De (On)mogelijkheden van machine learning voor het verminderen van bias en discriminatie bij personeelsbeslissingen

Hiemstra, Annemarie M.F.; Cassel, Tatjana; Born, Marise Ph; Liem, Cynthia C.S.

DOI
10.5117/2020.033.004.002

Publication date
2020

Document Version
Final published version

Published in
Gedrag en Organisatie

Citation (APA)
Hiemstra, A. M. F., Cassel, T., Born, M. P., & Liem, C. C. S. (2020). De (On)mogelijkheden van machine learning voor het verminderen van bias en discriminatie bij personeelsbeslissingen. *Gedrag en Organisatie*, *33*(4), 279-299. https://doi.org/10.5117/2020.033.004.002

Boom

info@boomamsterdam.nl
www.boomuitgeversamsterdam.nl

ARTIKELEN

De (on)mogelijkheden van machine learning voor het verminderen van bias en discriminatie bij personeelsbeslissingen

Annemarie M. F. Hiemstra, Tatjana Cassel, Marise Ph. Born & Cynthia C. S. Liem[*]

In dit artikel zetten we uiteen wat de toepassing van machine learning-algoritmes voor personeelsselectie inhoudt en hoe deze data-gedreven werkwijze overeenkomt met en verschilt van de klassieke selectiepsychologie. Aansluitend bespreken we of, en op welke manier, er bias en discriminatie kan optreden bij het gebruik van algoritmes gebaseerd op machine learning voor personeelsselectie. Hiervoor voerden we een literatuurstudie uit (periode 2016-2019), waarbij we 41 artikelen selecteerden. De resultaten geven aan dat algoritmes mogelijkerwijs leiden tot minder (indirecte) discriminatie vergeleken met sommige andere selectiemethoden. Dat is een van de redenen waarom de ontwikkeling van algoritmes voor selectie zo snel is gegaan en het aantal aanbieders is toegenomen. Het is echter onvoldoende mogelijk om vast te stellen of de belofte ook wordt ingelost. Dit komt deels doordat algoritmes vaak bedrijfsgeheim zijn (geen transparantie) en vanwege onduidelijkheden over de validiteit en betrouwbaarheid van data die gebruikt worden om algoritmes te ontwikkelen. Selectiepsychologische vraagstukken rondom diversiteit en validiteit zijn (nog) niet opgelost via de ontwikkeling van algoritmes. De toenemende aandacht voor het onderwerp, getuige de sterke groei van het aantal publicaties, stemt hoopvol. We besluiten met aanbevelingen voor het traceren en verminderen van bias en discriminatie bij het gebruik van algoritmes voor selectie.

1 Inleiding

In de nabije toekomst zijn selectiecommissies overbodig: de computer kan dan beslissen wie de meest geschikte persoon is voor een baan. Geen tijdrovende interviews meer, of intensieve discussies over de geschiktheid van een kandidaat, maar juist snel en efficiënt genomen beslissingen. Zonder bias! Dit toekomstbeeld is al de optimistische realiteit, volgens sommige aanbieders van beslissingsondersteunende software, zoals Harver, Hirevue en Pymetrics (Raghavan, Barocas, Kleinberg, & Levy, 2019). Zij stellen vooral dat het gebruik van artificiële intelligentie voor personeelsvraagstukken, zoals algoritmes gebaseerd op 'machine lear-

* Annemarie M.F. Hiemstra, Tatjana Cassel en Marise Ph. Born zijn verbonden aan de Erasmus Universiteit Rotterdam, Erasmus School of Social and Behavioral Sciences, Department of Psychology, Education and Child Sciences. Cynthia C.S. Liem is verbonden aan de Technische Universiteit Delft, Faculteit Elektrotechniek, Wiskunde en Informatica, afdeling Intelligent Systems, Multimedia Computing Group. Correspondentie naar aanleiding van dit artikel kan gericht worden aan Annemarie M.F. Hiemstra, hiemstra@essb.eur.nl.

ning', kan leiden tot meer inclusieve werkomgevingen. Verschil in baankansen en promotie voor werkzoekenden en werknemers vanuit diverse minder gerepresenteerde groepen, zoals cultureel diverse werknemers, is immers nog steeds veelvoorkomend en een complex probleem.

Er wordt in toenemende mate wetenschappelijk onderzoek gedaan naar de mogelijkheden van het gebruik van algoritmes voor selectie (Liem et al., 2018). We geven in deze bijdrage een overzicht van de stand van zaken in de literatuur over het toenemende gebruik van algoritmes gebaseerd op machine learning voor selectie. Daarbij gaan we in het bijzonder in op vraagstukken rondom discriminatie, eerlijkheid, 'adverse impact' en bias (zie onder 3).

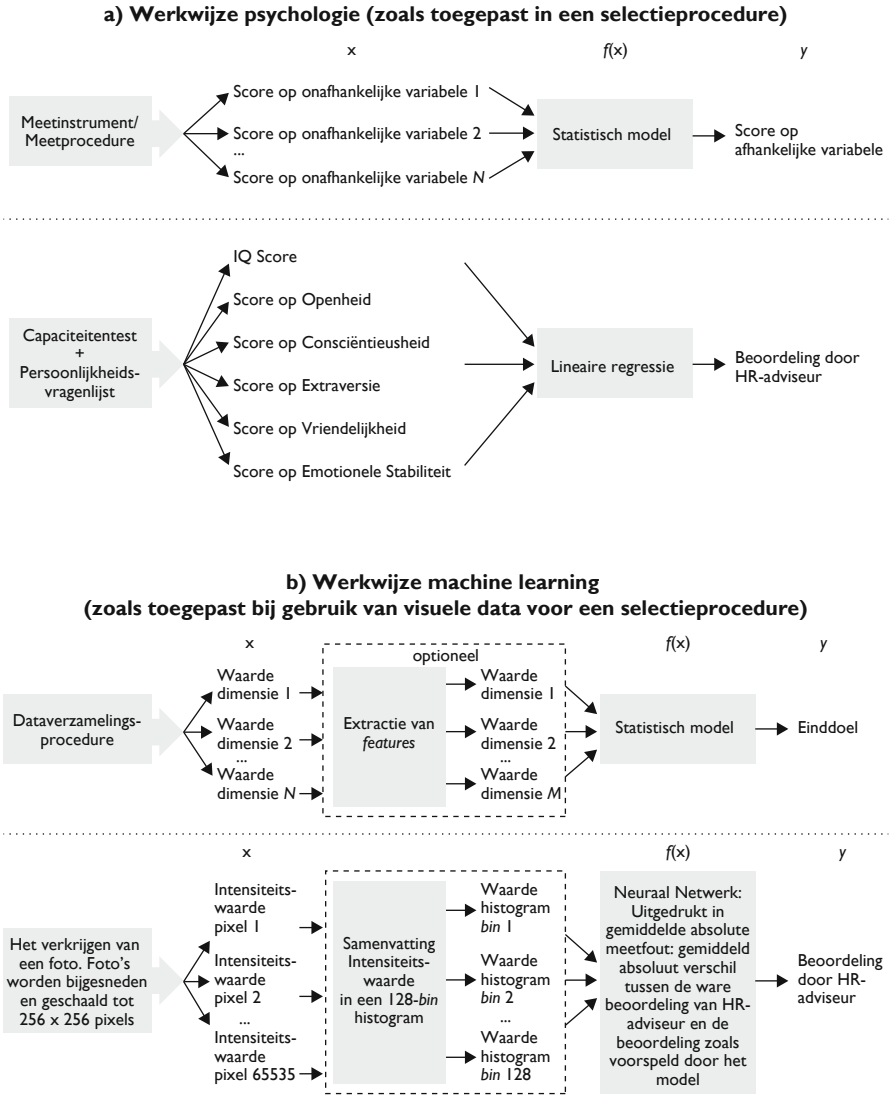
De inhoud van deze bijdrage is gebaseerd op de uitkomsten van het Erasmus + consortium 'Big Data in Psychological Assessment' (www.bdpa.eu). Binnen dit consortium werken computerwetenschappers en selectiepsychologen uit wetenschap en praktijk samen. Dit heeft geresulteerd in onderwijsmaterialen die ingezet kunnen worden in opleidingen en trainingen voor selectieadviseurs. In dit artikel richten we ons op sociale wetenschappers en psychologen, werkzaam in de praktijk, met een niet-computerwetenschappelijke achtergrond. Om op een verantwoorde wijze gebruik te kunnen maken van algoritmes gebaseerd op machine learning bij personeelsselectie, is het voor psychologen belangrijk om kennis te hebben van de (on)mogelijkheden van het gebruik van dit type nieuwe technologie voor selectie. Alleen op die manier kunnen we als beroepsgroep optimaal gebruikmaken van de kansen die ons geboden worden in deze tijd van snelle technologische ontwikkelingen. Dit artikel beoogt daaraan bij te dragen.

