CENTER FOR
SAFE
ARTIFICIAL
INTELLIGENCE
TU/e

Algoritmes en Discriminatie

De (on)zin van kwantitatieve methodes voor het meten en voorkomen van discriminatie door de inzet van algoritmes.

Hilde Weerts (Technische Universiteit Eindhoven)

Januari 2026

Auteur: Hilde Weerts

Dit onderzoek is uitgevoerd in opdracht van het College voor de Rechten van de Mens.

Inhoudsopgave

Lijst van gebruikte afkortingen

Afkorting	Betekenis
AUC	Area Under the receiver operating characteristic Curve
AIV	Artificiële Intelligentie Verordening
AVG	Algemene Verordening Gegevensbescherming
Awgb	Algemene Wet Gelijke Behandeling
CBS	Centraal Bureau voor de Statistiek
College	College voor de Rechten van de Mens
HvJ-EU	Hof van Justitie van de Europese Unie

Gebruikte definities

Begrip	Definitie/toelichting
Algoritme	Een reeks stappen die een computer automatisch kan uitvoeren.
Beschermd persoonskenmerk	Een persoonskenmerk waarop (behoudens objectieve rechtvaardiging) geen onderscheid gemaakt mag worden. Zie ook: *discriminatiegrond*.
Demografische pariteit	Een *fairness metric* die stelt dat de groepsspecifieke selectieratio's (i.e. de proportie leden van een groep die is geselecteerd) gelijk moeten zijn tussen groepen.
Discriminatiegrond	Juridisch beschermde persoonskenmerken waarop (behoudens objectieve rechtvaardiging) geen onderscheid gemaakt mag worden. Voorbeelden van discriminatiegronden zijn 'ras'[1], nationaliteit, religie, geslacht, seksuele gerichtheid, handicap of chronische ziekte.
Directe discriminatie	Een behandeling van een persoon of groep personen op andere wijze dan een ander in een vergelijkbare situatie op grond van een beschermd persoonskenmerk (discriminatiegrond) zonder dat hiervoor een objectieve rechtvaardiging bestaat.
Doelvariabele	De variabele die men door middel van een voorspellend algoritme probeert te voorspellen. In klassiek statistisch jargon wordt deze variabele ook de *afhankelijke variabele* genoemd.
False negative rate ("valsnegatievenratio")	Het aantal valsnegatieven gedeeld door het aantal daadwerkelijke positieven.
False positive rate ("valspositievenratio")	Het aantal valspositieven gedeeld door het aantal daadwerkelijke negatieven.
Fairness metric ("eerlijkheidsmetriek")	Een metriek ontwerpen om 'eerlijkheid' te kwantificeren, meestal op basis van een vergelijking van de uitkomsten van een algoritme voor verschillende (groepen) personen.
Group fairness	Een verzamelnaam voor *fairness metrics* welke eerlijkheid beogen te kwantificeren op groepsniveau.
Indirecte discriminatie	Een ogenschijnlijk neutrale bepaling, maatstaf of handelwijze die personen met een bepaald beschermd persoonskenmerk (discriminatiegrond) in vergelijking met andere personen in het bijzonder benadeelt, zonder dat hiervoor een objectieve rechtvaardiging bestaat.
Inputvariabele	Een variabele op basis waarvan het algoritme de doelvariabele voorspelt. In klassiek statistisch jargon wordt dit ook wel een *onafhankelijke* variabele genoemd.
Label	Mogelijke uitkomst of categorie van de doelvariabele en voorspellingen.

[1]De grond 'ras' is in het recht opgenomen om bescherming te bieden tegen het maatschappelijke verschijnsel dat mensen in de realiteit worden gediscrimineerd op grond van geracialiseerde kenmerken. Het bestaan van juridische discriminatiegrond 'ras' betekent niet dat het recht uitgaat van het feitelijk 'bestaan' van 'ras' als biologische categorie. Zie in deze zin overweging (6) van de Preambule van EU- Richtlijn 2000/43.

