijven mogen immers ook mensen beslissingen laten nemen terwijl zij mogelijk morele fouten maken, of een onjuist ethisch protocol hanteren zolang het niet direct in strijd is met de wet. Het verschil is echter dat als er ethisch iets fout gaat in het ontwerp van een algoritme, deze fout elke keer dat het algoritme gebruikt wordt zal terugkeren, terwijl mensen

elke keer op andere manier tot een besluit kunnen komen waardoor een fout niet structureel hoeft te worden.

Heroïsche deugden

Wiltshire (2015, in Mittelstadt et al., 2016: 11) suggereert een basis van heroïsche deugden[13]. Wiltshire (2015: 61) definieert heroïsme voornamelijk als altruïstisch of onbaatzuchtig handelen, en noemt een aantal voorbeelden van fictie waarbij robots heroïsch handelen, om te laten zien dat het niet conceptueel ondenkbaar is om deze vorm van ethiek in algoritmes te incorporeren. Hierbij vergeet hij echter dat een algoritme in tegenstelling tot een meer fysieke robot uit een actiefilm, geen eigen belangen kan hebben, omdat het in besluitvormingsprocessen geen risico's voor zichzelf loopt. Het algoritme zou dus hooguit heroïsch kunnen handelen ten opzichte van belangen van het bedrijf of de organisatie, want dit is de enige manier waarop het ''eigen'' belangen kan behartigen en daar onzelfzuchtig over kan zijn. Hiervoor moet het dus ook eerst mogelijk zijn om de belangen van het bedrijf of de organisatie aan het algoritme duidelijk te maken. Als dit al mogelijk is, lopen we tegen het volgende probleem aan. Heroïsche acties vallen meestal onder *''supererogatory''* handelingen: handelingen die een morele plus zijn, geen morele verplichting. Hoewel het zeer positief te noemen is, lijkt dit veel gevraagd van organisaties en bedrijven.

Intuïtief ethisch redeneren

Bello & Bringsjord (2012 in Mittelstadt et al., 2016: 12) stellen dat als we moraliteit willen incorporeren in algoritmes we dit niet moeten baseren op de klassieke ethische theorieën omdat die niet representatief zijn voor hoe mensen in werkelijkheid morele besluiten nemen. Ze pleiten ervoor dat algoritmes moeten leren hoe ze intuïtief tot morele besluiten kunnen komen, zoals mensen dit doen. Hiervoor is dan eerst verder onderzoek nodig naar hoe mensen tot deze besluitvorming komen. Dit leidt tot het probleem dat mensen niet altijd ethisch handelen. Het is dus maar de vraag is of het goed is om volledig te steunen op menselijke intuïties voor het maken van ethische overwegingen. Daarbij is het onduidelijk of het menselijke ethische denken volledig

[13] Als een manier om deugdethiek in algoritmes te incorporeren.

intuïtief is, of dat die intuïties voor een groot deel cultureel bepaald zijn. Ook dan zal de bias van ontwikkelaars deel kunnen worden van de morele beslissingen van het algoritme. Zoals ik eerder noemde stel ik met deze kritiek hogere eisen aan het gebruik van algoritmes dan aan het inzetten van werknemers, maar is dit nodig omdat een algoritme op meer structurele manier verkeerd kan blijven handelen. Een mens kan bijvoorbeeld inconsistent handelen of betere morele overtuigingen ontwikkelen. Het is nog maar de vraag of een ethisch algoritme die ruimte voor verbetering heeft, aangezien het onderhevig is aan geprogrammeerde regels.

Hoewel de besproken manieren van het incorporeren van ethiek in algoritmes niet perfect zijn, kunnen ze sterk bijdragen aan het voorkomen van discriminatie en is het uiterst urgent om dit zo snel mogelijk toepasbaar te maken. Toch heb ik geprobeerd aan te tonen dat ook met deze technieken voor ethische besluitvorming, de verantwoordelijkheid voor de uitkomsten van algoritmes niet weggenomen wordt bij de ontwikkelaars en gebruikers. Zowel de ethiek die gebruikt wordt als de manier waarop dit gebeurt blijven een keuze en daarom de verantwoordelijkheid van de ontwikkelaar, en van de gebruikers als de ethiek van het bedrijf ervoor gebruikt wordt. Ook zijn deze technieken niet waterdicht en doorzichtig waardoor er nog steeds een gebrek aan controle over de algoritmes is waar ontwikkelaars en gebruikers verantwoordelijk voor blijven.

3. Oplossingen

In paragraaf 2 heb ik beargumenteerd wie op welke manieren verantwoordelijkheid dragen voor toegepaste profilerende algoritmes. Dit biedt een basis voor een handelingsperspectief. Als we deze verdeling van verantwoordelijkheid toepassen op de problematiek en oplossingen die besproken zijn aan paragraaf 1, vinden we een (beginnend) handelingsperspectief vanuit de discipline van morele filosofie.

3.1 Diversiteit

Allereerst is diversiteit nodig onder programmeurs en gebruikers. De verantwoordelijkheid die ontwikkelaars en gebruikers hiervoor kunnen nemen, is bijvoorbeeld door binnen het bedrijf een diversiteitsquotum[14] in te stellen en advertenties voor vacatures inclusief te maken. Opleidingen tot programmeur dragen de verantwoordelijkheid om een diverse groep studenten aan te trekken.

3.2 Onafhankelijke controle

Er zou gebruik gemaakt moeten worden van een derde onafhankelijke partij die bestaat uit een divers team van mensen die gespecialiseerd zijn in de werking van algoritmes. Zij kunnen dan met steekproeven het gebruik van algoritmes controleren. Het volgende probleem is nu wie er verantwoordelijk is om zo'n team op te richten. We kunnen ons afvragen of kleine bedrijven en organisaties hier überhaupt toe in staat zijn, omdat een gespecialiseerd team samenstellen veel tijd en geld kan kosten. Grote bedrijven zijn hier in dat opzicht wel toe in staat, maar de vraag is dan hoe onafhankelijk een team nog is als ze door het bedrijf betaald worden om controle uit te oefenen. Het komt er dus op neer dat het niet binnen de controle van bedrijven en organisaties zelf ligt om een *onafhankelijk* team samen te stellen, dus kunnen we ze daar niet voor verantwoordelijk maken. Dit neemt niet weg dat ze nog steeds verantwoordelijk zijn om deze controle uit te laten voeren voordat ze een algoritme inzetten, of het nou mogelijk is of niet. Dit vormt zo een praktisch probleem die vraagt om een oplossing. Mijn voorstel is dat daarmee de verantwoordelijkheid voor het oprichten van het team verplaatst wordt naar overheden. Overheden dragen immers een verantwoordelijkheid om grondrechten te waarborgen en discriminatie is daar een schending van. Controle op algoritmes zou dan ook wettelijk verplicht moeten zijn. Daarnaast draagt een overheid ook verantwoordelijkheid over het stimuleren van een economie en het bedrijfsleven, dus moeten ze het gebruik van algoritmes met een verplichte onafhankelijke controle daarop ook werkbaar maken. Vanwege de omvang van het probleem en omdat overheden in staat zijn om een onafhankelijk team te ontwikkelen of te laten ontwikkelen

[14] Hierbij moet gekeken worden naar diversiteit in gender, seksualiteit, leeftijd en etniciteit.

door een non-profit organisatie, valt dit binnen hun controle en daarmee verantwoordelijkheid. De kosten van een controle blijven voor de rekening van het bedrijf of de organisatie, omdat zij het profilerende algoritme voor belangen van het bedrijf willen gebruiken. De overheid is dus alleen verantwoordelijk voor de verplichting van onafhankelijke controle en de mogelijkheid ertoe.

3.3 Informatie voor datasubjecten

Datasubjecten moeten beter geïnformeerd worden over waar hun data voor gebruikt wordt. Vaak is de informatie wel aanwezig maar is dit complex geformuleerd in een lange tekst, die mensen niet lezen als ze bijvoorbeeld iets opzoeken op het internet. Hier zou een preview gemaakt van moeten worden die in enkele zinnen stelt wat profilering door data inhoudt, die zichtbaar is voordat mensen op ‘’akkoord’’ klikken. Hier valt tegenin te brengen dat ook dan datasubjecten deze informatie vaak niet lezen. Dit is echter een gemakkelijker manier om datasubjecten beter te informeren met een grotere kans dat ze het lezen en begrijpen, dus is het alsnog de verantwoordelijkheid van ontwikkelaars en gebruikers om dit te doen.

3.4 Informatie voor gebruikers

Niet alleen de datasubjecten moeten geïnformeerd worden over de mogelijke gevolgen. Ook de gebruikers zijn zich hier vaak niet van bewust, terwijl ontwikkelaars dit wel zijn. Het is dan ook de verantwoordelijkheid van de ontwikkelaars om dit in samenwerkingscontracten op te nemen en duidelijk te informeren. Indien het nodig is om daar een werknemer op te zetten die gespecialiseerd is in communicatie omdat ze het zelf niet kunnen overbrengen (wat daarmee buiten hun controle zou vallen), moeten ze dat doen (want dit valt wel binnen hun controle). Ook moet er hierbij op gewezen worden dat er altijd nog ethische reflectie nodig is over de uitkomst van een (niet-ethisch) algoritme, vanwege Hume’s Law. Daar zijn gebruikers zelf verantwoordelijk voor. Of deze controle in de praktijk werkbaar is verandert niets aan de verantwoordelijkheid. Daarom is het ook voor de gebruikers van belang dat hier verder

onderzoek naar gedaan wordt en ook op andere manieren discriminatie door algoritmes voorkomen wordt, zodat de verantwoordelijkheid hiervoor krimpt. Dit zou hen kunnen motiveren om bijvoorbeeld meer te investeren in het team van ontwikkelaars zodat deze meer geld en tijd krijgen om het algoritme zo eerlijk mogelijk te laten werken.