Allereerst zetten we uiteen wat artificiële intelligentie, en met name algoritmes gebaseerd op machine learning, voor selectiedoeleinden nu eigenlijk behelst (vgl. Liem et al., 2018). We vergelijken de selectie met behulp van machine learning met de reguliere psychologische selectiepraktijk tijdens de verschillende fasen van het selectieproces: data verzamelen, data combineren en beslissen. Aansluitend bespreken we, op basis van een literatuurstudie, of en op welke manier er (indirecte) discriminatie kan optreden bij het gebruik van algoritmes gebaseerd op machine learning (zie bijvoorbeeld Ajunwa, Friedler, Scheidegger, & Venkatasubramanian, 2016; Barocas & Selbst, 2016; Chander, 2017; Raghavan et al., 2019; Raub, 2018). We besluiten de bijdrage met concrete aanbevelingen voor het verminderen en voorkomen van discriminatie bij het gebruik van algoritmes gebaseerd op machine learning voor selectie.

1.1 Machine learning-toepassingen voor personeelsbeslissingen

Machine learning is een techniek afkomstig uit de computerwetenschappen. Learning betreft hier het op geautomatiseerde wijze leren herkennen van patronen in data (Liem et al., 2018). Er kunnen meerdere vormen van machine learning worden onderscheiden, maar zogeheten supervised learning lijkt het best inzetbaar voor personeelsselectie. Het doel bij supervised machine learning is om een wiskundige functie te maken die het beste de relatie weergeeft tussen onafhankelijke variabelen (bijv. geschreven data, audiodata of videodata) en een afhankelijke variabele (bijv. baanprestatie). Het is belangrijk om te onderkennen dat we binnen de sociale wetenschappen de relatie tussen de onafhankelijke variabele *x*

en afhankelijke variabele *y* vaak willen aantonen (via het toetsen van hypothesen), maar dat binnen machine learning juist wordt aangenomen dat de relatie tussen *x* en *y* bestaat. De relatie tussen *x* en *y* wordt dus beschouwd als een vaststaand gegeven. Het doel is om een wiskundige vergelijking te maken die de onafhankelijke en afhankelijke variabelen zo optimaal mogelijk met elkaar verbindt. Dit lichten we toe aan de hand van Figuur 1 (gebaseerd op Liem et al., 2018).



Figuur 1 ***Werkwijze assessmentpsycholoog (a) en via machine learning (b) om tot een oordeel te komen over de geschiktheid van een sollicitant (gebaseerd op Liem et al., 2018)***

In Figuur 1a is de klassieke werkwijze van de assessmentpsycholoog weergegeven. De psycholoog gaat hypothese-toetsend te werk. Dat wil zeggen dat er een aanname is over de samenhang van de onafhankelijke variabelen (voorspellers, zoals persoonlijkheid en intelligentie) en de afhankelijke variabele (prestaties in een baan). Op basis van de bekende meta-analyse van Schmidt en Hunter (1998) gaan psychologen er bijvoorbeeld van uit dat intelligentie en consciëntieusheid goede voorspellers zijn van werkprestatie in allerlei soorten banen. Bij het beoordelen van de geschiktheid van de sollicitant zal een assessmentpsycholoog dan ook op zoek gaan naar informatie over de kenmerken van de sollicitant. De informatie (data) kan via meerdere bronnen worden verzameld, zoals via vragenlijsten, interviews en rollenspellen. De uitkomsten van deze dataverzameling worden vervolgens gecombineerd. Dat kan op een statistische manier gedaan worden (regressievergelijking door de computer) of klinisch (de psycholoog vormt zich een oordeel op basis van de gegevens). De uitkomst is een inschatting van de geschiktheid van een sollicitant voor een bepaalde baan.

Figuur 1b geeft de werkwijze via machine learning weer. Ook hier wordt ernaar gestreefd om de afhankelijke variabele (prestaties in een baan) te voorspellen via onafhankelijke variabelen (bijv. geschreven, video- en/of audiodata). Het grote verschil met de assessmentpsycholoog is echter het ontbreken van hypothesen en de betekenis van de data. Bij machine learning wordt er vooraf geen betekenis gegeven aan de data. De data kunnen bestaan uit videobeelden waarbij de sollicitant in de camera vragen beantwoordt (bijv. met gebruikmaking van de software van ontwikkelaars zoals Hirevue). De ruwe videodata worden door computerwetenschappers beschouwd als een combinatie van beeldinformatie (bijv. opeenvolgende matrices met pixel-intensiteiten) en audio-informatie (bijv. een opeenvolging van frequenties). De pixels en frequenties hebben numerieke waarden, maar geen inhoudelijke betekenis. Om tot een voorspelling van baangeschiktheid te komen wordt er bij supervised learning gezocht naar patronen in deze data, die zo goed mogelijk te koppelen zijn aan een gegeven uitkomst (in dit geval baangeschiktheid). In sommige gevallen wordt hierbij een tussenstap gedaan, waarbij de ruwe data eerst worden omgezet naar 'features', oftewel kenmerken. Dit kunnen bijvoorbeeld indicaties van gezichtsuitdrukkingen zijn, zoals glimlachen of fronsen. Een glimlach of frons wordt in dat geval bijvoorbeeld gemodelleerd als een specifieke combinatie van pixels die binnen een bepaald tijdsframe veranderen van intensiteit en kleur. Deze kenmerken worden volgens een statistisch model gecombineerd om de data zo goed mogelijk te koppelen aan de uitkomst (baangeschiktheid). Er zijn ook machine learning-technieken beschikbaar waarbij deze tussenstap niet wordt uitgevoerd, met name wanneer technieken gebaseerd op 'deep learning' worden toegepast. In dat geval worden de relaties tussen de onafhankelijke en afhankelijke variabelen direct gebaseerd op de ruwe data, en is de aanname dat de leermodellen zelf de meest accurate tussenrepresentaties van de data kunnen leren. Deze technieken doen het vaak 'beter' (betere fit tussen x en y, waarbij de verklaarde variantie zo dicht mogelijk bij 1.0 komt), maar de resultaten zijn veel minder goed te interpreteren door mensen (wat is de betekenis van de data waarop het model is gebaseerd, en lijken tussenstappen op stappen die mensen zouden maken?). Omdat de resultaten veel minder goed te interpreteren

zijn voor mensen, wordt door Liem en collega's (2018) beargumenteerd dat deep learning technieken minder geschikt zijn voor de ontwikkeling van algoritmes voor selectie dan technieken waarbij handmatig kenmerken (features) worden gemodelleerd.

Zoals gezegd gaan machine learning-technieken ervan uit dat de onafhankelijke en afhankelijke variabelen aan elkaar gerelateerd zijn, en dus wordt er gezocht naar de best mogelijke classificatie of indeling door het algoritme (het zo goed mogelijk voorspellen van *y*; bijvoorbeeld wel/niet geschikt, of mate van geschiktheid). Het te gebruiken statistische model om tot deze voorspelling te komen kan een lineair model zijn (vergelijkbaar met de regressievergelijking uit de sociale wetenschappen), maar het kan ook een complexer niet-lineair model zijn (bijv. een neuraal netwerk, vectoren of beslisbomen; zie Liem et al., 2018). Het model waarbij de uiteindelijke parameters zijn vastgesteld, is dus doorgaans het model dat de data zo goed mogelijk weet te koppelen aan de afhankelijke variabele (best passende parameters). Het model definieert de stappen om op grond van de data en de parameters tot een voorspelling van de afhankelijke variabele te komen, en vormt hiervoor een algoritme (set van instructies die een machine kan uitvoeren). Om tot dit algoritme te komen, wordt een leerproces doorlopen (learning) dat uit twee fases bestaat: de trainingsfase en de testfase. De dataset wordt daarvoor in twee delen gesplitst. De ene subset wordt gebruikt voor de trainingsfase, waarin in iteratieve stappen de optimale parameters voor het model worden berekend. Deze optimale parameters worden opgenomen in het algoritme. De andere subset wordt gebruikt voor de testfase, waarbij getest wordt of het algoritme ook op datapunten die niet gebruikt zijn tijdens het leerproces tot een goede voorspelling komt. Er wordt in de testfase dus gekeken of het algoritme in staat is om te generaliseren. Als ook daar een optimale fit wordt bereikt (de voorspelling van *y* op basis van *x* is zo accuraat mogelijk), dan wordt geconcludeerd dat het leren door de computer is gelukt. De wiskundige vergelijking die de uitkomst is van het machine learning-proces, hoeft geen theoretische of interpreteerbare betekenis te hebben. De uitkomst is slechts gericht op optimalisatie van de wiskundige vergelijking gebaseerd op de data. Dit is een belangrijk verschil in werkwijze tussen sociale wetenschappers (op zoek naar betekenis in de data, hypothese-toetsend) en computerwetenschappers (een zo goed mogelijke wiskundige vergelijking maken, ongeacht de betekenis van de data).

Een belangrijke (impliciete) aanname bij machine learning-technieken is hiermee dat de afhankelijke variabele (y) objectief correct is, betrouwbaar en valide. De afhankelijke variabele wordt bij machine learning dan ook aangeduid als 'ground truth', waartegen de accuratesse van de voorspellingen vanuit het model wordt getoetst. Dit is een groot verschil met de werkwijze van assessmentpsychologen. Zij zullen er doorgaans van uitgaan dat de betrouwbaarheid en validiteit van zowel de onafhankelijke variabelen (zoals intelligentie gemeten met een test) als van de afhankelijke variabele (het criterium, zoals baanprestatie) niet perfect zijn.