Risicoprofilering	Het categoriseren en/of selecteren van personen, op basis van persoonlijke, situationele, gedragsmatige of dossierkenmerken, voor opsporings- of controlebevoegdheden, die plaatsvindt aan de hand van algemene, verondersteld voorspellende (risico)criteria, die geacht worden te correleren met de kans op overtreding van een bepaalde norm, bij afwezigheid van concrete of specifieke informatie over feitelijke normovertredingen of daders.
Trainingsdata	De gegevens die worden gebruikt om een zelflerend algoritme te trainen. De trainingsdata bestaat uit de voorbeelden waarin het zelflerende algoritme patronen herkent om voorspellingen te doen.
Valsnegatief	Een positief voorbeeld dat is gemisclassificeerd als negatief.
Valspositief	Een negatief voorbeeld dat is gemisclassificeerd als positief.
Confusion matrix (“verwarringsmatrix”)	Een matrix die het absolute aantal (of de proportie) valspositieven, valsnegatieven, echtpositieven, en echtnegatieven weergeeft.
Zelflerend algoritme	Vorm van kunstmatige intelligentie waar op basis van grote hoeveelheden data statistische verbanden worden geïdentificeerd (*machine learning*).

1. Introductie

Organisaties maken steeds vaker gebruik van algoritmes voor selectie en categorisering om de effectiviteit en efficiëntie van werkprocessen te bevorderen. Het gebruik van algoritmes brengt echter ook significante risico's mee en heeft in de afgelopen jaren herhaaldelijk tot discriminatie geleid. Bekende voorbeelden zijn de kinderopvangtoeslagenaffaire, waar ouders met een niet-Nederlandse afkomst harder werden getroffen door de fraudeaanpak van de belastingdienst dan mensen van Nederlandse afkomst[2], en de controle uitwonendenbeurs door DUO, waar studenten met een niet-Europese migratieachtergrond waren oververtegenwoordigd in controles[3].

In de computerwetenschappen is het afgelopen decennium een veelvoud aan kwantitatieve methodes ontwikkeld voor het meten en mitigeren van onderscheid en (mogelijke) discriminatie. Denk aan statistieken voor het meten van onderscheid en technische "*de-biasing*" methodes om zelflerende algoritmes bij te sturen. In reactie op zorgen over discriminatie door algoritmes hebben ook veel publieke organisaties de afgelopen jaren kaders ontwikkeld om *bias* of vooringenomenheid te identificeren en voorkomen. Zo is in 2021 in opdracht van het Ministerie van Binnenlandse Zaken de *Handreiking Non-discriminatie by design*[4] ontwikkeld, heeft de gemeente Amsterdam in 2022 *The Fairness Handbook* geschreven[5], en zijn verscheidene praktische handvatten sinds 2024 gebundeld in het Algoritmekader[6]. Een recente inspanning is de ontwikkeling van een Nederlandse Technische Afspraak (NTA), waarin toetsingsmethodes worden beschreven die ondersteunen bij het voorkomen van discriminatie bij de inzet van profileringsalgoritmes.[7]

Ondanks de ontwikkeling van deze kaders blijft het voor zowel juristen als data scientists vaak onduidelijk in hoeverre kwantitatieve methodes overlappen met de juridische discriminatietoets en waar ze verschillen.[8] Het doel van dit rapport is om een overzicht te geven van gangbare kwantitatieve methodes voor het meten en voorkomen van onderscheid en discriminatie. Er wordt nader ingegaan op in hoeverre deze methodes overlappen met de juridische discriminatietoets. Ook wordt aangestipt waar op dit moment de grootste uitdagingen liggen in de praktijk.

1.1. Achtergrond en aanpak

Het College voor de Rechten van de Mens heeft de Technische Universiteit Eindhoven verzocht om onderzoek te doen naar kwantitatieve methodes voor het meten en voorkómen van discriminatie in algoritmes.