3.5 Ethische algoritmes

Als het mogelijk wordt om normatieve regels op te nemen in de besluitvormingsstructuur van een algoritme, zal deze eerder tot rechtvaardige uitkomsten komen. Ethische reflectie op de uitkomsten wordt daardoor minder dringend. Een algoritme baseert zich dan namelijk niet meer alleen op informatie uit het verleden[15], maar ook op morele regels. Een algoritme werkt dan niet meer geheel amoreel[16]. Daarbij krijgen ontwikkelaars en gebruikers op deze manier meer controle over de werking en uitkomsten van het algoritme waardoor er meer mogelijkheden ontstaan om onrechtvaardige uitkomsten tegen te gaan. De verantwoordelijkheid verandert hier dus mee: De verantwoordelijkheid voor het gebruik van een algoritme waar ze weinig controle over hebben neemt af, maar de verantwoordelijkheid voor de inhoud van de uitkomsten neemt toe. In paragraaf 2.2 heb ik laten zien dat de voorgestelde methodes voor ‘’ethische algoritmes’’ niet foutloos zijn en veelal gekleurd worden door de waarden van de ontwikkelaars en gebruikers. De eisen voor het gebruik van algoritmes moet echter niet ineens hoger gesteld worden dan de eisen voor het inzetten van werknemers. Een algoritme werkt misschien niet ethisch foutloos en onafhankelijk, menselijke[17] werknemers doen dat ook niet. Werkgevers

[15] Zie paragraaf 1.1 over Hume’s Law

[16] Amoreel betekent dat er geen sprake is van een morele afweging.

[17] Grappig genoeg probeerde het programma Google Docs het woord ‘’menselijke’’ hier te verbeteren naar ‘’mannelijke’’ als bijvoeglijk naamwoord bij ‘’werknemers’’. Hier zien we hoe taalalgoritmes stereotyperingen kunnen versterken. We kunnen ons afvragen of dit niet simpelweg gebeurde omdat ‘’menselijke werknemers’’ een ongebruikelijke woordcombinatie is omdat we er meestal wel vanuit gaan dat werknemers menselijk zijn, en ‘’menselijk’’ door de ‘’m’’ vrij veel op ‘’mannelijke’’ lijkt. Het is dus misschien ook wel logisch dat het deze verbetering suggereert, en niet bijvoorbeeld ‘’vrouwelijke’’ of een totaal ander bijvoeglijk naamwoord. Het kan zelfs positief zijn dat het blijkbaar vaak voorkomt dat we ‘’mannelijke’’ voor ‘’werknemers’’ zetten, in plaats van er per definitie vanuit te gaan dat werknemers mannelijk zijn. Echter kan ik hier weer wijzen op de ondoorzichtigheid van algoritmes: we weten zo met het blote oog niet of het ‘’mannelijke’’ suggereerde omdat het een sterke correlatie ziet tussen mannen en werknemers, of doordat dit vaker als bijvoeglijk naamwoord voor ‘’werknemers’’ wordt gebruikt. Desalniettemin zien we hier dat suggesties door algoritmes ons bepaalde richtingen op kunnen

hebben uiteraard een vorm van verantwoordelijkheid over de handelingen en besluiten van hun werknemers, maar werknemers mogen ook met beperkte controle worden vertrouwd. Het gebruik van ‘’ethische algoritmes’’ komt op moreel vlak dus al iets dichter in de buurt van het inzetten van menselijke werknemers, en is daarom minder bezwaarlijk en beter uitvoerbaar. Helaas geldt voor algoritmes niet precies hetzelfde als bij menselijke werknemers, omdat een algoritme sneller structurele fouten blijft maken, terwijl een mens kan variëren in besluitvorming en zichzelf kan beteren. De verantwoordelijkheid voor algoritmes blijft dus groter dan voor werknemers, maar met ‘’ethische algoritmes’’ wordt dit verschil wel kleiner gemaakt en kan discriminatie door algoritmes worden verminderd.

3.6 Oplossingen buiten de morele filosofie

Vanwege de grote verantwoordelijkheid die ontwikkelaars en gebruikers dragen, is het voor hen belangrijk om deze oplossingen zoveel mogelijk toe te passen. Hier stopt het echter niet. Er zijn oplossingen die buiten het vakgebied van de ethiek vallen, of buiten de blik van dit hoofdstuk. Aangezien het binnen de mogelijkheden van de ontwikkelaars en gebruikers ligt om deze andere oplossingen uit verschillende vakgebieden toe te passen, zijn ze er ook verantwoordelijk voor om dit te doen. Het theoretische handelingsperspectief van morele verantwoordelijkheid is zo ook toepasbaar op zaken rondom profilerende algoritmes die in dit hoofdstuk niet aan bod gekomen zijn.

Conclusie

In dit hoofdstuk zijn allereerst vijf ethische problemen bij het gebruik van profilerende algoritmes besproken en aangetoond hoe deze tot discriminatie kunnen leiden. Hieruit is duidelijk geworden dat algoritmes niet tot morele uitkomsten kunnen komen omdat ze zich op data uit het verleden baseren, wat in strijd is met *Hume’s Law* en ongelijkheden uit de

duwen die onze associaties of stereotyperingen (zoals het idee dat werknemers meestal mannelijk zijn) kunnen versterken.

samenleving in stand kan houden. Daarnaast zijn algoritmes niet vrij van bias en is dat moeilijk te doorzien door de complexe werking. Door de onderkende bias die aanwezig is in de uitkomsten van een algoritme kunnen transformatieve effecten plaatsvinden waarin stereotypes gerealiseerd en versterkt worden.

Het zesde ethische probleem van verantwoordelijkheid is vervolgens opgepakt om een basis voor een handelingsperspectief te vinden. De definitie van verantwoordelijkheid die gebruikt is, stelt dat iemand ergens verantwoordelijk voor is als hij/zij er een zekere mate van controle over heeft en er sprake is van intentie achter de handeling. Er is vervolgens beargumenteerd dat de verantwoordelijkheid voor het toegepaste algoritme onlosmakelijk verbonden is met de ontwikkelaar en de gebruiker, omdat zij altijd een vorm van controle hebben over het feit *dat* ze het algoritme inzetten en *hoe* ze dit doen. Omdat ze zich bewust zijn van de risico's en dus bewuste keuzes maken, is er ook sprake van intentie. De ontwikkelaars en gebruikers zijn dan ook moreel verplicht om discriminatie door hun algoritmes zoveel mogelijk te voorkomen.

Door de verdeling van verantwoordelijkheid toe te passen op de morele bezwaren tegen algoritmes en de mogelijke oplossingen, zijn een aantal concrete oplossingen gevonden die in combinatie met de verdeling van verantwoordelijkheid een handelingsperspectief bieden: Er moet gewerkt worden aan het waarborgen van diversiteit onder ontwikkelaars; er moet sprake zijn van een onafhankelijke derde partij die steekproefsgewijs controle op de besluitvorming van het algoritme uitvoert; er is betere informatieoverdracht nodig van ontwikkelaars aan zowel datasubjecten als gebruikers; ethische reflectie voordat uitkomsten van algoritmes ingezet worden is essentieel; en de ontwikkeling van ''ethische algoritmes'' kan ervoor zorgen dat algoritmes minder snel discrimineren en beter vertrouwd kunnen worden. Los van de oplossingen die in dit filosofische literatuuronderzoek naar voren gekomen zijn, zijn er ook oplossingen mogelijk vanuit andere vakgebieden die door ontwikkelaars meegenomen moeten worden. De verantwoordelijkheid van ontwikkelaars en gebruikers ten aanzien van oplossingen buiten dit hoofdstuk kunnen telkens aan de gegeven definitie van verantwoordelijkheid getoetst worden om zo te bepalen wie de plicht heeft actie te ondernemen.

Hoofdstuk 4: Integratie

Emma Sofie Achterberg, Margot Hoogerwerf, Eline Kempkes

In de vorige hoofdstukken is vanuit drie verschillende disciplines een antwoord gegeven op de onderzoeksvraag: "*Welk handelingsperspectief kan worden afgeleid uit de oorzaken van de discriminatie jegens zwarte vrouwen door profilerende algoritmes?"*
In dit integratieve hoofdstuk zullen eerst de drie bevindingen van deze disciplines worden besproken. Vervolgens worden de conflicten die tussen deze inzichten bestaan geïdentificeerd en weggenomen door een *common ground* van aangepaste begrippen, aannames en theorieën te creëren. Door middel van deze *common ground* wordt uiteindelijk een *more comprehensive understanding* geformuleerd: een interdisciplinair antwoord op de hoofdvraag.

4.1 Disciplinaire conclusies

Geschiedenis

In het disciplinaire hoofdstuk van Geschiedenis is beschreven hoe historisch verklaard kan worden dat juist zwarte vrouwen door profilerende algoritmes worden gediscrimineerd. Er is betoogd dat de oorsprong van deze discriminatie gevonden kan worden in de achttiende eeuw, de eeuw van slavernij en surveillance. In deze eeuw raakten 'ras' en surveillance onlosmakelijk met elkaar verbonden. Dat bleven deze concepten ook na de achttiende eeuw. Dit kwam onder meer door het creëren en het voortbestaan van stereotypes over zwarte mensen, in het bijzonder over zwarte vrouwen. Deze stereotypes – zoals de Jezebel, de Mammy en de Sapphire – werden door witte machthebbers ingezet ter rechtvaardiging van hun gewelddadige en controlerende behandeling van zwarte mensen.

Het hoofdstuk laat vervolgens zien hoe historische praktijken van *racializing surveillance* kunnen doorwerken in hedendaagse digitale technologieën. Een verklaring hiervoor is gezocht in denkbeelden en eigenschappen van de ontwikkelaars van technologieën. Dit zijn voornamelijk witte mannen die zich onbewust zijn van hun eigen stereotypes over zwarte vrouwen en van de aanwezigheid van deze zelfde stereotypes in data. In dit hoofdstuk is beargumenteerd dat deze

stereotypes over zwarte vrouwen ook één van de redenen vormen waarom er zo weinig zwarte vrouwen in de ICT werken. Vanwege het stereotype van 'technoprimitiviteit' zouden zwarte vrouwen namelijk niet geschikt worden geacht om te werken in de ICT.