Samenvattend kan worden gesteld dat machine learning-specialisten zich richten op het verbeteren van het mathematische model (het learning-gedeelte) om een zo optimaal mogelijk fit tussen *x* en *y* te realiseren (van pixels naar kenmerken naar baangeschiktheid; zie Figuur 1b). Psychologen, daarentegen, richten zich

meer op het aantonen van een relatie tussen de onafhankelijke en de afhankelijke variabelen (van data over persoonlijkheid en intelligentie naar baangeschiktheid; zie Figuur 1a) en het verklaren van deze (mogelijke) samenhang.

1.2 Bias en discriminatie bij het gebruik van machine learning voor selectie

In de voorgaande sectie hebben we toegelicht welke werkwijze wordt gebruikt wanneer via machine learning-toepassingen tot een aanbeveling wordt gekomen over de geschiktheid van een kandidaat voor een bepaalde baan. In dit deel bespreken we of en op welke manier er bias en (indirecte) discriminatie kunnen optreden bij het gebruik van algoritmes gebaseerd op machine learning. Ter verduidelijking beschrijven we hier eerst kort wat we bedoelen met discriminatie, eerlijkheid, adverse impact en bias.

Onder discriminatie verstaan we het weigeren van sollicitanten op basis van hun groepslidmaatschap. Indirecte discriminatie verwijst naar (onbedoelde) verschillen in selectie-uitkomsten voor verschillende groepen. Volgens artikel 1 van de Nederlandse grondwet is discriminatie op basis van ras, sekse, leeftijd, religie of seksuele oriëntatie niet toegestaan. Wettelijke bescherming is vooral gericht op groepen die ondergerepresenteerd zijn op de arbeidsmarkt (minderheidsgroepen, bijvoorbeeld migranten, vrouwen). Er zijn allerlei manieren om te definiëren of een selectie-uitkomst eerlijk is. Een selectieprocedure kan als eerlijk worden beschouwd als iedereen op gelijke wijze wordt behandeld (gelijke behandeling) of als de uitkomsten voor verschillende groepen vergelijkbaar zijn (gelijke uitkomsten). In lijn met de Society for Industrial and Organizational Psychology (SIOP, 2018) stellen we dat eerlijkheid in de eerste plaats een sociaal concept is.

Adverse impact (ook wel 'disparate impact' genoemd) verwijst naar een substantieel verschil in selectie-uitkomsten voor sollicitanten uit een minderheidsgroep vergeleken met sollicitanten uit de meerderheidsgroep (uitgedrukt in selectieratio's voor beide groepen). De selectieratio van sollicitanten uit de minderheidsgroep mag niet te veel afwijken van de selectieratio van sollicitanten uit de meerderheidsgroep (minimaal 80% ofwel de 4/5-regel). Deze 4/5-regel is een heuristiek die gebruikt wordt (in met name de Verenigde Staten) om een juridische casus op te bouwen. Als niet voldaan wordt aan de 4/5-regel, dan is er voldoende grond voor een rechtszaak. Adverse impact kan verschillende oorzaken hebben, waaronder bias (vertekening).

We onderscheiden twee soorten bias in selectie: predictieve bias en 'measurement bias'. Er is sprake van predictieve bias als er systematische verschillen zijn in prestatie op het criterium (bijv. baanprestatie), terwijl de gebruikte selectie-instrumenten (predictoren) gelijke scores laten zien. Er is sprake van measurement bias als door irrelevante oorzaken er sprake is van variatie in testscores (bijv. variaties in scores op een rekentest vanwege verschillen in taalvaardigheid, of variaties in interviewbeoordeling vanwege vooroordelen bij een menselijke interviewer).

Verschillende aanbieders van geautomatiseerde methoden voor selectie, zoals Harver, Hirevue, Pymetrics en Seedlink, maken gebruik van de onder 2 beschreven basisprincipes vanuit machine learning (Raghavan et al., 2019; Woods, Ahmed, Nikolaou, Costa, & Anderson 2019). Bij de toepassingen van Hirevue, bijvoorbeeld,

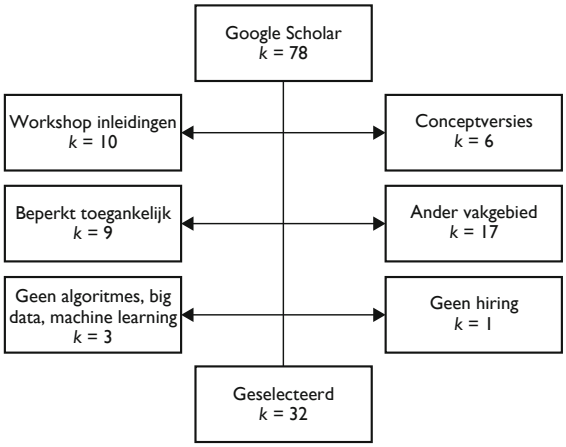
worden antwoorden van sollicitanten op interviewvragen automatisch geanalyseerd via een machine learning-algoritme. Het slim en op mathematisch geavanceerde wijze beoordelen van data kan tot betere personeelsbeslissingen leiden, met name tijdens het combineren van gegevens (Chamorro-Premuzic, Akhtar, Winsborough, & Sherman, 2017; Kuncel, Klieger, Connelly, & Ones, 2013). Bovendien spelen persoonlijke voorkeuren geen rol bij de totstandkoming van de aanbevelingen; de werkwijze is immers gelijk voor elke sollicitant. Tot slot kan een computer op een veel fijnmaziger niveau patronen herkennen, vergeleken met menselijke beoordelaars. Het is dus niet verwonderlijk dat aanbieders van machine learning-toepassingen voor selectie zich erop laten voorstaan dat bias en discriminatie minder kans krijgen bij deze toepassingen (Raghavan et al., 2019). De algoritmes die gebruikt worden, zijn echter vaak bedrijfsgeheim, waardoor het voor zowel gebruikers als wetenschappers moeilijk, zo niet onmogelijk, is om deze uitspraken te toetsen (Ajunwa et al., 2016; Barocas & Selbst, 2016; Chander, 2017; Liem et al., 2018; Raghavan et al., 2019). Bovendien moeten we constateren dat de data waarop machine learning wordt toegepast, via het ontwikkelen van een zo optimaal mogelijk algoritme, verzameld zijn door mensen. De afhankelijke variabele kan gebiast zijn vanwege discriminatie op het werk (bijv. negatievere beoordelingen, minder vaak promotie, of een lager salaris van medewerkers uit minderheidsgroepen). Daarnaast kunnen de data die gebruikt worden om het algoritme te ontwikkelen bias bevatten, bijvoorbeeld door onderrepresentatie van verschillende groepen sollicitanten, of door verschillen in meetfouten bij groepen (minder accurate voorspellingen bij minderheidsgroepen). Deze groepsverschillen in de data kunnen dan opgepikt en gereproduceerd worden bij machine learning-toepassingen.

Om inzicht te krijgen in of en hoe bias en (indirecte) discriminatie een rol kunnen spelen bij het gebruik van machine learning-algoritmes voor personeelsbeslissingen, hebben we een literatuurstudie uitgevoerd waarin wetenschappelijke artikelen over deze thematiek zijn geïntegreerd. Het verkrijgen van een overzicht van wetenschappelijke kennis over de (on)mogelijkheden van machine learning-algoritmes om bias en discriminatie te verminderen bij personeelsbeslissingen is noodzakelijk om deze algoritmes te kunnen verbeteren en om bij te kunnen dragen aan eerlijke, inclusieve selectieprocedures.

2 Methode

Via twee stappen verzamelden we relevante literatuur voor dit literatuuroverzicht. De eerste stap bestond uit de selectie van artikelen vanuit zoekopdrachten in een literatuurdatabase (Google Scholar). In de tweede stap voegden we artikelen toe via referenties in de artikelen vanuit stap 1. De gebruikte zoektermen waren 'algorithm', 'big data' en 'machine learning' om ervoor te zorgen dat de geselecteerde artikelen gericht waren op geavanceerde technologie. Vervolgens werd de zoekterm 'hiring' toegevoegd om artikelen te selecteren gericht op werving en selectie. Tot slot voegden we de zoekterm 'bias' toe. In mei 2017 voerden we voor de eerste keer deze zoekopdracht uit met de volgende syntax: 'algorithm', 'hiring', 'bias',

'machine-learning' OR 'machine learning', 'big data'. Dit resulteerde in 218 resultaten, waarbij we uiteindelijk 22 artikelen includeerden (De Neijs, 2017). Bij het opnieuw uitvoeren van de zoekopdracht in 2019 voor dit artikel leverde dezelfde zoekopdracht 1730 resultaten op (periode 2016-2019). Om tot een betere selectie te kunnen komen voegden we enkele zoektermen toe, namelijk 'adverse impact' en 'fairness'. Dit leidde tot de volgende zoek-syntax die is gebruikt voor dit artikel: 'algorithm', 'hiring', 'bias', 'machine-learning' OR 'machine learning', 'big data', 'adverse impact', 'fairness'.
Naast de syntax is het mogelijk om via Google Scholar op een meer uitgebreide manier te zoeken. In de 'advanced search option' selecteerden we de optie 'anywhere in article'. Dit betekent dat de zoektermen op zijn minst één keer moesten voorkomen in een artikel. Vanwege de recente, snelle technologische ontwikkelingen besloten we om alleen artikelen te selecteren vanaf 2016. Tot slot werden patenten en referenties uitgesloten van selectie. Deze procedure leverde 78 resultaten op (zie Figuur 2).