Onderzoeksvragen

Het College stelt de volgende vragen:

1. Wat zijn gangbare kwantitatieve methodes voor het meten en voorkomen van discriminatie in algoritmes?
2. In hoeverre overlappen deze methodes met de discriminatietoets?
3. Hoe gebruiken ontwikkelaars verschillende methodes en welke uitdagingen ondervinden zij hierbij?

Deze vragen worden in dit rapport beantwoord op basis van wetenschappelijke literatuur en interviews met 14 data professionals in de Nederlandse (semi)publieke sector. Meer details over de interviews zijn beschikbaar in de **Bijlage A: Interviewstudie.** De inzichten uit de interviews zijn door het rapport heen los vermeld waar relevant voor het beantwoorden van één van de onderzoeksvragen.

---

[2] Zie bijvoorbeeld Autoriteit Persoonsgegevens (2020) *Belastingdienst/Toeslagen: De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag*, Onderzoeksrapport z2018-22445; College voor de Rechten van de Mens (2022). *Vooronderzoek naar de vermeende discriminerende effecten van de werkwijzen van de Belastingdienst/Toeslagen*.
[3] Algorithm Audit (2024). *Vooringenomenheid voorkomen - Aanbevelingen voor risicoprofilering in het Controle Uitwonendenbeurs-proces: een kwantitatieve en kwalitatieve analyse*.
[4] B. van der Sloot et al. (2021). *Handreiking non-discriminatie by design*. Rijksoverheid.
[5] S. Muhammad (2022). *The Fairness Handbook*. Gemeente Amsterdam.
[6] Het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2024) *Het Algoritmekader*
[7] De Nederlandse normcommissie (NEN) begeleidt dit proces. Een groep deelnemers van meer dan dertig partijen van de overheid, het bedrijfsleven, de wetenschap, en belangenorganisaties nemen deel aan de ontwikkeling van de NTA.
[8] H. Weerts & R. Xenidis et al. (2023). *Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree.* In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 805-816).

1.2. Algoritmes en kunstmatige intelligentie

Dit rapport gaat over kwantitatieve methodes om ongewenst onderscheid te meten en voorkomen bij het gebruik van (datagedreven) algoritmes en kunstmatige intelligentie.

Algoritmes en kunstmatige intelligentie

Een algoritme is een reeks stappen die een computer automatisch kan uitvoeren. Een voorbeeld van een simpel algoritme in werving en selectie is de volgende als-dan-instructie: *als* een sollicitant een bepaald universitair diploma heeft, *dan* wordt deze uitgenodigd voor een gesprek. Tegenwoordig wordt steeds vaker gebruik gemaakt van zelflerende algoritmes, een vorm van kunstmatige intelligentie waar op basis van grote hoeveelheden data statistische verbanden worden geïdentificeerd (*machine learning*). Een zelflerend algoritme voor cv-selectie kan bijvoorbeeld gebruik maken van kenmerken zoals opleidingsniveau, werkervaring, en eerdere werkprestaties (de inputvariabelen) om te voorspellen of iemand binnen twee jaar promotie maakt (de doelvariabele). Zo zou het algoritme bijvoorbeeld op basis van historische data kunnen ontdekken dat personen met een universitair diploma in het verleden vaker een promotie hebben gekregen. Op basis van dergelijke patronen voorspelt het algoritme vervolgens welke sollicitanten een grotere kans hebben om promotie te maken.

Er bestaat een grote variëteit in algoritmes die kunnen worden ingezet door publieke en private organisaties. In dit rapport is gekozen voor een focus op categoriserende, statistisch onderbouwde algoritmes.

Categoriserende en statistisch onderbouwde algoritmes

Algoritmes kunnen resulteren in onderscheid: een verschil in behandeling tussen personen in een vergelijkbare situatie. De kans op onderscheid – in de juridische zin – is met name hoog bij categoriserende of selecterende besluitvormingsprocessen met directe invloed op individuen, zoals risicoprofilering bij normovertredingen of het toekennen van (hulp)middelen zoals een uitkering of toegang tot een bankrekening. In dit rapport is daarom gekozen op dergelijke algoritmes te focussen te focussen.