Uiteindelijk is in dit disciplinaire hoofdstuk betoogd dat, om de discriminatie van zwarte vrouwen door algoritmes tegen te gaan, het zwarte vrouwen ten eerste makkelijker moet worden gemaakt om in de ICT te gaan werken en dat ten tweede witte mannelijke programmeurs bewust moeten worden gemaakt van hun stereotypes jegens zwarte vrouwen.

Kunstmatige Intelligentie

Literatuuronderzoek vanuit de Kunstmatige Intelligentie heeft de problematiek rondom het definiëren van een eerlijk profilerend algoritme onderzocht. Deze problematiek geeft een verklaring waarom er discriminatie jegens gemarginaliseerde groepen door profilerende algoritmes kan bestaan. Er is namelijk in de literatuur nog geen consensus bereikt over de definitie van een eerlijk algoritme, waardoor er nog geen algoritme bestaat dat in alle opzichten echt eerlijk is. Centraal in dit onderzoek staan de drie verschillende definities van eerlijkheid en hun voor- en nadelen. Omdat een definitie van 'eerlijk' voor alle gemarginaliseerde groepen moet gelden, richt dit hoofdstuk zich niet specifiek op zwarte vrouwen maar op gemarginaliseerde groepen in het algemeen.

In het hoofdstuk is betoogd dat geen enkele wiskundige definitie van eerlijkheid echt eerlijk is: elke definitie kent zo zijn zwakheden waardoor er alsnog discriminatie kan plaatsvinden als een algoritme dat volgens de definitie is gebouwd in de praktijk wordt ingezet. Er worden drie overkoepelende problemen besproken, die voor alle definities van eerlijkheid gelden. Het eerste probleem is dat er een compromis tussen eerlijkheid en accuratesse bestaat: wanneer het algoritme eerlijker wordt, wordt het doorgaans slechter in voorspellen. Het tweede probleem is dat de verschillende definities onverenigbaar zijn met elkaar: een algoritme kan in de praktijk niet op alle drie de manieren eerlijk zijn. Er zal moeten worden gekozen voor één van de definities. Ten slotte wordt het gebrek aan intersectionaliteit besproken: een algoritme kan niet seksistisch of racistisch *lijken*, maar toch discriminerend *zijn* jegens zwarte vrouwen. Bij het bespreken van deze verschillende problemen werden waar mogelijk oplossingen gepresenteerd,

maar er is nog veel onderzoek nodig naar de technische mogelijkheden rondom een eerlijk algoritme.

Filosofie

In dit hoofdstuk is onderzocht welk handelingsperspectief afgeleid kan worden uit de morele bezwaren tegen profilerende algoritmes en de morele verantwoordelijkheid ervoor. De definitie van verantwoordelijkheid die hierin gebruikt is, stelt dat iemand ergens verantwoordelijk voor is als hij/zij er een zekere mate van controle over heeft en er sprake is van intentie achter de handeling. Er is vervolgens beargumenteerd dat de verantwoordelijkheid voor het toegepaste algoritme onlosmakelijk verbonden is met de ontwikkelaar en de gebruiker, omdat zij altijd een vorm van controle hebben over het feit *dat* ze het algoritme inzetten en *hoe* ze dit doen. Omdat ze zich bewust zijn van de risico's en dus bewuste keuzes maken, is er ook sprake van intentie. De ontwikkelaars en gebruikers zijn dan ook moreel verplicht om discriminatie door hun algoritmes zoveel mogelijk te voorkomen.

In het hoofdstuk wordt ook duidelijk hoe die verantwoordelijkheid verdeeld wordt tussen gebruikers en ontwikkelaars. Door de verdeling van verantwoordelijkheid toe te passen op de morele bezwaren tegen algoritmes en de mogelijke oplossingen, zijn een aantal concrete oplossingen gevonden die een handelingsperspectief bieden: Er moet gewerkt worden aan het waarborgen van diversiteit onder ontwikkelaars; er moet sprake zijn van een onafhankelijke derde partij die steekproefsgewijs controle op de besluitvorming van het algoritme uitvoert; er is betere informatieoverdracht nodig van ontwikkelaars aan zowel datasubjecten als gebruikers; ethische reflectie voordat uitkomsten van algoritmes ingezet worden is essentieel; en de ontwikkeling van ''ethische algoritmes'' kan ervoor zorgen dat algoritmes minder snel discrimineren en beter vertrouwd kunnen worden. Los van de oplossingen die in dit filosofische literatuuronderzoek naar voren gekomen zijn, zijn er ook oplossingen mogelijk vanuit andere vakgebieden die door ontwikkelaars meegenomen moeten worden. De verantwoordelijkheid van ontwikkelaars en gebruikers ten aanzien van oplossingen buiten dit hoofdstuk kunnen telkens aan de gegeven definitie van verantwoordelijkheid getoetst worden om zo te bepalen wie de plicht heeft actie te ondernemen.

4.2. *Common Ground*

<u>Begrippen en concepten:</u>

1. **Classificatie/*social sorting*/profileren/uitkomsten**

In de disciplinaire hoofdstukken worden de begrippen 'classificatie', 'social sorting', 'profilering' en 'uitkomsten' door elkaar gebruikt, hetgeen leidt tot een conceptueel conflict. Met het begrip 'classificatie' doelt de Kunstmatige Intelligentie[18] op het indelen van individuen in bepaalde categorieën, op basis van gegevens over deze individuen. De Kunstmatige Intelligentie stelt dat profilerende algoritmes *classifiers* zijn: entiteiten die het proces van 'classificatie' uitvoeren. De classificatie door algoritmes vindt plaats nadat het algoritme met eerdere data is getraind.

De Geschiedenis blijkt een onderscheid te maken tussen 'classificatie' en 'profilering', omdat het gebruikmaakt van het begrip 'social sorting'. *Social sorting* is volgens de Geschiedenis een manier om identiteiten vast te stellen, maar ook om risico's en waarden toe te kennen aan mensen. *Social sorting* omvat zowel classificatie als profilering: classificatie is het bepalen van de categorieën en de bijbehorende eigenschappen, profilering is het plaatsen van individuen in deze categorieën. In de Kunstmatige Intelligentie wordt dit onderscheid niet gemaakt: profilering wordt gezien als onderdeel van classificatie, waarbij classificatie wordt gedefinieerd als het indelen van individuen in categorieën.

De Filosofie maakt wel een onderscheid tussen 'correlaties' en 'profilering', maar benoemd het niveau van 'classificatie' in haar geheel niet. 'Profilering' wordt daarnaast vaak aangeduid met de woordcombinatie 'uitkomst van het algoritme'. In het hoofdstuk wordt met deze woordcombinatie gedoeld op de toepassing van classificaties: wanneer individuen worden ingedeeld in categorieën die bepaald zijn op basis van correlaties.

[18] In dit hoofdstuk wordt regelmatig gesteld dat ''de Filosofie / Geschiedenis / Kunstmatige Intelligentie'' iets stelt. Hiermee wordt uiteraard verwezen naar de disciplinaire hoofdstukken, niet noodzakelijk naar de disciplines in het algemeen.

Om dit terminologische conflict te overbruggen maken wij gebruik van de integratietechniek *organisatie*. Deze techniek creëert een *common ground* door te verhelderen hoe bepaalde begrippen zich tot elkaar verhouden (Repko & Szostak, 2017). De disciplinaire inzichten uit de Geschiedenis bieden een aanknopingspunt om de relaties tussen de begrippen te verhelderen, door *social sorting* te definiëren als 'classificatie' plus 'profilering'. De processen van classificatie en profilering vinden altijd achter elkaar plaats: eerst vindt classificatie plaats, daarna profilering. Een schematische weergave van deze *common ground* die wij op basis van de techniek *organisatie* hebben gecreëerd ziet er als volgt uit: *social sorting* = classificatie + profilering.

Daarnaast is de methode van *intensie* (Alkema & Hamer, 2016) nog nodig om te zorgen dat ook voor Kunstmatige Intelligentie classificatie het identificeren van groepen met bepaalde eigenschappen op basis van correlaties betekent. Profilering valt hier dus nog niet onder en wordt een gescheiden concept.

Voor de filosofie wordt een nieuw concept geïntroduceerd, namelijk classificatie. Zo wordt het begrip 'classificatie' dat in de hoofdstukken vanuit Kunstmatige Intelligentie en Geschiedenis naar voren kwamen uitgebreid volgens de methode van *extensie* (Repko & Szostak, 2017) naar het domein van de filosofie.

2. Classificatie/stereotypering

In de disciplinaire stukken van de Kunstmatige Intelligentie en Geschiedenis worden twee verschillende begrippen gebruikt waarmee hetzelfde bedoeld wordt. De Kunstmatige Intelligentie gebruikt het begrip 'classificatie', waarmee wordt gedoeld op het in hokjes plaatsen van individuen in bepaalde categorieën, zoals in het vorige conflict is aangeduid. De Geschiedenis doelde met het begrip 'stereotypering' op hetzelfde concept: het in bepaalde categorieën plaatsen van individuen.

Dit conflict lossen wij op met de integratietechniek *distinguishing*, die is geïntroduceerd door Van der Lecq (2012). Met deze techniek wordt juist het verschil tussen de begrippen classificatie en stereotypering benadrukt, om verwarring te voorkomen en zo een *common ground* te creëren. Wij komen tot het onderscheid dat stereotypering bestaat uit denkbeelden van

mensen, terwijl classificatie en profilering wordt gedaan door algoritmes en door de correlaties die algoritmes maken. In het vorige conflict hebben we laten zien dat de Kunstmatige Intelligentie met 'classificatie' doelt op 'profilering'.

3. Discriminatie

In de verschillende disciplines wordt de term 'discriminatie' net iets anders gebruikt, waardoor er een conflict optreedt. Zowel het hoofdstuk vanuit Geschiedenis als dat vanuit Filosofie maakt een onderscheid tussen onrechtvaardige profilering en discriminatie. Onrechtvaardige profilering vindt plaats op het moment dat een individu beoordeeld wordt als onderdeel van een groep die op onrechtvaardige wijze gestereotypeerd wordt. Op het moment dat er naar deze profilering wordt gehandeld vindt er discriminatie plaats.