Figuur 2 ***Overzicht van zoekresultaten literatuurstudie***
NB. Aan de selectie voegden we nog 9 artikelen toe die niet direct voortkwamen uit de zoekopdracht (k = 41; zie de methodesectie voor een toelichting).

Van deze resultaten screenden we de titel en de samenvatting. Van sommige artikelen was het meteen duidelijk dat zij afkomstig waren uit een ander vakgebied, zoals natuurwetenschappen of bouwkunde; deze werden niet meegenomen (*k* = 17). Tien resultaten werden uitgesloten omdat het ging om inleidingen op workshops of seminars (*k* = 10). Nog eens zes 'dubbele' artikelen werden uitgesloten omdat dit conceptversies waren (gepubliceerd op Arxiv.org) van later gepubliceerde artikelen die ook in de zoekresultaten werden gevonden (*k* = 6). Vanwege beperkte toegankelijkheid besloten we ook een aantal artikelen en/of boekhoofdstukken niet mee te nemen, bijvoorbeeld vanwege betaling of beschikbaarheid van het boek (*k* = 9). Een klein aantal artikelen had algoritmes, big data of machine

learning (k = 4) niet als thema, één artikel had werving en/of selectie (hiring) niet als thema, waarbij het thema slechts kort werd genoemd, bijvoorbeeld als een mogelijke toepassing voor algoritmes. Dit type artikelen namen we ook niet mee, vanwege gebrek aan diepgang en focus op de relevante onderwerpen. Dit betekent dat er 32 artikelen overbleven die gebruikt zijn voor deze review. Een deel van de zoekresultaten overlapte met de resultaten van de eerste zoekopdracht in 2017, maar enkele zoekresultaten uit 2017 ontbraken vanwege de aanpassing van de syntax. Sommige van deze eerdere artikelen voldeden wel aan de inclusiecriteria (artikelen met als thema machine learning, bias en hiring). Vanwege de inhoud van deze artikelen besloten we deze ook te includeren (k = 8). Tot slot werden we door een anonieme reviewer gewezen op een artikel dat ook geïncludeerd kon worden (k = 1). De in totaal 41 geselecteerde artikelen dateren van de periode tussen 2010 en 2019. Op basis van de inhoud van deze 41 artikelen identificeerden we thema's (samengevat in Tabel 1). De resultaten worden hieronder beschreven. Omdat artikelen uit verschillende vakgebieden afkomstig waren en soms zowel een beschrijvend als een empirisch deel bevatten, hebben we voor de leesbaarheid van de resultaten besloten om de uitkomsten van de literatuurstudie te structureren aan de hand van het algoritme-ontwikkelingsproces. Dit proces is weergegeven in Figuur 3.

Tabel 1 ***Overzicht van de geselecteerde manuscripten***

Type artikel	**Thema**	**Aantal artikelen**	**Voorbeeld referenties**
Beschrijvend	Introductie en uitleg over betekenis algoritmes gebaseerd op machine learning; identificatie van risico's	18	Barocas & Selbst (2016); MacCarthy (2017)
	Ethiek en wetgeving	8	Chander (2017), Hacker (2018), Kim (2016)
Empirisch	Ontwikkeling en/of toepassing van algoritme voor selectiebeslissingen; 'ont-biasen' van algoritmes	8	Žliobaitė en Custers (2016)
	Gebruik van op algoritmes gebaseerde aanbevelingen door menselijke beslissers	3	Ebrahimi (2018)
	Reacties van sollicitanten op eerlijkheid geautomatiseerde selectiemethoden	2	Gonzalez et al. (2019)
	Evaluatie, op basis van publieke informatie (websites), van aanbieders van op algoritmes gebaseerde selectiemethoden	2	Raghavan et al. (2019)

NB. Met de aanduiding 'empirisch' bedoelen we dat de auteurs gebruikmaken van kwantitatieve of kwalitatieve data om hun onderzoeksvraag te beantwoorden of om voorbeelden te illustreren. Slechts vier artikelen toetsen hypotheses. Voor een uitleg over het verschil in werkwijzen tussen computer- en sociale wetenschappen verwijzen we naar König et al. (2020) en Liem et al. (2018).

3 Resultaten literatuurstudie naar discriminatie en bias in machine learning-algoritmes voor werving en selectie

In het eerste deel van dit artikel hebben we algoritmes, gebaseerd op machine learning voor selectievraagstukken, geïntroduceerd en hebben we laten zien hoe deze data-gedreven werkwijze overeenkomt en verschilt van de werkwijze van de selectiepsycholoog. Oorspronkelijk werden algoritmes voor personeelsselectie (deels) ontwikkeld om bias en discriminatie te verminderen of zelfs helemaal uit te sluiten (Barocas & Selbst, 2016; Chander, 2017; MacCarthy, 2017 Raghavan et al., 2019; Savage & Bales, 2016; Tambe, Cappelli, & Yakubovitch, 2019). De argumentatie hierbij is dat een algoritme logische beslissingen neemt, gebaseerd op feiten, getallen en data, en niet op bewuste en deels onbewuste processen bij menselijke beoordelaars (Ajunwa et al., 2016; Crawford & Calo, 2016). Vanuit die redenering bevat een algoritme de belofte van objectieve en accurate beslissingen. Een nadere beschouwing echter toont dat de data op basis waarvan een algoritme wordt gemaakt, niet (geheel) gebaseerd hoeven te zijn op objectieve gegevens. Een prestatiebeoordeling door een supervisor, bijvoorbeeld, is een vorm van subjectieve data. Daarnaast is het belangrijk om te realiseren dat het menselijke (subjectieve) aspect voor het nemen van een selectiebeslissing wordt verschoven van de beoordelaar naar de machine learning-specialist die het algoritme ontwikkelt, op basis van bestaande data (Kim, 2016). Er is dan ook steeds meer aandacht voor bias en discriminatie bij selectie gebaseerd op machine learning-algoritmes (bijv. Chander, 2017; Crawford & Calo, 2016; Hacker, 2018; Kim, 2016; Lipton, McAuley, & Chouldechova, 2018; Raghavan et al., 2019; Raub, 2018).

3.1 Directe discriminatie

Net als binnen de selectiepsychologische literatuur, wordt bij onderzoek naar machine learning-algoritmes onderscheid gemaakt tussen directe discriminatie (gedefinieerd als het bewust uitsluiten van mensen op basis van groepslidmaatschap, ongeacht hun kwalificaties; SIOP, 2018) en indirecte discriminatie (selectieuitkomsten gebaseerd op het algoritme leiden onbedoeld tot verschillen in selectiekansen voor leden van de minderheids- en meerderheidsgroepen). In de meeste gevallen is er geen eenduidige evidentie voor directe discriminatie (Barocas, Bradley, Honavar, & Provost, 2017; Hirsch, 2016).

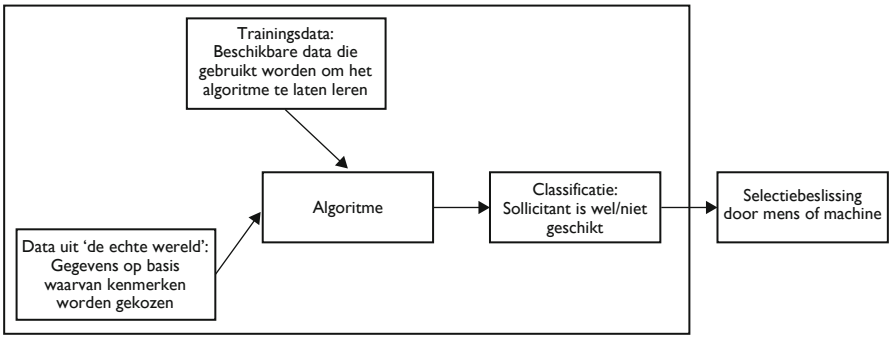
Auteurs van verschillende geselecteerde artikelen zijn het erover eens dat directe discriminatie veel moeilijker te detecteren is in machine learning-toepassingen dan in traditionele werkwijzen om tot een selectiebeslissing te komen, omdat discriminatie nu vermomd is in eindeloze rijen (van op zichzelf betekenisloze) codes (Ajunwa et al., 2016; Barocas & Selbst, 2016; Hirsch, 2016). Barocas en Selbst (2016) gaan zelfs zover dat zij menen dat algoritmes directe discriminatie in de hand werken, omdat personen met een intentie om direct te discrimineren dit via het algoritme kunnen maskeren. Maskeren kan bijvoorbeeld worden gedaan door indirecte variabelen te kiezen die bepalen of iemand tot een minderheidsgroep behoort. Barocas en Selbst (2016) en Chander (2017) stellen dat overheden vooralsnog onvoldoende zijn toegerust om discriminatie op basis van machine learning-algoritmes te reguleren. De ontwikkelingen op dit vlak gaan echter snel (Hacker, 2018; MacCarthy, 2017).