Daarnaast ligt het zwaartepunt van dit rapport op statistisch onderbouwde algoritmes, waarbij kwantitatieve, gegevens worden gebruikt voor het categoriseren van personen. Hoewel de nadruk ligt op algoritmes die tot stand gekomen zijn door middel van zelflerende algoritmes (*machine learning*), zijn de bevindingen ook grotendeels van toepassing op ‘handgemaakte’ regelgebaseerde algoritmes zonder zelflerend component.

Hoewel sommige bevindingen ook van toepassing kunnen zijn op niet-categoriserende algoritmes, worden deze buiten beschouwing gelaten. Niet-categoriserende algoritmes zijn bijvoorbeeld algoritmes die patronen ontdekken zonder vooraf gedefinieerde categorieën, zoals het clusteren van burgers in groepen met vergelijkbare kenmerken of het signaleren van ongebruikelijke patronen die mogelijk kunnen wijzen op fraude. Ook generatieve AI, een vorm van AI waarmee content zoals tekst, afbeeldingen, of audio kan worden gegenereerd, valt buiten de reikwijdte van dit rapport.

1.3. Meten en voorkómen van onderscheid en mogelijke discriminatie

In de computerwetenschappen is onder de noemer *algorithmic fairness* (“algoritmische eerlijkheid”) de afgelopen decennia een rijke wetenschappelijke literatuur ontstaan waarin (kwantitatieve) methodes voor het meten en voorkómen

van “oneerlijkheid” worden voorgesteld. Deze zogenoemde *fairness* metrieken en methodes zijn veelal geïnspireerd door juridische concepten zoals (indirecte) discriminatie.[9]

Kwantitatieve metrieken vervangen nooit de juridische discriminatietoets door een rechterlijke instantie

Hoewel sommige kwantitatieve metrieken gedeeltelijk overlappen met de juridische discriminatietoets, kunnen ze nooit ter vervanging dienen voor de (kwalitatieve) contextuele weging vereist voor de discriminatietoets zoals uitgevoerd door een rechterlijke instantie. In dit rapport spreken we daarom van het meten en voorkomen van ongewenst onderscheid en *mogelijke* discriminatie.

1.4. Leeswijzer

In **Discriminatie en** algoritmes wordt nader ingegaan op het discriminatieverbod en algoritmes. **Het meten** van onderscheid biedt een overzicht van gangbare kwantitatieve methodes voor het meten en detecteren van ongewenst onderscheid. Daarnaast wordt ingegaan in hoeverre deze methodes overlappen met de discriminatietoets en welke uitdagingen bestaan in de praktijk. In **Het voorkómen** van discriminatie wordt vervolgens ingegaan op technische methodes die zijn ontwikkeld met het doel mogelijke discriminatie te voorkomen. Er wordt nader ingegaan op de vraag in hoeverre deze methodes bruikbaar zijn in de context van het Nederlandse non-discriminatierecht. De belangrijkste conclusies worden samengevat in **Conclusies**.

[9] J. Adams-Prassl, R. Binns, A. Kelly-Lyth (2023). *Directly discriminatory algorithms*. The Modern Law Review, 86(1), 144-175.

2. Discriminatie en algoritmes

2.1. Het discriminatieverbod

Discriminatie is het behandelen van een persoon (of groep personen) op andere wijze dan een ander in een vergelijkbare situatie op grond van een beschermd persoonskenmerk (discriminatiegrond), zonder dat hiervoor een objectieve rechtvaardiging bestaat.

Het non-discriminatierecht is vastgelegd in Artikel 1 van de Grondwet en verschillende nationale en Europese wetten en internationale verdragsbepalingen.[10] Deze wetten zijn nader ingevuld door jurisprudentie. Het discriminatieverbod verbiedt sommige vormen van onderscheid – verschil in behandeling in een vergelijkbare situatie – op basis van één of meerdere discriminatiegronden. De Nederlandse gelijkebehandelingswetgeving verbiedt zowel direct als indirect onderscheid.