In het hoofdstuk vanuit Kunstmatige Intelligentie wordt onrechtvaardige profilering op zichzelf al als discriminatie gezien. Het is logisch dat dit in de Kunstmatige Intelligentie zo gezien wordt, omdat deze discipline zich bezighoudt met de werking van technologieën en zich dus ook vooral richt op hoe deze technologieën onrechtvaardige profilering kunnen voorkomen. Er ligt dus weinig focus op de vraag of de technologieën ook toegepast worden.

Het hoofdstuk vanuit de Filosofie laat zien dat het belangrijk is om een onderscheid te maken tussen discriminatie en onrechtvaardige profilering, omdat het besluit om een onrechtvaardig profilerend algoritme toch in te zetten een belangrijke factor is in het toewijzen van morele verantwoordelijkheid en daarmee in het ontwikkelen van handelingsperspectieven. Deze factor wordt minder zichtbaar op het moment dat onrechtvaardige profilering gelijk wordt gesteld aan discriminatie. Om deze reden hebben wij ervoor gekozen de integratietechniek *intensie* (Alkema & Hamer, 2016) toe te passen op de definitie van discriminatie van de Kunstmatige Intelligentie. Hiermee verkleinen we het concept 'discriminatie' naar de definitie die in de hoofdstukken vanuit Filosofie en Geschiedenis gebruikt wordt. Discriminatie vindt dus pas plaats als mensen op basis van een onrechtvaardige profilering op een andere (nadelige) manier behandeld worden.

4. Rechtvaardigheid

Het laatste conflict tussen concepten speelt rondom ‘eerlijkheid’ en ‘rechtvaardigheid’. In het disciplinaire hoofdstuk van de Kunstmatige Intelligentie werd het begrip eerlijkheid gebruikt, als vertaling van het Engelse *fairness*. In dit hoofdstuk werd eerlijkheid gedefinieerd als ‘niet-discriminerend’. In het hoofdstuk van de Filosofie wordt de term ‘rechtvaardigheid’ impliciet gebruikt volgens de klassieke opvatting uit de Filosofie, waarbij het vooral gaat om vrijheid en gelijkheid in kansen. Het negatief beoordelen van een individu aan de hand van groepen waaronder het individu valt, wordt dan ook als onrechtvaardig aangewezen omdat een individu dan anders behandeld wordt op basis van die groep. Dit is in strijd met gelijke en vrije kansen. De vakgebieden Kunstmatige Intelligentie en Filosofie gebruikten dus de verschillende termen ‘eerlijkheid’ en ‘rechtvaardigheid’ voor dezelfde betekenis.

Doordat rondom het begrip discriminatie een *common ground* is gecreëerd, kan eerlijkheid nu worden gedefinieerd als ‘rechtvaardige profilering’ binnen het onderwerp van dit onderzoek. Om tot deze *common ground* te komen, hebben wij gebruikgemaakt van de integratietechniek *herdefinitie*, die de gemeenschappelijke betekenis van begrippen naar boven haalt (Repko & Szostak, 2017). Om verdere verwarring te voorkomen, zal vanaf nu het begrip ‘rechtvaardigheid’ worden gebruikt voor ‘eerlijkheid’.

Aannames

1. Vooroordelen in de data

Zowel in het hoofdstuk van de Kunstmatige Intelligentie als dat van de Filosofie wordt beweerd dat er onbewuste vooroordelen zitten in de data die algoritmes gebruiken. Het disciplinaire hoofdstuk van Geschiedenis licht toe hoe de vooroordelen over zwarte vrouwen zijn ontstaan. De bevindingen van het stuk van Geschiedenis zijn dus een aanname in de stukken van Kunstmatige Intelligentie en Filosofie.

Om dit conflict tussen aannames op te lossen, wordt de techniek *organisatie* gebruikt (Repko & Szostak, 2017). Hiermee wordt verduidelijkt hoe bepaalde fenomenen met elkaar in verband staan. De relatie is dat de Kunstmatige Intelligentie en Filosofie aannemen dat er vooroordelen bestaan in de data en dat de Geschiedenis hier een verklaring en bewijs voor geeft.

2. **Verantwoordelijkheid ontwikkelaars**

Uit het hoofdstuk vanuit de Filosofie wordt duidelijk dat in het hoofdstuk vanuit de Kunstmatige Intelligentie een aanname wordt gedaan. De Filosofie beargumenteerd waarom en in hoeverre ontwikkelaars verantwoordelijkheid dragen voor de onrechtvaardige uitkomsten en discriminatie door het algoritme. In het hoofdstuk vanuit Kunstmatige Intelligentie wordt deze vraag niet gesteld. Er werd direct gekeken naar strategieën die door de ontwikkelaars toegepast kunnen worden om onrechtvaardige profilering door een algoritme tegen te gaan. Dit impliceert de aanname dat ontwikkelaars verantwoordelijk zijn.

Om dit conflict op te lossen passen we de *organisatie*-techniek van Repko en Szostak (2017) toe: de stellingname vanuit Filosofie dat ontwikkelaars verantwoordelijkheid dragen gaat aan het hoofdstuk vanuit Kunstmatige Intelligentie vooraf.

<u>Theorieën:</u>

1. **Hume's Law en Intersectionaliteit**

In het hoofdstuk van de Kunstmatige Intelligentie wordt de theorie over intersectionaliteit aangehaald. Deze theorie stelt onder andere dat er systematische gecombineerde onderdrukkingen in de maatschappij zijn zoals racisme en seksisme, waardoor er maatschappelijke verschillen ontstaan. Profilerende algoritmes gebruiken daardoor *biased* data om nieuwe voorspellingen te doen. Onderdeel van de theorie van intersectionaliteit is dat dit niet gerechtvaardigd is, omdat de gegevens gebaseerd zijn op oneerlijke maatschappelijke processen. Door op basis daarvan nieuwe voorspellingen te doen, zijn deze op hun beurt ook onrechtvaardig.

Er is een overeenkomst op te merken met het probleem van onvoldoende bewijs uit de Filosofie. Hierin wordt allereerst gesteld dat algoritmes onrechtvaardig kunnen werken omdat ze causaliteit (zoals een geschiedenis van ongelijkheid) niet mee kunnen nemen in de berekening van correlaties, waardoor uitkomsten incorrect en onrechtvaardig uit kunnen pakken. Vervolgens wordt dit bij de Filosofie gecombineerd met *Hume's Law*, die stelt dat hoe iets is, niet per se is

hoe het zou moeten zijn. In feite zegt de theorie over intersectionaliteit dat omdat de situatie nu zo in elkaar zit en dit vaak op onrechtvaardige manier tot stand gekomen is, het ook niet gerechtvaardigd is om op basis daarvan nieuwe voorspellingen te maken. Hierin komen de theorieën dus met elkaar overeen, maar met *Hume's Law* is de theorie van de Filosofie net wat breder.

Om deze twee theorieën met elkaar te integreren gebruiken wij de techniek *extensie* (Repko & Szostak, 2017). Hiermee wordt het bereik van de ene theorie verbreed, om zo het conflict op te lossen. De theorie over intersectionaliteit verbreden wij met *Hume's Law*: hoe iets is, is niet per se hoe het zou moeten zijn. Dit hoeft dus niet noodzakelijk af te hangen van ongelijkheden uit de samenleving, al is dat wel (zoals ook in de Filosofie beargumenteerd wordt) een erg duidelijke en schadelijke manier waarop dit tot onrechtvaardige profilering en discriminatie kan leiden.

2. Verantwoordelijkheid van ontwikkelaars/gebruikers om 'alles' dat mogelijk is te doen

In het hoofdstuk vanuit de Filosofie wordt vrij stellig gesteld dat ontwikkelaars en gebruikers alles moeten doen dat mogelijk is om onrechtvaardige profilering en discriminatie door algoritmes te voorkomen. Hoewel het hoofdstuk duidelijk maakt dat het alleen gaat om oplossingen die ook daadwerkelijk mogelijk zijn, wordt er onvoldoende rekening gehouden met het feit dat die oplossingen erg beperkt zijn. In het hoofdstuk vanuit Kunstmatige Intelligentie zijn verschillende problemen geanalyseerd omtrent een eerlijk algoritme en is het duidelijk geworden dat deze oplossingen inderdaad nog erg beperkt zijn. Dit botst met de conclusie vanuit de Filosofie, daar wordt onvoldoende rekening gehouden met de technische mogelijkheden. Dit is echter simpelweg een gevolg van een gebrek aan kennis over de technische werking van algoritmes binnen de discipline Filosofie.

We kunnen nu met de integratietechniek *herdefinitie* 'alles dat mogelijk is doen' een nieuwe betekenis geven, namelijk: 'het optimale doen' (Repko & Szostak, 2017). We kiezen hier voor *herdefinitie* omdat uit de argumentatie vanuit de Filosofie duidelijk wordt dat als deze technische kennis bekend was geweest, de woordkeuze van 'het optimale doen' gekozen was.

Dit is af te leiden uit het feit dat de Filosofie stelt dat alleen alles gedaan moet worden dat mogelijk is en dat discriminatie daarmee zoveel mogelijk voorkomen moet worden. Uit de technische grenzen blijkt dat het in die context moet gaan om het optimale, niet om 'alles'. Het gaat er dus vooral om dat deze nuance bij de Filosofie nodig is om de argumenten juist op te vatten.

3. Cybertyping en transformatieve effecten

De Filosofie maakt gebruik van een theorie over 'transformatieve effecten'. Deze theorie houdt in dat algoritmes een versterkende werking kunnen hebben op denkbeelden en/of verhoudingen in de samenleving. De Filosofie laat zien dat stereotypes of ongelijkheden in de data terecht kunnen komen waardoor het algoritme deze opnieuw toepast op personen. Wanneer dat gebeurt kunnen mensen in bepaalde rollen geduwd worden, waarmee stereotypes in de praktijk worden bevestigd. Ook kunnen stereotypes in denkbeelden van mensen bevestigd worden door algoritmes die bepaalde stereotyperende informatie laten zien, bijvoorbeeld bij suggesties in zoekmachines op het internet.