3.2 Indirecte discriminatie

Hoewel verschillende auteurs benadrukken dat algoritmes ruimte bieden voor het verdonkeremanen van directe discriminatie (Ajunwa et al., 2016; Barocas & Selbst, 2016; Hirsch, 2016), geven dezelfde auteurs aan dat indirecte discriminatie een grotere bron van zorg is (Barocas & Selbst, 2016; Kim, 2016; Raghavan et al., 2019; Raub, 2018). Er zijn mogelijkheden om te bepalen of er bias is die kan leiden tot indirecte discriminatie vanwege de wijze waarop het algoritme is vastgesteld. Hirsch (2016) stelt dan ook dat algoritmes niet zozeer bias uitsluiten, maar dat de bias eerder 'ondergronds' gaat, waarbij de bias eerder versterkt terugkomt in het algoritme en tegelijk juist moeilijker te detecteren is. Ook wordt wel gesteld dat juist indirecte discriminatie een groter negatief effect heeft op minderheidsgroepen, uitgedrukt in adverse impact (Ajunwa et al., 2016; Mainka, 2018).

De reden dat de auteurs aangeven dat indirecte discriminatie een grotere zorg is dan directe discriminatie, is dat indirecte discriminatie veel meer voorkomt dan directe discriminatie (Barocas & Selbst, 2016; Žliobaitė & Custers, 2016). Factoren die kunnen leiden tot indirecte of onbedoelde discriminatie, kunnen makkelijk over het hoofd worden gezien door de ontwerper van een algoritme. Daar kan nog aan worden toegevoegd dat checks op bias kunnen leiden tot valse negatieven ('false negatives'; het onterecht afwijzen van geschikte kandidaten), waardoor de bias niet opgemerkt wordt (Barocas & Selbst, 2016).

De machine learning-specialist die algoritmes voor personeelsselectie maakt, kan mogelijk vele vormen van bias over het hoofd zien (Chander, 2017; Kim, 2016). Hieronder zetten we uiteen wat er mis kan gaan, op basis van de geselecteerde literatuur. Met 'misgaan' bedoelen de meeste auteurs dat een algoritme onbedoeld tot adverse impact leidt. We structureren deze beschrijving aan de hand van de verschillende fases van het algoritme-ontwikkelingsproces. Dit is weergegeven in Figuur 3. Bij deze beschrijving maken we onderscheid in bias bij het bepalen van de features (vanuit de beschikbare data uit 'de echte wereld') en in de beschikbare trainings- en testdata.



Figuur 3 Onderdelen van het algoritme-ontwikkelingsproces (gebaseerd op Chander, 2017)

3.3 Data uit 'de echte wereld' – bias bij het selecteren van data en kenmerken

Aan het begin van het machine learning-proces dat leidt tot een algoritme, wordt bepaald welke data meegenomen worden, en welke features (betekenisvolle eenheden; kenmerken) gekozen worden (zie Figuur 1b en Figuur 3). Sommige kenmerken zijn makkelijker te meten en vast te stellen dan andere (Barocas & Selbst, 2016) en sommige kenmerken kunnen zo gemeten worden dat bepaalde groepen in de maatschappij buitengesloten worden. Informatie over tijdig aflossen van leningen, bijvoorbeeld, kan niet worden vastgesteld bij personen/groepen die geen lening hebben. Crawford en Calo (2016) noemen dit de 'schaduwzones' binnen grote datasets (big data), waarbij minderheidsgroepen over het hoofd kunnen worden gezien. Van deze mensen worden minder data verzameld, waardoor zij minder worden gerepresenteerd in datasets die gebruikt worden om machine learning op toe te passen. Een berucht voorbeeld is het algoritme van Google, gemaakt om figuren op plaatjes te herkennen, gebaseerd op data van beschikbare plaatjes op het internet. Daar worden blanke mensen meer vertegenwoordigd dan mensen met een donkere huidskleur. Het algoritme was daardoor minder goed in staat om gezichten van mensen uit minderheidsgroepen te herkennen. Zo werd een Afro-Amerikaans persoon ten onrechte geclassificeerd als een 'gorilla'. Deze fout zal niet snel door een menselijke beoordelaar worden gemaakt (Raub, 2018). Bij selectiesituaties valt te denken aan kenmerken van mensen die succesvol zijn in een baan. Op basis van die kenmerken worden data gekozen om mee te nemen in het algoritme. Het is echter bekend dat mensen uit minderheidsgroepen minder vaak hoogbetaalde en zichtbare functies bekleden, vaker een tijdelijk contract hebben en vaker na korte tijd weer vertrekken (Benton, Fratzke, & Sumpton, 2014). Personen uit minderheidsgroepen worden dus minder vertegenwoordigd in data die gebruikt kunnen worden om een algoritme voor personeelsselectie te ontwikkelen. De data waarop het algoritme wordt ontwikkeld, geven dus vaak de 'status quo' weer. Terugkerend naar het voorbeeld over het terugbetalen van leningen: mensen die geen lening hebben, kunnen niet worden meegenomen in het algoritme dat ontwikkeld wordt om kredietwaardigheid te bepalen, simpelweg omdat daarover geen gegevens beschikbaar zijn. Dit heeft weer tot gevolg dat deze groep bij een volgende aanvraag ook geen lening kan krijgen, omdat zij niet als kredietwaardig worden geclassificeerd door het algoritme. Er kan dus sprake zijn van een selffulfilling prophecy.

Een andere bron voor bias bij de keuze van data en kenmerken is dat minderheden op sommige kenmerken die wel goed voorspellen, minder goed scoren dan leden van de meerderheidsgroep (Barocas & Selbst, 2016; Calders & Žliobaitė, 2013). Dit is een bekend fenomeen in de selectiepsychologie, dat ook wel het diversiteit-validiteitsdilemma wordt genoemd. Dat wil zeggen dat de meest valide predictoren van werkprestatie (bijv. intelligentie) ook de grootste groepsverschillen laten zien (etnische minderheden scoren bijvoorbeeld gemiddeld lager op intelligentietests dan de etnische meerderheid). Daarnaast kunnen sommige kenmerken op andere wijze worden gemeten dan direct via het werkelijke kenmerk. Strahilevitz (2008) geeft een voorbeeld van hoe ras soms wordt gebruikt als (zeer imperfecte) schatting van criminele geschiedenis van sollicitanten, vanwege de statistisch aangetoonde samenhang tussen ras en veroordelingen. Los van het feit dat het

om directe discriminatie gaat als deze gegevens worden gebruikt, is het bovendien een hoogst onbetrouwbare manier om van een individu vast te stellen of deze persoon een strafblad heeft.

Andere data, zoals schoolprestaties, kunnen op het eerste gezicht worden aangemerkt als kenmerken die niet tot directe discriminatie leiden. Dergelijke data kunnen echter wel tot indirecte discriminatie leiden. Dit kan zo zijn als deze kenmerken dienen als een proxy voor andere kenmerken die groepslidmaatschap bepalen. Dat wil zeggen dat dezelfde kenmerken die gebruikt worden om te voorspellen welke sollicitanten zullen excelleren in een baan, hen ook indeelt op groepslidmaatschap (Barocas & Selbst, 2016). Onderzoekers binnen het veld van de machine learning noemen dit probleem 'redundante codering', 'proxy-variabelen' of het 'redlining-effect'. Dit betekent kort gezegd dat lidmaatschap van een (beschermde) minderheidsgroep onbedoeld gecodeerd is in andere relevante data (Barocas & Selbst, 2016; Barocas et al., 2017; Žliobaitė, 2017). Onbedoeld kunnen ontwikkelaars van machine learning-algoritmes dan ook bestaande sociale ongelijkheden in stand houden, zoals verminderde baankansen voor etnische minderheden vergeleken met de meerderheid (Ajunwa et al., 2016; Barocas et al., 2017; Chander, 2017; Kim, 2016; Tambe et al., 2019).

We lichten het bovenstaande toe met twee voorbeelden (Hirsch, 2016; Žliobaitė, 2017). Het bedrijf Evoly gebruikte machine learning-algoritmes om vast te stellen hoe verschillende features (kenmerken) samenhangen met hoelang iemand bij Evoly bleef werken. Het bedrijf ontdekte zo dat medewerkers die binnen vijf mijl van het werk woonden ook 20% langer bij hen bleven werken. Het bedrijf besloot dit kenmerk echter niet te gebruiken, omdat het discrimineerde ten nadele van sollicitanten uit achterstandswijken. Deze wijken waren verder van het centrum gelegen, waar het bedrijf gevestigd was. Een andere proxy-variabele is de school of universiteit waar een sollicitant is afgestudeerd. In de Verenigde Staten is, vergeleken met Nederland, het onderwijs sterker gesegregeerd. De school of universiteit op zichzelf kan dus al een duiding zijn voor het ras van de sollicitant (Barocas & Selbst, 2016).