Er is sprake van **direct onderscheid** wanneer personen in een vergelijkbare situatie verschillend worden behandeld op basis van een discriminatiegrond. Op enkele specifieke uitzonderingen na, mogen discriminatiegronden geen expliciete factor zijn in besluitvormingsprocessen die vallen onder het discriminatieverbod. Voorkeursbeleid is bijvoorbeeld één van de uitzonderingen op het verbod op direct onderscheid.[11] Voorkeursbeleid heeft als doel feitelijke achterstanden van beschermde groepen teniet te doen of te verminderen. De Awgb voorziet een uitzondering op het verbod op onderscheid als het "*een specifieke maatregel betreft die tot doel heeft vrouwen of personen behorende tot een bepaalde etnische of culturele minderheidsgroep een bevoorrechte positie toe te kennen ten einde feitelijke nadelen verband houdende met de gronden ras of geslacht op te heffen of te verminderen en het onderscheid in een redelijke verhouding staat tot dat doel*".

Indirect onderscheid vindt plaats wanneer een ogenschijnlijk neutrale handeling of beleid personen uit een beschermde groep in het bijzonder treft, in vergelijking met anderen in een vergelijkbare situatie. Indirect onderscheid is verboden, tenzij er een objectieve rechtvaardiging bestaat. Deze rechtvaardiging hangt af van het doel dat met het onderscheid wordt nagestreefd en het middel dat daarvoor wordt ingezet. Hierbij geldt dat het doel legitiem moet zijn en het ingezette middel passend en noodzakelijk. Een middel is *passend* als het geschikt is om het beoogde doel te bereiken. Het middel is *noodzakelijk* als het belang in verhouding staat tot de inbreuk (*proportionaliteit*) en er geen minder ingrijpend middel is waarmee het doel bereikt kan worden (*subsidiariteit*). Onderscheid dat op deze wijze gerechtvaardigd kan worden is juridisch gezien dus geen discriminatie. De mogelijkheid tot een objectieve rechtvaardiging is sterk contextafhankelijk en de proportionaliteitstoets wordt door rechters strikter (of minder strikt) ingevuld afhankelijk van bijvoorbeeld het domein of de discriminatiegrond.

2.2. Discriminatie door de inzet van algoritmes

Er zijn verschillende manieren waarop de inzet van algoritmes kan leiden tot onderscheid tussen groepen mensen met een bepaald (beschermd) persoonskenmerk. Binnen de overheid worden algoritmes vaak ingezet voor selectie. Denk bijvoorbeeld aan de selectie van burgers voor het toekennen van bepaalde voorzieningen (zoals extra hulp bij het zoeken naar een baan), kansen (bijvoorbeeld een uitnodiging voor een sollicitatiegesprek), of informatie (zoals een melding over de mogelijkheid om belasting terug te vragen). Ook worden algoritmes ingezet voor risicoprofilering in het kader van handhaving, bijvoorbeeld als personen voor (nadere) controles worden geselecteerd met algoritmes. Wanneer personen met een bepaald beschermd persoonskenmerk vaker (of juist minder vaak) worden geselecteerd dan anderen in een vergelijkbare situatie, kan de inzet van een selectiealgoritme leiden tot (in)direct onderscheid.