De Geschiedenis wijst op een ontwikkeling van versterking van raciale stereotypes op het internet en duidt dit aan met de theorie 'cybertyping'. In deze theorie worden genderstereotypes wel genoemd, maar de focus ligt voornamelijk op raciale stereotypes. De theorie blijft hier dus wat onduidelijk over, dus zullen we dat in deze integratie verduidelijken.

Uit vergelijking blijkt dat de theorie van 'transformatieve effecten' een meer algemene theorie is over hoe algoritmes denkbeelden en structuren in de samenleving kunnen realiseren en versterken. In het hoofdstuk vanuit de Filosofie is dit toegepast op discriminatie jegens zwarte vrouwen. De theorie van 'cybertyping' vanuit Geschiedenis is een specifieke theorie die gaat over de versterking van voornamelijk raciale stereotypes op het internet. Dit omvat zowel de werking van algoritmes, als internetgebruik van personen die zelf stereotypes uitdragen. Om common ground te creëren is daarom zowel *extensie* als *intensie* (Repko & Szostak, 2017) van 'cybertyping' nodig en *intensie* (Alkema & Hamer, 2016) van 'transformatieve effecten'.

'Cybertyping' moet allereerst worden uitgebreid naar versterking van alle vormen van stereotypes door alle soorten profilerende algoritmes. Vanuit de Filosofie hebben we immers

gezien dat stereotypes niet alleen herbevestigd worden op het internet maar ook in bijvoorbeeld recruitmentsystemen. Daarbij gebeurt dit niet alleen ten aanzien van raciale stereotypes maar kan dit ook ten aanzien van bijvoorbeeld genderongelijkheid optreden of de combinatie daartussen, zoals we ook bij de Kunstmatige Intelligentie gezien hebben.

Vervolgens is *intensie* (Alkema & Hamer, 2016) nodig op 'cybertyping' omdat dit integratieve onderzoek gericht is op de werking van algoritmes. 'Cybertyping' wordt veroorzaakt door algoritmes, maar ook doordat mensen stereotypes uitdragen op het internet. Deze tweede oorzaak is voor dit onderzoek alleen relevant in hoeverre dit de verwerking van algoritmes beïnvloedt, maar die invloed is al aanwezig omdat de versterkende effecten van algoritmes onder andere veroorzaakt worden door *biased* data. De invloed van mensen die stereotypes uitdragen op het internet los daarvan, is dus niet relevant voor een integratieve theorie over algoritmes. Dat het buiten de nieuwe theorie over algoritmes valt die we hier formuleren, betekent natuurlijk niet dat het geen relevant inzicht is.

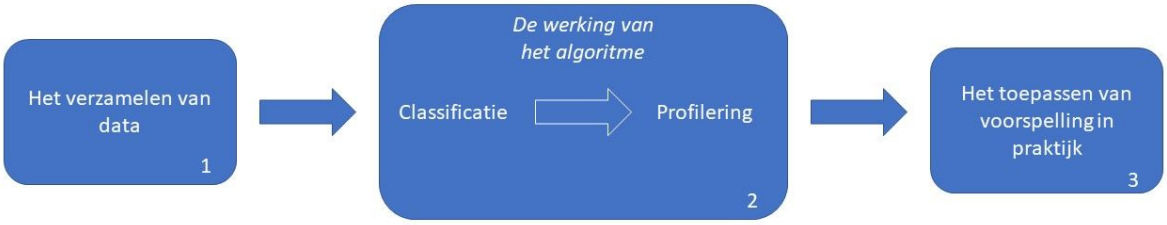
De theorie van 'transformatieve effecten' heeft *intensie* nodig (Alkema & Hamer, 2016) , omdat de theorie op zichzelf vooral gaat over hoe algoritmes veranderende effecten op de samenleving kunnen hebben. Het is hiermee dus nog niet specifiek gericht op schadelijke stereotypes die leiden tot discriminatie, terwijl de Filosofie deze stap wel tracht te maken in het hoofdstuk. Bij deze wordt de theorie daartoe gespecificeerd.

Om de aangepaste theorieën vervolgens te verenigen onder één naam en dit te onderscheiden van de oorspronkelijke theorieën waar het uit voortgekomen is, is *herdefinitie* (Repko & Szostak, 2017) nodig. De geïntegreerde theorie zal vanaf nu 'algoritmische herprofilering' worden genoemd. Daarmee bedoelen wij het effect dat bestaande schadelijke stereotypes en ongelijkheden uit de samenleving door het gebruik van profilerende algoritmes herbevestigd en zelfs versterkt worden, waardoor discriminatie optreedt als gevolg van algoritmes zodra de algoritmes toegepast worden in de praktijk.

4.3. More comprehensive understanding

Op basis van inzichten uit de disciplinaire delen en uit de *common ground* hebben wij een model gevormd dat de discriminerende werking van algoritmes in kaart brengt. Dit model is een visuele weergave van onze *more comprehensive understanding*: een interdisciplinair antwoord op de onderzoeksvraag.

Het uiteindelijke model (zie figuur 3) bestaat uit drie lagen, namelijk:1) het gebruik van een algoritme, 2) discriminatie door algoritmes, 3) oplossingen voor discriminatie door algoritmes. We zullen de verschillende lagen waaruit het schema bestaat één voor één toelichten, om zo antwoord te geven op onze hoofdvraag: *Welk handelingsperspectief kan worden afgeleid uit de oorzaken van de discriminatie jegens zwarte vrouwen door profilerende algoritmes?*

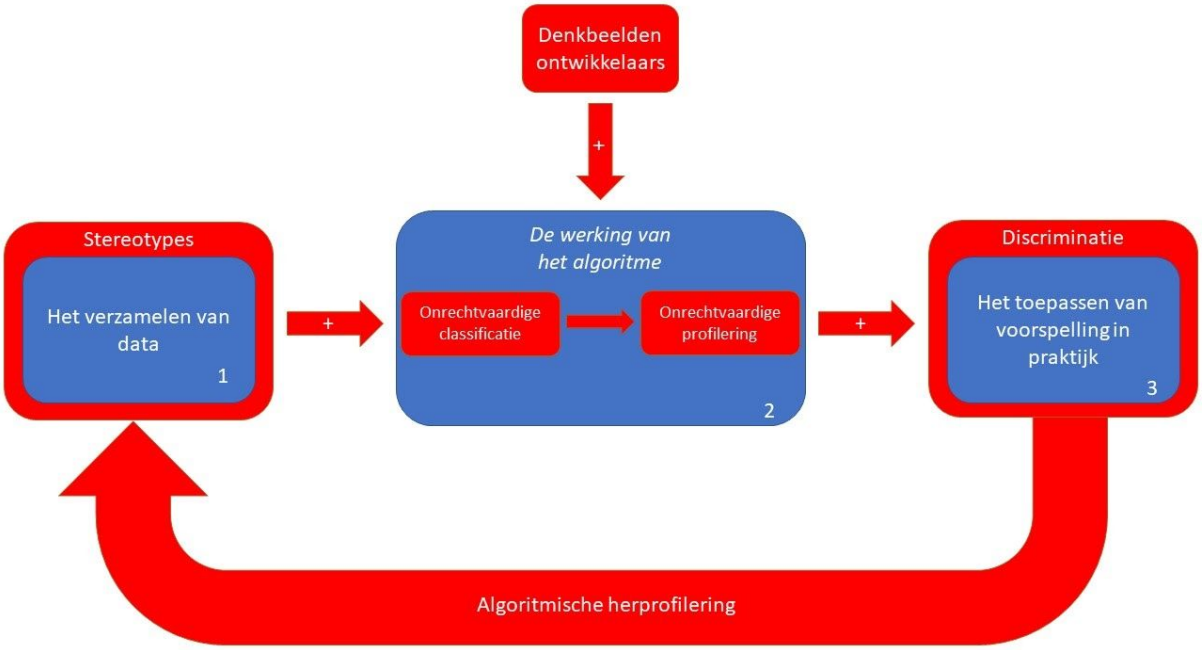


Afbeelding 4.3.1: de eerste laag van het model

Laag 1: het gebruik van algoritmes (blauw in het model)

De eerste laag illustreert het gebruik van algoritmes. De eerste stap hierin betreft het verzamelen van data. Stap twee omvat de werking van het algoritme. Het gaat hierbij om hoe het algoritme ontworpen is om data te verwerken door groeperingen (classificatie) en op basis daarvan

voorspellingen te doen (profilering). De derde stap betreft het toepassen van de voorspellingen van het algoritme in de praktijk: bedrijven en organisaties handelen naar de voorspellingen van het algoritme. De pijlen in deze laag illustreren enkel en alleen het lineaire tijdsverloop van de verschillende stappen, oftewel hoe de stappen elkaar in tijd opvolgen.



Afbeelding 4.3.2: de tweede laag van het model: discriminatie door algoritmes

Laag 2: discriminatie door algoritmes (rood in het model)

De tweede laag toont de discriminatie door algoritmes. Niet alle algoritmes discrimineren, maar in de praktijk gebeurt dit regelmatig. Centraal in deze laag staat de positieve terugkoppeling: de verschillende stappen in deze laag versterken elkaar. De pijlen geven hier dus oorzaak-gevolgrelaties aan. Iedere rode pijl met een plus versterkt nu discriminatie, vandaar dat er een plusteken in staat. Een rode pijl zonder plus geeft, net zoals in de eerste laag, een tijdsverloop aan.

De eerste stap uit de tweede laag betreft het gegeven dat er stereotypes zitten in de data die verzameld worden. Deze stereotypes zijn al in de samenleving aanwezig en komen daardoor in data terecht. Het hoofdstuk van Geschiedenis heeft laten zien dat de stereotypes voortkomen uit een geschiedenis van ongelijkheid.