Verschillende auteurs (Ajunwa et al., 2016; Barocas & Selbst, 2016; Barocas et al., 2017; Žliobaitė & Custers, 2016) geven aan dat de hiervoor beschreven bias de reden is dat het niet zo eenvoudig is om simpelweg data (bijv. van etniciteit of geslacht) te verwijderen uit de dataset, om discriminatie te voorkomen. Bij het verwijderen van gegevens over etniciteit, bijvoorbeeld, is het nog steeds mogelijk om dit te coderen via proxy-kenmerken, zoals postcodegebied (Žliobaitė, 2017) of de wijk waarin iemand woont (Chander, 2017; Hirsch, 2016).

Bias kan zich ook op andere manieren manifesteren, door de manier waarop kenmerken worden gekozen, bijvoorbeeld de keuze van het gebruik van data op basis van cv's, motivatiebrieven of videosollicitaties. Savage en Bales (2016) benoemen dat de keuze van trefwoorden (keywords) in motivatiebrieven en cv's ten onrechte kan leiden tot het afwijzen van geschikte kandidaten, omdat zij niet de juiste trefwoorden hebben gebruikt. Savage en Bales (2016) betogen dat, gemiddeld genomen, het niet noemen van de juiste trefwoorden de kans op een uitnodiging voor een selectie-interview verlaagt met 75%, zelfs als de sollicitant hooggekwalificeerd is. Sollicitanten uit minderheidsgroepen gebruiken mogelijk andere trefwoorden, en

gebruiken de voor het algoritme relevante trefwoorden minder vaak dan sollicitanten uit de meerderheidsgroep. Hierbij kan bijvoorbeeld worden gedacht aan oudere sollicitanten die minder ervaring hebben met online solliciteren of met het spelen van een serious game voor selectie (Barocas & Selbst, 2016; Savage & Bales, 2016), of personen die solliciteren in een andere taal dan hun eigen moedertaal.
Op basis van voorgaande voorbeelden kunnen we stellen dat het niet meenemen van classificatiegegevens (zoals naam of leeftijd) bij de ontwikkeling van algoritmes onvoldoende bescherming zal bieden voor sollicitanten uit minderheidsgroepen, vanwege de aanwezigheid van proxy-kenmerken (het hiervoor genoemde redlining-effect; Ajunwa et al., 2016; Žliobaitė, 2017; Kim, 2016). Tot slot zorgt het verwijderen van sensitieve data, zoals over etniciteit, ervoor dat discriminatie op basis van etniciteit ook niet meer kan worden vastgesteld (Žliobaitė & Custers, 2016).

3.4 Bias in de trainingsdata en in het te voorspellen einddoel

Naast de keuze van de data en de selectie van kenmerken kan de bron van de data die gebruikt wordt om het algoritme te trainen (zie Figuur 3), zoals historische data van selectieprocedures binnen een organisatie, een van de grootste oorzaken van bias zijn (Ajunwa et al., 2016; Barocas & Selbst, 2016; Barocas et al., 2017; Chander, 2017; Hirsch, 2016; Žliobaitė & Custers, 2016). Als de data waarop learning plaatsvindt al bias bevatten, dan zal het machine learning-proces resulteren in een algoritme dat deze bias heeft 'geleerd' en geïncorporeerd. Dit probleem speelt vooral als een organisatie alleen de eigen historische data gebruikt om een algoritme te ontwikkelen voor selectie. Dit illustreren we met een voorbeeld: Een organisatie wil graag een algoritme ontwikkelen dat bias ten aanzien van geslacht voorkomt. Ze gebruikt hiervoor beschikbare data, zoals de huidige prestatiebeoordelingen van recent aangenomen sollicitanten. Het algoritme leert dan op basis van data die gebaseerd zijn op eerdere beslissingen van recruiters (aangenomen sollicitanten). Zelfs als er geen overduidelijke bias is in het huidige medewerkersbestand, is wel te verwachten dat het gebruik van historische data uit het bedrijf zal leiden tot selectie van sollicitanten die lijken op de op dat moment werkzame medewerkers, waardoor eventueel al lang bestaande sociale ongelijkheden in stand worden gehouden en mogelijk worden vergroot. Het is dus lastig om op deze manier als organisatie een verandering te bewerkstelligen in het type kandidaat dat wordt aangenomen, zelfs als dat een expliciete beleidskeuze is (bijv. vanuit diversiteitsoverwegingen).
Barocas en Selbst (2016) geven een voorbeeld hiervan. Het social media-platform LinkedIn biedt de mogelijkheid om bij sommige kandidaten een organisatie aan te bevelen. Deze aanbeveling is gebaseerd op een algoritme dat is ontwikkeld op basis van historische gedragsdata van de werkgever. Als een werkgever in het verleden de voorkeur gaf aan een bepaald type kandidaat, dan zal op basis van het algoritme een aanbeveling worden gedaan van nieuwe kandidaten die vergelijkbaar zijn. Als werkgevers, bewust of onbewust, een bias hebben ten nadele van bepaalde groepen en hun gedrag op LinkedIn negeren, door bijvoorbeeld aanbevelingssuggesties voor deze kandidaten te negeren, dan wordt hen vervolgens ook weer minder aangeboden. Dit vermindert weer de kans van leden van (beschermde) minderheidsgroepen om gezien te worden en om uitgenodigd te worden om te solliciteren.

3.5 Ondervertegenwoordiging in de trainingsdata

Naast data die in zichzelf gebiast zijn, kan ook een scheve verdeling binnen de dataset een bron van bias zijn. Ondervertegenwoordiging van een groep in een dataset kan ervoor zorgen dat het algoritme classificeert of wegingen maakt die gebaseerd zijn op een (te) kleine steekproefgrootte van de ondervertegenwoordigde groepsleden, dat wil zeggen een steekproef die niet ten volle de vaardigheden dan wel variabiliteit in vaardigheden van deze groepsleden weergeeft (Barocas & Selbst, 2016; Raghavan et al., 2019). Als bepaalde groepen totaal afwezig zijn in de beschikbare data, dan lopen deze groepen het risico om volledig genegeerd te worden in het uiteindelijke algoritme gebaseerd op machine learning. Naast onderrepresentatie kan juist ook overrepresentatie tot adverse impact leiden (Barocas & Selbst, 2016). Managers en specialisten die in opdracht van managers het algoritme ontwikkelen, kunnen een relatief groot deel van de tijd en aandacht laten uitgaan naar het perfectioneren van het algoritme gebaseerd op overgepresenteerde groepen. Als er dus een fout in het algoritme zit (bijv. door onjuiste codering van een deel van de data) of als het algoritme verder geoptimaliseerd kan worden, dan valt dat eerder op voor overgepresenteerde groepen in de dataset.

Het slecht bijhouden van historische data die gebruikt worden voor de ontwikkeling van een algoritme voor selectie, kan ook een bron van bias zijn. Als organisaties hun eigen beschikbare data gebruiken, dan is het wel essentieel dat deze data accuraat en zo compleet mogelijk zijn. Er kunnen immers fouten in zitten, en fouten in data van ondervertegenwoordigde groepen hebben een grotere invloed op het machine learning-resultaat vergeleken met het gemiddelde van de populatie in de database (Ajunwa et al., 2016; Barocas & Selbst, 2016; Kim, 2016). Het kan lastig zijn om fouten te detecteren en, zoals hierboven wordt geïllustreerd, kunnen fouten mogelijk eerder worden opgemerkt voor data die zijn gebaseerd op de meerderheid, waardoor sociale verschillen groter kunnen worden (Barocas & Selbst, 2016; Kim, 2016).

3.6 Classificatie en manieren om bias en adverse impact te verminderen

Op basis van de gekozen kenmerken en training komt het algoritme tot een classificatie: de sollicitant is wel of niet geschikt (zie Figuur 3). Er zijn verschillende manieren om met mogelijke (indirecte) discriminatie op basis van de classificatie door een algoritme om te gaan. Het is daarbij belangrijk om te definiëren wanneer er sprake is van onterechte discriminatie (Lipton et al., 2018; Hacker, 2018; MacCarthy, 2017). Met andere woorden: wanneer is de selectie-uitkomst op basis van de classificatie door een algoritme oneerlijk? Er zijn verschillende manieren om eerlijkheid vast te stellen, bijvoorbeeld via gelijke behandeling, of bij gelijke uitkomsten. In overeenstemming met de SIOP-richtlijnen (2018) die gangbaar zijn binnen de selectiepsychologie, stellen ook de auteurs uit verschillende disciplines in deze review vast dat eerlijkheid een sociaal concept is (MacCarthy, 2017). De meeste auteurs stellen vast dat een algoritme eerlijk is als er geen sprake is van adverse impact (bijv. Raub, 2018). Zoals beschreven houdt adverse impact in dat de selectieratio van sollicitanten uit de minderheidsgroep niet te veel mag afwijken van de selectieratio van sollicitanten uit de meerderheidsgroep (moet minimaal 80% zijn). Aanbieders van algoritmes voor selectie, zoals Pymetrics en

Hirevue, richten zich bij het 'ont-biasen' van hun algoritmes dus op het voldoen aan deze 80%-regel (Raub, 2018; Raghavan et al., 2019).