Ook wanneer een algoritme niet expliciet wordt ingezet voor selectie, kunnen verschillen in voorspelkracht leiden tot onderscheid in de kwaliteit van dienstverlening of producten. Denk bijvoorbeeld aan gezichtsherkenningssoftware die

[10] Nederlandse wetgeving over non-discriminatie is onder andere te vinden in Artikel 1 van de Grondwet en gelijkebehandelingswetten zoals de *Algemene wet gelijke behandeling,* de *Wet gelijke behandeling van mannen en vrouwen,* en de *Wet gelijke behandeling op grond van leeftijd bij arbeid*. Deze nationale wetten zijn gedeeltelijk implementaties van verschillende EU-richtlijnen, zoals richtlijnen voor rassengelijkheid (2000/43/EC) en gendergelijkheid (2004/113/EC). Daarnaast bevatten ook het *Verdrag betreffende de werking van de Europese Unie* relevant en het *Handvest van de grondrechten van de Europese Unie* bepalingen gerelateerd aan het recht op non-discriminatie.
[11] Algemene Wet Gelijke Behandeling, Artikel 2.3.

minder goed werkt voor personen met een donkere huidskleur ten op zichtte van personen met een lichtere huidskleur. Daarnaast kan oververtegenwoordiging van gemarginaliseerde groepen stereotypes versterken en daardoor stigmatiserend werken. Selectieve controles op een specifieke bevolkingsgroep, als gevolg van risicoprofilering, kunnen leiden tot meer opsporing van overtredingen binnen deze groep. Dit kan vervolgens leiden tot een vertekend en stigmatiserend beeld van de betreffende bevolkingsgroep. Ook een dergelijk collectief nadeel kan als discriminatie worden gekwalificeerd.

Discriminatie is geen "technisch" probleem

In dit hoofdstuk wordt dieper ingegaan op technische aspecten in het ontwerp van algoritmes die het risico op discriminatie kunnen vergroten. Dit wil niet zeggen dat discriminatie een "technisch" probleem is.

Discriminatie is een vaak complexe samenloop van factoren, waarin zowel mensen en instituten als techniek een rol spelen. Om te beginnen wordt techniek ontworpen door mensen. Menselijke vooroordelen en vooringenomenheid kunnen invloed hebben op ontwerpkeuzes, van dataverzameling en modellering tot evaluatie en implementatie. Ook is het niet enkel van belang *wat* een algoritme doet, maar ook *hoe* het wordt ingezet. Wat is de impact van selectie voor personen die worden geselecteerd? Wordt het algoritme toegepast op al gemarginaliseerde groepen? Is er een toegankelijke en effectieve mogelijkheid om bezwaar te maken waardoor onterechte besluiten tijdig worden hersteld? Is het besluit volledig geautomatiseerd of vindt er ook een menselijke beoordeling plaats? Dergelijke contextuele factoren hebben grote invloed op het risico op discriminatie.

Kwantitatieve metrieken en technische interventies zijn daarom nooit afdoende om discriminatie te detecteren en voorkomen. Dat gezegd hebbende, zijn er een aantal technische aspecten, specifiek voor algoritmes, die het risico op discriminatie vergroten.

In de rest van dit hoofdstuk wordt dieper ingegaan op twee oorzaken van ongewenst onderscheid in datagedreven algoritmes: verschillen in voorspelkracht tussen beschermde groepen en een statistische verband tussen een discriminatiegrond en doelvariabele.

Voorspelkracht

Wanneer een algoritme minder accuraat is voor sommige groepen ten opzichte van andere groepen, kan dit leiden discriminatie.[12] Zo oordeelde het College in een tussenoordeel in 2022 dat de inzet van antispieksoftware voor het surveilleren van tentamens kan leiden tot indirect onderscheid op grond van ras, als de gezichtsregistratie-software slechter presteert voor personen met een donkere huidskleur.[13] Bij datagedreven algoritmes hangt voorspelkracht nauw samen met ontwerpkeuzes tijdens het ontwikkelingsproces, zoals de keuze voor een specifieke trainingsdataset, keuzes in datavoorbereiding, en de keuze voor een specifiek type zelflerend algoritme.