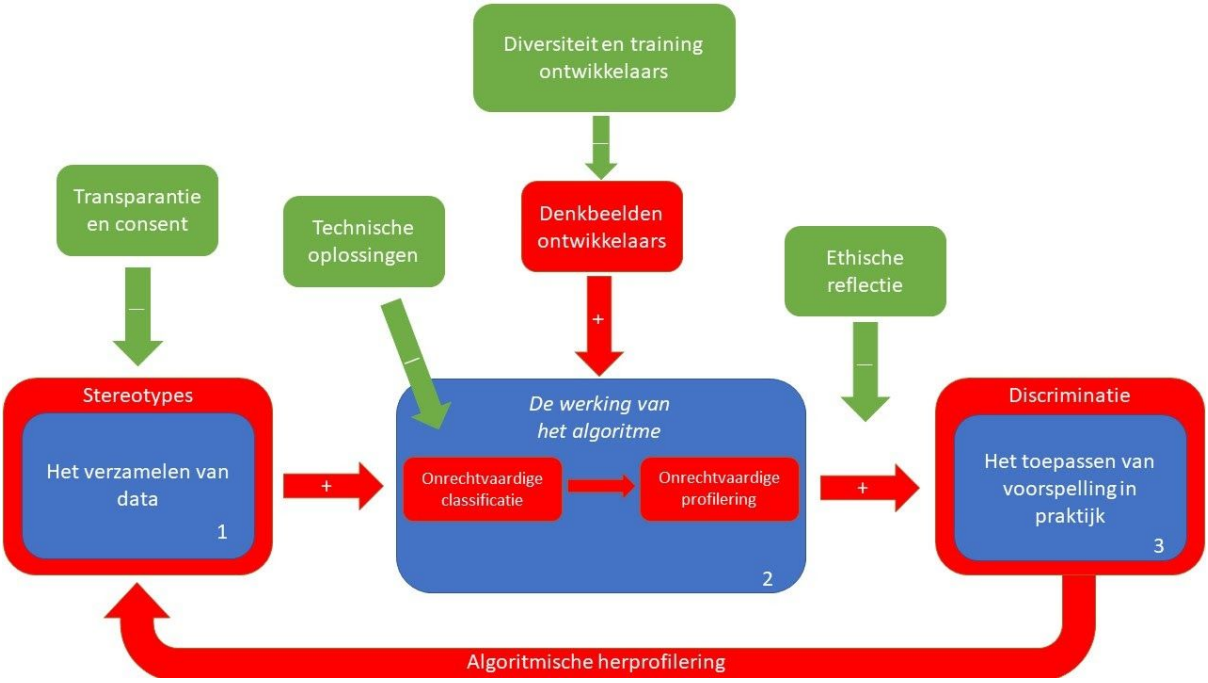
De rode pijl tussen de eerste en tweede stap illustreert dat data die stereotypes bevatten gebruikt worden door algoritmes voor classificatie. Doordat deze data met stereotypes gebruikt worden, is de classificatie door algoritmes onrechtvaardig. Dit is in het model zichtbaar in de tweede stap. Deze betreft de onrechtvaardige classificatie door algoritmes, gevolgd door onrechtvaardige profilering. Deze stap wordt behalve door stereotypes in de data ook beïnvloed door een nieuw blok: de denkbeelden van de programmeurs. Deze denkbeelden kunnen bevooroordeeld zijn jegens zwarte vrouwen. Een samenwerking binnen een homogene groep programmeurs kan ertoe leiden dat deze denkbeelden eenzijdig zijn en dat doorwerking van stereotypes niet opgemerkt wordt. Dit versterkt de onrechtvaardige werking van het algoritme, wat is aangegeven met de rode pijl met een plusteken.

De pijl tussen de tweede en derde stap illustreert hoe de onrechtvaardige werking van een algoritme kan leiden tot discriminatie, wat gebeurt in stap drie. Discriminatie is in de *common ground* gedefinieerd als het toepassen van onrechtvaardige profilering. De derde stap (discriminatie) vindt dus plaats als een onrechtvaardig algoritme wordt toegepast.

In deze laag is er een laatste pijl bijgekomen, die de derde stap verbindt aan de eerste stap. Dit maakt de processen in de tweede laag tot een positieve terugkoppeling*:* niet alleen bestaat de laag uit verschillende processen die elkaar versterken, maar het laatste proces beïnvloedt uiteindelijk ook weer het eerste proces, en daarmee ook weer de andere stappen. Deze pijl beschrijft het fenomeen ‘cybertyping’. Dat houdt in dat stereotypes die door discriminerende

algoritmes in de praktijk worden gebracht, er ook toe leiden dat die stereotypes blijven bestaan in de samenleving. Als gevolg daarvan komen die stereotypes ook weer in de data terecht en is er dus sprake van een positieve terugkoppeling.

De gehele hierboven beschreven tweede laag kan gezien worden als een hedendaagse uiting van *racializing surveillance*. Dit begrip wordt door historici gebruikt om te verwijzen naar de manier waarop surveillancepraktijken normen produceren over 'ras'. Surveillanceprakijken zijn verschillende manieren waarop mensen gecontroleerd en gedisciplineerd worden, bijvoorbeeld door vanuit wachttorens toezicht op mensen te houden of door mensen te verplichten een identificatiedocument bij zich te dragen. De uitkomst van *racializing surveillance* is vaak een discriminerende behandeling van mensen van een bepaald 'ras', doordat de normen die volgen uit surveillancepraktijken bepalen wat zij wel en niet kunnen doen. *Racializing surveillance* vond niet alleen plaats in het verleden. Ook vandaag de dag blijken mensen van een bepaald 'ras' vaak gediscrimineerd te worden middels surveillancepraktijken. Denk aan zwarte mensen die door bewakingssystemen van de politie worden benadeeld. Als discriminatie door algoritmes betrekking heeft op mensen van een bepaald 'ras', kan dit daarom gezien worden als een hedendaagse uiting van *racializing surveillance.*



Afbeelding 4.3.3: de derde laag van het model: een handelingsperspectief

Laag 3: een handelingsperspectief (groen in het model)

Onze analyse van de werking en oorzaken van discriminerende algoritmes biedt nu een handelingsperspectief om in de toekomst algoritmes discriminatie-vrij te maken. In de disciplinaire delen van deze scriptie zijn deze oplossingen al voorbij gekomen, maar nu zullen we deze ook relateren aan de stappen die wij in de tweede laag van ons schema hebben onderscheiden om het zo tot één geheel te maken. De derde laag maakt duidelijk waar stappen ondernomen moeten worden en biedt daarmee een geïntegreerd handelingsperspectief.

In deze derde laag zijn de oplossingen weergeven met groene blokken, waaruit groene pijlen vertrekken. In deze pijlen is een minteken geplaatst, om te illustreren dat ze de positieve terugkoppeling verminderen en uiteindelijke discriminatie door algoritmes tegengaan.

Tijdens de eerste stap, verzameling van data, is het van belang om informatie over de risico's van profilerende algoritmes beter bekend te maken aan datasubjecten, oftewel aan mensen van wie data verzameld wordt. Een eerste oplossing is dus 'Transparantie en consent'. In deze oplossing worden de datasubjecten geïnformeerd over het verzamelen van hun gegevens. Op deze manier wordt een extra stap van controle mogelijk gemaakt over de verwerking van data

door profilerende algoritmes, omdat datasubjecten op dit punt de mogelijkheid krijgen om hun data aan het algoritme te onttrekken en zich op z'n minst bewust worden van de risico's. Zo wordt in ieder geval aan hun recht voldaan om een geïnformeerde toestemming te geven. Zo wordt het ook makkelijker voor datasubjecten om bezwaar aan te tekenen tegen het gebruik van een algoritme, dan wanneer ze niet eens op de hoogte zijn van de risico's.

Vervolgens is het in de tweede stap belangrijk om zoveel mogelijk diversiteit in teams van programmeurs te waarborgen. Een divers team merkt stereotypes in de data namelijk eerder op en neemt zelf ook minder eenzijdige stereotypes mee in het ontwerpen van algoritmes. Ook moeten de programmeurs bewust worden gemaakt van hun eigen vooroordelen door middel van trainingen. Tezamen vormen deze oplossingen een nieuwe pijl 'Diversiteit en training ontwikkelaars' die de rode pijl 'Denkbeelden van ontwikkelaars' afzwakt.

Tijdens het ontwerpen van algoritmes moeten ook technische oplossingen worden toegepast om onrechtvaardige uitkomsten tegen te gaan. Dit kan worden gedaan door normatieve regels in algoritmes te implementeren (wanneer dat mogelijk is) en om strategieën voor rechtvaardigheid toe te passen uit de Kunstmatige Intelligentie. In dit disciplinaire hoofdstuk is gesteld dat er nog meer onderzoek nodig is naar het ontwerpen van een rechtvaardig algoritme. Er is echter al wel vooruitgang geboekt en er zijn pogingen gedaan tot een rechtvaardig algoritme, die zouden kunnen worden ingezet in deze stap.

Na de tweede stap, als het algoritme is ontworpen en voorspellingen heeft gemaakt, zouden de voorspellingen onderworpen moeten worden aan een 'ethische reflectie'. Deze reflectie kan helpen bij het maken van morele keuzes, bijvoorbeeld bij de beslissing om het algoritme wel of niet in de praktijk te gaan toepassen. Deze ethische reflectie is een samenwerking tussen verschillende partijen. Ten eerste moeten ontwikkelaars die het algoritme hebben gemaakt zich verantwoorden tegenover organisaties en bedrijven over de werking van hun algoritme. Welke keuzes hebben zij bij het ontwerpen van het algoritme gemaakt en waarom? Het is belangrijk dat bedrijven en organisaties die het algoritme gebruiken hiervan op de hoogte zijn, zodat zij vervolgens hun eigen morele afwegingen kunnen maken. Als ze deze ethische reflectie niet uitvoeren maar ze de algoritmes wel inzetten, dan kunnen ze verantwoordelijk worden gehouden voor de mogelijk discriminerende uitwerking van het

inzetten van onrechtvaardige algoritmes. In de praktijk blijkt ethische reflectie op de algoritmes voor bedrijven en organisaties moeilijk, omdat zij niet altijd de precieze technische details van het algoritme kunnen snappen. Ook hebben zij eigen belangen die soms boven het ethisch gebruik van algoritmes worden geplaatst. Het zou daarom goed zijn als ethische reflectie ten tweede wordt uitgevoerd door een onafhankelijke derde partij die gespecialiseerd is in algoritmes en discriminatie. Deze derde partij kan eens in de zoveel tijd ter controle steekproeven uitvoeren bij gebruikers die algoritmes inzetten. Mocht er dan bevonden worden dat het algoritme discriminerende effecten heeft, dan moeten de ontwikkelaars en gebruikers zorgen dat ze op zoek gaan naar manieren om het algoritme rechtvaardiger te maken.