Initiatieven en aanbevelingen door onderzoekers en ontwikkelaars zijn dus veelal gericht op het voorkomen van adverse impact bij personeelsselectie. Om adverse impact te voorkomen, worden zogeheten disparate learning-processen (DLP's) ingezet om het algoritme zó te laten leren, dat het geen kenmerken meer meeneemt die direct of indirect (via correlaties) gerelateerd zijn aan groepslidmaatschap (bijv. geslacht of ras; Lipton et al., 2018; Raghavan et al., 2019; Raub, 2018). Disparate learning-processen zijn erop gericht om beslisregels te maken die de selectie-uitkomsten voor de groepen ongeveer gelijktrekken (gelijke selectieratio's), zonder disparate treatment (positieve discriminatie) op het individuele niveau. DLP's leggen dus als het ware beperkingen op aan het algoritme dat wordt ontwikkeld, zodat adverse impact bij de uitkomst kan worden vermeden. In de praktijk betekent dit vooral dat features die correleren met beschermde groepskenmerken (zoals geslacht, ras of seksuele oriëntatie) worden verwijderd uit het model, net zolang totdat de uitkomsten (wettelijk) acceptabel zijn. Dit proces wordt ook wel algoritmische de-biasing genoemd (Bogen, Rieke, & Ahmed, 2019; Raghavan et al., 2019). Het proces wordt als geslaagd beschouwd als de uitkomst geen adverse impact meer laat zien. Als er toch sprake blijft van adverse impact, dan treedt een stappenproces in werking dat gelijk is voor allerlei typen selectie-instrumenten (met name in de Verenigde Staten; MacCarthy, 2017; Raub, 2018). De gebruiker moet aantonen dat het algoritme valide is (baangerelateerd) en dat er geen goed alternatief bestaat voor het gebruik van het algoritme (EEOC, 1978). We benadrukken dat het voor ontwikkelaars vaak expliciet het doel is om eerlijke en bias-vrije selectie-instrumenten te bieden. Dat is zelfs vaak de reden van oprichting geweest (zie bijv. Raub, 2018). Bij het 'ont-biasen' van algoritmes via het voldoen aan de 80%-regel zijn echter verschillende kanttekeningen te plaatsen, die niet altijd eenvoudig te verhelpen zijn. Ten eerste ontbreken vaak gegevens over groepslidmaatschap (bijv. seksuele oriëntatie). In sommige landen is registratie van beschermde groepskenmerken bovendien verboden. Discriminatie en bias zijn dan dus niet goed inzichtelijk te maken. Ten tweede is adverse impact niet de enige indicatie van bias. De kwaliteit van de uitkomsten hangt immers af van de kwaliteit van de data (validiteit, representativiteit). Voldoen aan de 80%-regel geeft nog geen antwoord op de vraag waarom er verschillen bestaan. Door het ontbreken van adverse impact te beschouwen als het ontbreken van bias wordt de problematiek onterecht gesimplificeerd en mogelijk ook gebagatelliseerd (Raghavan et al., 2019).

Verschillende auteurs geven aan dat er steeds meer juridische discussie is over 'algoritmische positieve actie' (algorithmic affirmative action; Hacker, 2018; Lipton et al., 2018; MacCarthy, 2017; Mainka, 2018; Raghavan et al., 2019; Raub, 2018; Tambe et al., 2019). Zij pleiten voor het instellen van kwaliteitsaudits en het uitgeven van licenties voor algoritmes (Barocas & Selbst, 2016, Kim, 2016; Raub, 2018). Ook in Nederland wordt door het ministerie van Sociale Zaken en Werkgelegenheid wetgeving, en de handhaving daarvan, voorbereid, die gericht is op de vermindering van discriminatie bij werving en selectie, waartoe ook selectie via algoritmes wordt gerekend (TNO, 2018).

3.7 Percepties van eerlijkheid

Tot slot bespreken enkele auteurs percepties van eerlijkheid van selectiebeslissingen op basis van classificaties door machine learning-algoritmes (zie Tabel 1 en Figuur 3), bij beslissers (bijv. Ebrahimi, 2018) en/of bij sollicitanten (Gonzalez, Capman, Oswald, Theys, & Tomczak, 2019; Suen, Chen, & Lu, 2019). Uit een serie studies van Ebrahimi (2018) blijkt dat percepties van eerlijkheid bij beslissers afhangen van de transparantie van de demografische kenmerken waarop de aanbeveling (classificatie) tot stand is gekomen. Bij meer transparantie over het gebruik van demografische gegevens wordt een mogelijk discriminatoire aanbeveling eerder herkend door beslissers. Sollicitanten zouden geautomatiseerde selectiebeslissingen (op basis van video-interviews) even eerlijk vinden als beslissingen genomen door een menselijke beoordelaar (Suen et al., 2019). Gonzalez et al. (2019) concluderen op basis van een panelstudie (N = 192) dat sollicitanten een selectie-beslissing door een geautomatiseerd systeem juist minder eerlijk vinden dan een selectiebeslissing door een mens.

4 Discussie en aanbevelingen

We hebben beschreven wat machine learning voor selectiedoeleinden inhoudt en hoe deze werkwijze overeenkomt en verschilt met de gebruikelijke werkwijze van assessmentpsychologen. Algoritmes kunnen mogelijkerwijs leiden tot minder (indirecte) discriminatie in selectie. Dat is een van de redenen waarom de ontwikkeling van algoritmes voor selectie zo snel is gegaan. Op basis van het hier beschreven literatuuroverzicht concluderen we dat het op dit moment nog onvoldoende mogelijk is om vast te stellen of deze belofte ook wordt ingelost. Bovendien wordt in de maatschappij in toenemende mate erkend dat er risico's kleven aan het gebruik van machine learning-toepassingen voor selectie (Harwell, 2019). Mogelijk hebben wij op basis van de door ons gebruikte zoektermen relevante artikelen gemist. Toekomstig onderzoek kan andere zoektermen gebruiken, bijvoorbeeld gericht op specifieke selectie-instrumenten zoals het interview of persoonlijkheidsvragenlijsten. De toenemende aandacht voor het onderwerp, getuige de grote groei van het aantal publicaties binnen de computerwetenschappen, arbeids- en organisatiepsychologie, rechtsgeleerdheid en politicologie, stemt in ieder geval hoopvol. Raghavan en collega's (2019) geven vijf aanbevelingen voor beleidsmakers. Deze aanbevelingen vatten de uitkomsten van de review samen en zijn ook relevant voor gebruikers en ontwikkelaars/verkopers van software. Daarom nemen wij de richting van deze aanbevelingen hieronder over.

Ten eerste onderstrepen we het niet te onderschatten belang van transparantie, die nodig is om de systemen en algoritmes te begrijpen (vgl. Ajunwa et al., 2016; Barocas & Selbst, 2016; Liem et al., 2018; Raub, 2018). Dit is vaak niet eenvoudig, omdat aanbieders van software niet geneigd zijn om informatie te delen; het algoritme is immers hun bedrijfsgeheim. De transparantie is wel noodzakelijk om tot beter beleid te kunnen komen (vgl. Barocas & Selbst, 2016).

Ten tweede is adverse impact niet de enige indicatie van bias. Ook in de data en tijdens de ontwikkeling van het algoritme kan bias optreden, zoals we in het voor-

gaande stuk hebben uiteengezet. Verkopers, gebruikers en beleidsmakers moeten dit proces ook monitoren en hun inspanningen niet alleen richten op de uitkomst (wel/geen adverse impact), maar bijvoorbeeld ook op het onderzoeken van differentiële predictie en verschillen in errorvariantie binnen groepen.
Ten derde kunnen statistische indicatoren, zoals adverse impact en differentiële predictie, alleen goed worden vastgesteld als de beschikbare data representatief zijn voor de verschillende groepen sollicitanten. Bovendien wordt met statistische indicatoren niet bepaald in hoeverre individuele predictoren ook geschikt zijn om te gebruiken. Data uit gestructureerde interviews zijn doorgaans bijvoorbeeld betere predictoren dan data op basis van ongestructureerde interviews. Tot slot ontbreekt er vaak informatie over groepskenmerken, waardoor het moeilijk is om vast te stellen of er sprake is van adverse impact en/of differentiële predictie. Het is dus belangrijk om te streven naar een zo groot mogelijke representativiteit en power van de data.
Ten vierde pleiten verschillende auteurs voor het herzien van wettelijke standaarden voor selectie met de opkomst van machine learning, in zowel de Verenigde Staten (Raghavan et al., 2019) als in Europa (Hacker, 2018). Herzieningen zijn nodig omdat via machine learning significante relaties worden vastgesteld die wij niet altijd goed kunnen interpreteren (zie Figuur 1b). Een statistisch valide assessment is niet per se ethisch acceptabel.
Ten vijfde, en tot slot, pleiten auteurs voor het verder verkennen en stimuleren van best practices waarbij uitgevers en gebruikers van werving-en-selectie-software actief zoeken naar alternatieven die niet, of zo min mogelijk, tot adverse impact leiden. Wetgeving moet hierbij zo duidelijk mogelijk zijn (TNO, 2018). Verschillende verkopers doen dit al (Raghavan et al., 2019) en dit kan verder worden aangemoedigd.
Omdat systemen, zoals algoritmes voor selectie, ontstaan door technische, politieke, juridische en maatschappelijke ontwikkelingen, is het belangrijk om bovenstaande aanbevelingen op een interdisciplinaire manier aan te pakken. In plaats van het tegenover elkaar plaatsen van vakgebieden pleiten wij ervoor om de sterktes vanuit de traditionele personeelsselectie en de computerwetenschappen te integreren ten behoeve van betere selectie- en assessmentprocedures. Wij hopen met dit artikel, een uitkomst van het interdisciplinaire project Big Data in Psychological Assessment, hieraan bij te dragen.