Wanneer sommige (minderheids)groepen ondervertegenwoordigd zijn in een dataset, kan dit leiden tot slechtere prestaties voor deze groepen. Een gezichtsherkenningssysteem dat primair getraind is op foto's van personen met een lichte huidskleur, kan bijvoorbeeld minder goed werken voor personen met een donkere huidskleur.[14]

Verschillen in prestaties kunnen ook ontstaan door de keuze om bepaalde kenmerken (inputvariabelen*)* wel of niet mee te nemen in de trainingsdata. Het is bijvoorbeeld bekend dat bepaalde risicofactoren, zoals langdurig roken en zwangerschapsdiabetes, belangrijke voorspellers zijn voor hart- en vaatziekten bij vrouwen, maar minder belangrijk (in het

[12] Het is van belang op te merken dat risicoscores en voorspelkracht altijd gaan over kansen die empirisch berekend worden over een *groep* personen. Empirisch berekende accuraatheid van een voorspelling op groepsniveau zegt enkel iets over de betrouwbaarheid op dat (geaggregeerde) groepsniveau en dus niet voor een individueel geval dat behoort tot die groep.
[13] Zie College voor de Rechten van de Mens (2022) *Tussenoordeel 2022-146*.
[14] J. Buolamwini & T. Gebru (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. In Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR.

geval van roken) of ongedefinieerd (in het geval van zwangerschapsdiabetes) voor (cisgender) mannen.[15] Wanneer deze risicofactoren niet worden opgenomen in de trainingsdataset, presteert het algoritme mogelijk minder goed voor vrouwen dan voor mannen.

Verschillen in voorspelkracht kunnen worden vergroot door ontwerpkeuzes tijdens de datavoorbereiding en modellering. Zo kan de keuze voor een zo simpel mogelijk model[16] ertoe leiden dat niet alle complexe relaties tussen variabelen even goed worden gevangen. Een mogelijk gevolg is dat het simpele model goed presteert voor de meerderheidsgroep, maar minder accuraat is voor de minderheidsgroep. Ook keuzes in de datavoorbereiding, zoals (geautomatiseerde) selectie van variabelen en het imputeren van missende waardes[17] kunnen leiden tot verschillen in prestaties.

Een belangrijk punt, genoemd door enkele geïnterviewden in de interviewstudie, betreft datakwaliteit. Data ondergaat vaak vele, soms handmatige, bewerkingen voordat zij gebruikt kan worden in een (zelflerend) algoritme, wat de accuraatheid niet altijd ten goede komt. Ook zijn datasets vaak niet compleet. Missende waardes kunnen bijvoorbeeld ontstaan doordat een burger, klant, of organisatie bepaalde gegevens niet heeft aangeleverd. Ook wanneer een dataset compleet is, kunnen er grote verschillen zijn in datakwaliteit. Beleidsprocessen binnen de overheid zijn vaak afhankelijk van extern aangeleverde gegevens. Zeer kleine organisaties (bijvoorbeeld particuliere verhuurders of gastouders), hebben veelal niet de capaciteit om datalevering even professioneel in te richten als grotere organisaties (bijvoorbeeld woningcorporaties of kinderdagverblijfketens). Dit kan ertoe leiden dat klanten van kleinere organisaties vaker worden gecontroleerd.

Statistisch verband

Zelflerende algoritmes maken gebruik van statistische verbanden tussen inputvariabelen en de doelvariabele om de doelvariabele zo goed mogelijk te voorspellen. Zo kan een cv-selectie-algoritme bijvoorbeeld worden getraind om op basis van jaren relevante werkervaring, behaalde diploma's, en persoonlijke verkoopcijfers (de inputvariabelen) te voorspellen of een sollicitant wel of niet wordt aangenomen (de doelvariabele). De doelvariabele bepaalt dus in grote mate waar het algoritme voor optimaliseert. **Als er een statistisch verband bestaat tussen de doelvariabele en een discriminatiegrond, zal een (accuraat) algoritme dit verband reproduceren.** De inzet van het algoritme kan hierdoor leiden tot onderscheid.

Er zijn verschillende redenen voor een (ongewenste) associatie tussen een discriminatiegrond en de doelvariabele. Enerzijds kan het een gevolg zijn van bestaande ongelijkheid in de samenleving. Anderzijds kan het veroorzaakt worden door keuzes in de dataverzameling.