De verschillende oplossingen die hierboven zijn gepresenteerd kunnen een handelingsperspectief bieden om de positieve feedbackloop van discriminerende algoritmes te doorbreken. Als deze oplossingen worden aangegrepen, dan zal de pijl 'algoritmische herprofilering' verzwakt worden en mogelijk zelfs geleidelijk verdwijnen. Onze *more comprehensive understanding* rechtvaardigt onze interdisciplinaire aanpak: zonder een interdisciplinaire aanpak had het drielaagsmodel niet opgesteld kunnen worden. De verschillende disciplinaire inzichten waren dan niet samengekomen tot één geheel van oorzaken en oplossingen.

Discussie

Intersectionaliteit

De drie disciplinaire hoofdstukken uit deze scriptie illustreren goed het intersectionele probleem van discriminatie door algoritmes. De Filosofie ging vooral in op gender, de Geschiedenis op 'ras'. De Kunstmatige Intelligentie benadrukte dat zwarte vrouwen juist door apart te focussen op racisme en seksisme vergeten kunnen worden, vanwege de interferentie tussen racisme en seksisme. Daarnaast focuste de Kunstmatige Intelligentie zich op gemarginaliseerde groepen in het algemeen, waar zwarte vrouwen ook onder vallen. Toch hadden wij misschien tot een nog sterkere integratie kunnen komen wanneer wij ons allemaal óf op dezelfde categorie van maatschappelijke uitsluiting hadden gefocust óf op de intersectionaliteit van deze factoren.

Overige relevante disciplines

Alhoewel de verschillende disciplines uit dit onderzoek elkaar zeer goed aanvullen, hadden inzichten uit verschillende andere disciplines ons onderzoek nog kunnen versterken. Vanuit onze eigen disciplines hebben wij in ons onderzoek meer gefocust op de rol van individuele programmeurs in de discriminatie door algoritmes. Echter, discriminatie door algoritmes betreft een institutioneel probleem en daarom was het nuttig geweest om inzichten vanuit de Bestuurs- en Organisatiewetenschappen bij ons onderzoek te kunnen betrekken. Discriminatie door algoritmes is namelijk een probleem dat sterk samenhangt met de beslissingen van bedrijven en organisaties. Zo worden algoritmes ontwikkeld in bedrijven en zijn er vervolgens andere bedrijven en organisaties die de algoritmes in de praktijk toepassen. Met onderzoek vanuit Bestuurs- en Organisatiewetenschappen zou wellicht bepaald kunnen worden hoe bedrijven en organisaties kunnen zorgen voor meer diverse teams. Ook zou dit één van de oplossingen die wij hebben geschetst, om een onafhankelijke controlerende derde partij samen te stellen die bedrijven en organisaties controleert, concreter kunnen maken.

Een tweede discipline die ons onderzoek had kunnen versterken is de Gender en Postkoloniale Studies. Deze discipline had meer inzichten kunnen bieden in de theorie over intersectionaliteit en hoe onderdrukking in de maatschappij werkt. Daarnaast had dit perspectief

meer context kunnen geven over hoe zwarte vrouwen benadeeld kunnen worden door stereotypering.

Een laatste discipline die ons onderzoek had kunnen versterken is de Toegepaste Cognitieve Psychologie, aangezien wij in ons onderzoek veel schrijven over stereotypes en vooroordelen. Wij stellen op basis van eerder onderzoek dat mensen stereotypes hebben over bijvoorbeeld zwarte vrouwen en dat deze stereotypes ook verborgen liggen in de data die algoritmes gebruiken. Vanuit de Psychologie was het mogelijk geweest om meer inzicht te krijgen in de manier waarop mensen aan vooroordelen komen en hoe ze er vanaf kunnen komen. Deze inzichten hadden kunnen worden gebruikt om meer concrete suggesties te doen voor bijvoorbeeld de training van programmeurs om hen van hun stereotypes te laten afkomen. Verder onderzoek daarnaar is dus nog nodig.

Suggesties voor vervolgonderzoek

Oorspronkelijk was het de bedoeling om deze scriptie te schrijven over discriminatie jegens vrouwen van kleur. Uiteindelijk hebben we ervoor gekozen om het onderzoek specifiek op zwarte vrouwen te richten. Deels hebben we deze overweging gemaakt op basis van de theorie van *colorisme*, en deels om grenzen te stellen aan het onderzoek om het werkbaar te houden. Zoals eerder genoemd kan met een concretere toepassing ook een sterkere integratie gerealiseerd worden. Echter was een andere reden ook dat er heel weinig literatuur bestaat over discriminatie jegens bijvoorbeeld vrouwen met een Aziatische achtergrond. Dit laat eigenlijk juist zien dat het belangrijk is dat ook daar onderzoek naar wordt gedaan. Hoewel dit onderzoek gericht is geweest op discriminatie ten aanzien van zwarte vrouwen, zijn de problemen en oplossingen die wij in kaart hebben gebracht echter wel te veralgemeniseren naar andere vormen van discriminatie. Ook daar kunnen de genoemde oplossingen dus bij helpen. Dit neemt niet weg dat verder onderzoek nodig is om hier diepgaande inzichten over te verkrijgen.

Zowel het onderzoek vanuit Kunstmatige Intelligentie als Filosofie hebben laten zien dat er nog veel technische ontwikkelingen nodig zijn om algoritmes eerlijker en ethischer tot voorspellingen te laten komen. Ook hier is verder onderzoek dringend nodig.

Bewustwording

Een belangrijk onderdeel van de oplossingen uit dit onderzoek is niet besproken, namelijk bewustwording. Als er weinig bewustzijn is ten aanzien van racisme, seksisme, de combinatie daartussen en van discriminatie veroorzaakt door algoritmes, komen onderzoeken zoals deze maar weinig aan het licht en wordt er weinig druk uitgevoerd op de verantwoordelijken om er iets aan te doen. De huidige Black Lives Matter beweging die na de dood van George Floyd is ontstaan, zorgt momenteel voor een groter bewustzijn. Daarom is het zo belangrijk dat onderzoeken zoals deze nu uitgevoerd worden, nu er grote aandacht voor is en mensen de tijd nemen om over institutioneel racisme te leren.

Bibliografie (APA citation style)

Voorkant, Voorwoord en Inleiding

Black Girls Code (2018). Black Girls Code Holiday Party [foto]. Geraadpleegd van https://secure.givelively.org/donate/black-girls-code-inc/black-girls-code-holiday-party-2018/fundraisers/new

Bolukbasi, T., Chang, K. W., Zou, J.Y., Saligrama, V. & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

Buolamwini, J. & Gebru, T.. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR* 81:77-91. Geraadpleegd van Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Cheney-Lippold, J. (2017). Categorization: Making Data Useful. In J. Cheney-Lippold (ed.), *We Are Data. Algorithms and the MAking of Our Digital Selves* (pp. 37-92). New York, USA: NYU Press.

Dastin, J. (2018, 10 oktober). 'Amazon scraps secret AI recruiting tool that showed bias against women'. *Reuters.* Geraadpleegd van:

Harding, S. G. (2004). *The feminist standpoint theory reader: Intellectual and political controversies*. New York, USA: Routledge. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Hijink, M. (2020, 11 juni). Amazon, IBM en Microsoft zetten rem op omstreden software voor gezichtsherkenning. *NRC Handelsblad*. Geraadpleegd van https://www.nrc.nl/nieuws/2020/06/11/amazon-en-ibm-zetten-rem-op-omstreden-software-voor-gezichtsherkenning-a4002551

Hill, K. (2020, 25 juni). Wrongfully Accused by an Algorithm. *The New York Times*. Geraadpleegd van https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

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.

ProPublica: Angwin, J., Larson, J., Mattu, S. & Kirchner. L. (2016, 23 mei). *Machine Bias: There's software used across the country to predict future criminals. And it's biased*

against blacks. Geraadpleegd van https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Repko, A. & Szostak, R. (2017). *Interdisciplinary Research: Process and Theory.* Thousand Oaks, USA: SAGE Publications

Squires, C. R. (2002). Rethinking the black public sphere: An alternative vocabulary for multiple public spheres. *Communication theory*, *12*(4), 446-468.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. Verkregen op 21-05-2020 van: https://www.nature.com/articles/d41586-018-05707-8?source=post_page-----817fa60d75e9----------&fbclid=IwAR0iajZ2WRp7mlrW4gy10X1jsmJLvRH-8blaSI0vGO_OMK05yJ

Hoofdstuk 1: Geschiedenis

Andrejevic, M. (2004). The work of watching one another: Lateral surveillance, risk, and governance. *Surveillance & Society,* 2(4), 479-497.

Baker Jr, H.A. (1994). Critical memory and the black public sphere. *Public Culture*, *7*(1), 3-33.

Brock, A. (2011). Beyond the pale: The Blackbird web browser's critical reception. *New Media & Society,* 13(7), 1085-1103.

Browne, S. (2015). *Dark matters: On the surveillance of blackness*. Durham, North Carolina: Duke University Press.

Collins, P.H. (1990). Black feminist thought in the matrix of domination. In P. H. Collins (ed.), *Black feminist thought: Knowledge, consciousness, and the politics of empowerment* (pp. 221-238). London, UK: HarperCollins.