Praktijkbox

Wat betekenen de resultaten voor de praktijk?

- Algoritmes gebaseerd op machine learning leiden mogelijkerwijs tot minder (indirecte) discriminatie in personeelsselectie. Het is in de selectiepraktijk echter nog niet mogelijk om vast te stellen of deze belofte ook wordt ingelost.
- Dit komt deels door een gebrek aan transparantie over de algoritmes en de kwaliteit van de data die gebruikt worden om algoritmes te ontwikkelen, maar ook doordat vaak alleen naar de selectie-uitkomsten wordt

gekeken (adverse impact) als indicatie van bias. Ook in de data en tijdens de ontwikkeling van het algoritme kan echter bias optreden.
- We raden assessmentpsychologen, verkopers van machine learning-algoritmes, en beleidsmakers aan om het ontwikkelproces van algoritmes gebaseerd op machine learning te monitoren. Dat is noodzakelijk om algoritmes op een verantwoorde wijze te kunnen inzetten voor personeelsselectie. Het is bovendien nodig vanwege herziening van de Nederlandse wettelijke standaarden voor werving en selectie.
- Om te kunnen monitoren is basiskennis nodig van de werkwijze van machine learning-specialisten. Deze kennis wordt onder andere toegankelijk gemaakt in lesmateriaal dat is ontwikkeld in het project Big Data in Psychological Assessment (www.bdpa.eu).

Literatuur

Ajunwa, I., Friedler, S., Scheidegger, C., & Venkatasubramanian, S. (2016). *Hiring by algorithm: Predicting and preventing disparate impact.* Retrieved from sorelle.friedler.net/papers/SSRN-id2746078.pdf

Barocas, S., Bradley, E., Honavar, V., & Provost, F. (2017). *Big data, data science, and civil rights.* Retrieved from van arxiv.org/ftp/arxiv/papers/1706/1706.03102.pdf

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104,* 671-732. doi:10.15779/Z38BG31

Benton, M., Fratzke, S., & Sumption, M. (2014). *Moving up or standing still? Access to middle-skilled work for newly arrived migrants in the European Union.* Geneva: ILO. doi: http://hdl.voced.edu.au/10707/342012

Bogen, M., Rieke, A., & Ahmed, S. (2019). Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. *ArXiv preprint.* Retrieved from arXiv:1912.06171.

Calders, T., & Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society* (pp. 43-57). Berlin/Heidelberg: Springer. doi:10.1007/978-3-642-30487-3_3

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (2017). The datafication of talent: How technology is advancing the science of human potential at work. *Current Opinion in Behavioral Sciences*, *18*, 13-16. doi:10.1016/j.cobeha.2017.04.007

Chander, A. (2017). The racist algorithm. *Michigan Law Review*, *115*, 1023-1045. Retrieved from repository.law.umich.edu/mlr/vol115/iss6/13

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*, 311-313. doi:10.1038/538311a

De Neijs, N. (2017). *Bias and discrimination through hiring algorithms and how to avoid them*. Bachelor scriptie (niet gepubliceerd). Rotterdam: Erasmus Universiteit Rotterdam.

Ebrahimi, S. (2018). *Demographic transparency to combat discriminatory data analytics recommendations* (Dissertation). Retrieved from hdl.handle.net/11375/23406

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*, 38290-38315.

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). 'Where's the IO?' Artificial Intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions*, *5*, 33-44. doi:10.25035/pad.2019.03.005

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, *55*(4), 1143-1185. Retrieved from papers.ssrn.com/sol3/papers.cfm?abstract_id=3164973

Harwell, D. (2019). Rights group files federal complaint against AI-hiring firm HireVue, citing 'unfair and deceptive' practices. *The Washington Post*. Retrieved from https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/

Hirsch, P. B. (2016). The Caliphate of numbers. *Journal of Business Strategy*, 37, 51-55. doi:10.1108/JBS-09-2016-0098

Kim, P. T. (2016). *Data-driven discrimination at work*. Retrieved from www.scholarship.law.wm.edu/wmlr/vol58/iss3/4

König, C. J., Demetriou, A. M., Glock, P., Hiemstra, A. M., Iliescu, D., Ionescu, C.,... & Vartholomaios, I. (2020). Some advice for psychologists who want to work with computer scientists on big data. *Personnel Assessment and Decisions*, *6*, 17-23. doi:10.25035/pad.2020.01.002

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology: An International Review*, *98*, 1060-1072. doi:10.1037/a0034156

Liem, C. C. S., Langer, M., Demetriou, A., Hiemstra, A. M. F., Sukma Wicaksana, A., Born, M. Ph., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, I. Guyon, & S. Escalera (Eds.), *The Springer Series on Challenges in Machine Learning* (pp. 197-253). Cham: Springer Nature.

Lipton, Z., McAuley, J., & Chouldechova, A. (2018). Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems* (pp. 8125-8135).

MacCarthy, M. (2017). Standards of fairness for disparate impact assessment of big data algorithms. *Cumberland Law Review*, *48*(1), 67-148. Retrieved from heinonline.org/HOL/P?h=hein.journals/cumlr48&i=75

Mainka, S. M. (2018). Algorithm-based recruiting technology in the workplace. *Texas A&M Journal of Property Law*, *5*, 801-822. Retrieved from heinonline.org/HOL/P?h=hein.journals/txamrpl5&i=359

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2019). *Mitigating bias in algorithmic employment screening: Evaluating Claims and Practices.* Retrieved from arxiv.org/abs/1906.09208

Raub, M. (2018). Bots, bias and big data: Artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Arkansas Law Review*, *71*, 529-570. Retrieved from heinonline.org/HOL/P?h=hein.journals/arklr71&i=549

Savage, D. D., & Bales, R. A. (2016). *Video games in job interviews: Using algorithms to minimize discrimination and unconscious Bias*. Retrieved from ssrn.com/abstract=2887757

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262-274. doi: 10.1037/0033-2909.124.2.262

SIOP (Society for Industrial and Organizational Psychology). (2018). *Principles for the validation and use of personnel selection procedures*. Retrieved from www.siop.org/_principles/principles.pdf

Strahilevitz, L. J. (2008). Privacy versus antidiscrimination. *University of Chicago Law Review*, 75, 363. Retrieved from chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1480&context=journal_articles

Suen, H. Y., Chen, M. Y. C., & Lu, S. H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, 93-101. doi:10.1016/j.chb.2019.04.012

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*, 15-42. doi:10.1177/0008125619867910

TNO. (2018). *Risico's voor discriminatie bij werving en selectie: Huidige gang van zaken en trends*. Rapport TNO-2018-R11086. Retrieved from https://publications.tno.nl/

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2019). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, 1-14. doi:10.1080/1359432X.2019.1681401

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 1-30. doi:10.1007/s10618-017-0506-1

Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*, 183-201. doi:10.1007/s10506-016-9182-5

The promises and perils of machine learning algorithms to reduce bias and discrimination in personnel selection procedures

Hiemstra, A. M. F., Cassel, T., Born, M. Ph., & Liem, C. C.S. (2020), Gedrag & Organisatie, volume 33, no. 4, pp. 279-299.

In this article, we describe the implementation of algorithms based on machine learning for personnel selection procedures and how this data-driven approach corresponds to and differentiates from classical psychological assessment. We discuss if, and in what way, bias and discrimination occur when using algorithms based on machine learning for personnel selection. For this reason, we conducted a literature review (covering 2016-2019) from which 41 articles were included. The results indicate that algorithms possibly lead to reduced (indirect) discrimination compared to some other selection methods. This is one of the reasons why the development of algorithms for personnel selection has increased quickly and the number of vendors has grown fast. It is insufficiently possible yet, however, to ascertain if the promise is kept. First, this is because algorithms are often trade secrets (lack of transparency). Second, the validity and reliability of data used for the development of algorithms are not always clear. Furthermore, psychological selection issues about diversity and validity cannot (yet) be solved by algorithms. The increasing attention for the topic, expressed by a large growth in publications, is hopeful. We conclude with recommendations for the detection and reduction of bias and discrimination when using machine learning algorithms for personnel selection.

Keywords: algorithms, personnel selection, bias, discrimination, machine learning