Collins, P.H. (1998). *Fighting words: Black women and the search for justice*. Minneapolis, Minnesota: University of Minnesota Press.

Daniels, J. (2015). "My Brain Database Doesn't See Skin Color" Color-Blind Racism in the Technology Industry and in Theorizing the Web. *American Behavioral Scientist*, *59*(11), 1377-1393.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. Cited in S.L. Gilman (1985). Black bodies, white bodies: Toward an iconography of female sexuality in late nineteenth-century art, medicine, and literature. *Critical Inquiry*, *12*(1), 219.

Dyer, R. (2000). The matter of whiteness. In Back, L., & Solomos, J. (eds.), *Theories of race and racism: A reader* (pp. 539-548). London, UK: Routledge.

Ferris State University, Jim Crow Museum of Racist Memorabilia (z.d.). Jezebel Stereotypes Gallery 02. https://www.ferris.edu/HTMLS/news/jimcrow/jezebel/jezebel-stereotype-gallery-02.htm

Foucault, M. (1975). Surveiller et punir. *Paris,* 1, 192-211.

Gilman, S.L. (1985). Black bodies, white bodies: Toward an iconography of female sexuality in late nineteenth-century art, medicine, and literature. *Critical Inquiry*, *12*(1), 204-242.

Gold, N. (2018, 18 september). *Too few women in tech? Blame 1980s marketing!* Geraadpleegd van https://wearetechwomen.com/too-few-women-in-tech-blame-1980s-marketing/

Google (2020, 5 mei). *Google Diversity Annual Report 2020*. Geraadpleegd van https://diversity.google/

Habermas, J. (1962) [1991]. *The Structural Transformation of the Public Sphere: an Inquiry into a Category of Bourgeois Society* [Strukturwandel der Öffentlicheit] (T. Burger & F. Lawrence, vert.). Cambridge, USA: MIT Press.

Habermas, J., Lennox, S., & Lennox, F. (1974). The public sphere: An encyclopedia article (1964). *New German Critique*, (3), 49-55.

Hardt, H. (1996). The Making of The Public Sphere: Class Relations and Communication in The United States. *Javnost-The Public,* 3(1), 7-23.

Historiek (2019, 4 december). *Hottentottententententoonstelling: herkomst van een vreemd woord.* Geraadpleegd van https://historiek.net/hottentottententententoonstelling-herkomst/80252/

Hooyman, P. (2019, 30 oktober). Het ware gezicht van gezichtsherkenningstechnologie. *Bits of Freedom*. Geraadpleegd van https://www.bitsoffreedom.nl/2019/11/21/het-ware-gezicht-van-gezichtsherkenning/

JOHO (z.d.). *Wat is etniciteit? Chapter 1*. Geraadpleegd van https://www.joho.org/nl/wat-etniciteit-chapter-1#website

Landes, J.B. (1988). Introduction and Conclusion. In J.B. Landes (ed.), *Women and the Public Sphere in the Age of the French Revolution* (pp. 1-13 and 201-206). Ithaca, USA: Cornell University Press.

Lyon, D. (2003). *Surveillance as social sorting*. London, UK: Routledge.

Nakamura, L. (2009). Don't Hate the Player, Hate the Game: the Racialization of Labor in World of Warcraft. *Critical Studies in Media Communication*, 26(2), 128-144.

Noble, S.U. (2018). Searching for Black Girls. In S.U. Noble (ed.), *Algorithms of Oppression: How Search Engines Reinforce Racism* (pp. 64-109). New York, USA: NYU Press.

NPR (2014, 21 oktober). *When Women Stopped Coding*. Geraadpleegd van https://www.npr.org/sections/money/2014/10/21/357629765/when-women-stopped-coding?t=1592062910954

Pilgrim, D. (2012a). The Mammy Caricature. *Ferris State University, Jim Crow Museum of Racist Memorabilia.* https://www.ferris.edu/jimcrow/mammies/

Pilgrim, D. (2012b). The Sapphire Caricature. *Ferris State University, Jim Crow Museum of Racist Memorabilia.* https://www.ferris.edu/HTMLS/news/jimcrow/antiblack/sapphire.htm

RoSa: kenniscentrum van gender en feminisme (z.d.). *RoSa bespreekt: Ain't I a Woman: Black Women and Feminism – bell hooks.* Geraadpleegd van https://rosavzw.be/site/bibliotheek-archief/rosa-bespreekt/781-rosa-bespreekt-ain-t-i-a-woman-black-women-and-feminism-bell-hooks

Siapera, E. (2010). Security, Surveillance and Safety. In E. Siapera (ed.), *Understanding New Media* (pp. 105-124). London, UK: Sage.

Sinclair, B. (2004). Integrating the histories of race and technology. In B. Sinclair (ed.), *Technology and the African American experience: Needs and opportunities for study* (pp. 1-17). Cambridge, USA: The MIT Press.

Van Schie, G. (2020, 23 maart). *Race and Technology: 2. Understanding Technology and Race* [YouTube]. Geraadpleegd van https://www.youtube.com/watch?v=pUcLQF2DFQs&list=PLFUkCaxh3298Y6wfYKFLxq1JyPhIvHmBP&index=3&t=373s

Yanow, D. (2013, 19 oktober). Ook Nederland hanteert nog ras als categorie. *Trouw*. Geraadpleegd van https://www.trouw.nl/nieuws/ook-nederland-hanteert-nog-ras-als-categorie~b54d3325/

Hoofdstuk 2: Kunstmatige Intelligentie

Alpaydin, E. (2020). *Introduction to machine learning*. (Eerste editie). Cambridge, Massachusetts: MIT press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. Verkregen op 21-05-2020 van: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., Hardt, M., & Narayanan, A. (2017). *Fairness in machine learning*. NIPS Tutorial.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *arXiv: 1712.03586*

Bolukbasi, T., Chang, K. W., Zou, J.Y., Saligrama, V. & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

Buolamwini, J. & Gebru, T.. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR* 81:77-91.

Corbett-Davies, S., & Goel S.. (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023.*

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 144-152).

Foulds, J., & Pan, S. (2018). An intersectional definition of fairness. *arXiv preprint arXiv:1807.08362*.

Hajian, S., Bonchi, F., & Castillo, C. (2016, August). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126).

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. In *Information, Communication & Society, 22:7,* (pp. 900-915).

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2017). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lahoti, P., Gummadi, K. P., & Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (pp. 1334-1345).

Malazita, J. W., & Resetar K. (2019). Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity* 30:4. (pp. 300-312).

Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency* (pp. 107-118).

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, *11*(3), (pp. 1193-1216).

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7).

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259.*

Khuzani, A. Z., Heidari, M., & Shariati, A. (2020). COVID-Classifier: An efficient machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images. *medRxiv*.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. Verkregen op 21-05-2020 van: https://www.nature.com/articles/d41586-018-05707-8?source=post_page-----817fa60d75e9----------&fbclid=IwAR0iajZ2WRp7mlrW4gy10X1jsmJLvRH-8blaSI0vGO_OMK05yJ

Hoofdstuk 3: Filosofie

Barocas, S. (2014, August). Data mining and the discourse on discrimination. *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining* (pp. 1-4).

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology 15(3):* 209–227.

Careerbuilder (2017, 18 mei). 'More Than Half of HR Managers Say Artificial Intelligence Will Become a Regular Part of HR in Next 5 Years'. Geraadpleegd van: http://press.careerbuilder.com/2017-05-18-More-Than-Half-of-HR-Managers-Say-Artificial-Intelligence-Will-Become-a-Regular-Part-of-HR-in-Next-5-Years

Dastin, J. (2018, 10 oktober). 'Amazon scraps secret AI recruiting tool that showed bias against women'. *Reuters.* Geraadpleegd van: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Hildebrandt, M. and Koops, B-J. (2010). The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review 73(3):* 428–460.

Hume, David. 1978 [1739]. A Treatise of Human Nature. 2*nd Edition; L. A. Selby-Bigge and P.H. Nidditch (eds.).* Oxford: Clarendon Press. Selections

Johnson, J.A. (2006). Technology and pragmatism: From value neutrality to value criticality. *SSRN Scholarly Paper, Rochester,* NY: Social Science Research Network.

Lewis, S.C. & Westlund, O. (2015) Big data and journalism. *Digital Journalism 3(3):* 447–466.

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.

Newell, S. & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems 24(1)*

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

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

Rubel, A. & Jones, K.M.L. (2014). Student privacy in learning analytics: An information ethics perspective. *SSRN Scholarly Paper*, Rochester, NY: Social Science Research Network.

Sandvig C, Hamilton K, Karahalios K, et al. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry.*

Schermer, B.W. (2011). The limits of privacy in automated profiling and data mining. *Computer Law & Security Review 27(1)*

The Census Bureau (2017). Computer Programmers. *Data USA.* Geraadpleegd van: https://datausa.io/profile/soc/computer-programmers#demographics

Turilli, M. & Floridi, L. (2009). The ethics of information transparency. E*thics and Information Technology 11(2):* 105–112

Zarsky, T.(2013). Transparent predictions. *University of Illinois Law Review 2013(4).*

Common Ground

Alkema, J.G. & Hamer, T. (2016) De Vernietiging van de Vernietiging: een interdisciplinair onderzoek naar de verandering van eigendom in een circulaire economie. [scriptie] Utrecht: Universiteit Utrecht.

Repko, A. & Szostak, R. (2017). *Interdisciplinary Research: Process and Theory.* Thousand Oaks, USA: SAGE Publications

van der Lecq, R. (2012). Why we talk: An interdisciplinary approach to the evolutionary origin of language. In A. F. Repko, W. H. Newell, & R. Szostak (Eds.), *Case studies in interdisciplinary research* (pp. 191–223). Thousand Oaks, CA: Sage.

Conclusie: more comprehensive understanding

Repko, A. & Szostak, R. (2017). *Interdisciplinary Research: Process and Theory.* Thousand Oaks, USA: SAGE Publications