Ongelijkheid in de samenleving

Als een bepaalde karakteristiek niet evenredig is verdeeld over de maatschappij, zal een accuraat algoritme dit patroon reproduceren. Neem bijvoorbeeld een algoritme dat wordt gebruikt om het risico op borstkanker te voorspellen. Omdat borstkanker vele malen vaker voorkomt bij vrouwen dan bij mannen, zal het algoritme als vanzelfsprekend de kans op borstkanker gemiddeld (veel) hoger inschatten voor vrouwen dan voor mannen. Als vrouwen op basis van deze scores vaker preventief gescreend worden dan mannen, is er sprake van onderscheid (dat objectief gerechtvaardigd kan worden). Ook is momenteel in Nederland migratieachtergrond gecorreleerd met sociaaleconomische positie, opleidingsniveau, en gezondheidsbeleving[18]. Besluitvorming op basis van (voorspelde) karakteristieken zoals inkomen of arbeidsmarktpositie kan dus leiden tot (indirect) onderscheid op basis van afkomst. Een dergelijk besluitvormingsproces is verboden, tenzij er sprake is van een objectieve rechtvaardiging.

---

[15] Zie bijvoorbeeld Y. Appelman et al. (2015). *Sex differences in cardiovascular risk factors and disease prevention*. Atherosclerosis, 241(1), 211-218.

[16] Een relatief simpel model kan bijvoorbeeld worden nagestreefd door te kiezen voor een simpel lineair model met een hoge regularisatieparameter.

[17] Het imputeren van missende waardes betekent dat ontbrekende gegevens in een dataset worden aangevuld met geschatte of vervangende waarden. Voor de variabele "leeftijd" kan bijvoorbeeld worden gekozen missende waardes te vervangen door de gemiddelde leeftijd of vaakst voorkomende leeftijd in de trainingsdata. Geavanceerdere methodes schatten missende waardes op basis van andere variabelen die correleren met de variabele waarvoor een waarde mist.

[18] Zie bijvoorbeeld CBS (2024). *Longread integratie en samenleven 2024*.

Datavaliditeit

Algoritmes worden vaak ingezet om complexe, moeilijk observeerbare constructen te voorspellen, zoals "zorgbehoeftes", "normovertredingen", of "kansen op de arbeidsmarkt". De keuze voor een specifieke doelvariabele on een dergelijk ingewikkeld construct te vangen kan belangrijke gevolgen hebben voor de mate waarin ongewenst onderscheid optreedt.

Doelvariabelen kunnen sterk verschillen in de mate waarin ze onderscheid veroorzaken. Zo voorspelde een veelgebruikt commercieel algoritme voor het toekennen van extra zorg in de Verenigde Staten "zorgkosten" als proxy[19] voor "zorgbehoeftes".[20] Deze ontwerpkeuze resulteerde in lagere risicoscores voor patiënten van kleur ten opzichte van even zieke witte patiënten. De reden: Amerikanen van kleur zijn gemiddeld minder goed verzekerd dan witte Amerikanen, waardoor er minder geld aan hen wordt uitgegeven in de zorg. "Zorgkosten" is hierdoor gemiddeld een minder goede maatstaf voor "zorgbehoeften" voor Amerikanen van kleur dan voor witte Amerikanen. Het gebruik van een andere, inhoudelijk relevantere doelvariabele, zoals de aanwezigheid van chronische aandoeningen, resulteerde in een minder scheve verdeling van risicoscores. Bij een risicoprofileringsalgoritme is bijvoorbeeld de definitie van "fraude" bepalend. Hoe is bepaald of er sprake was van opzettelijke fraude of een vergissing? Zijn er gevallen die mogelijk foutief als niet-frauduleus zijn gelabeld? Wanneer sommige groepen in de samenleving vaker een vergissing maken, bijvoorbeeld omdat Nederlands niet hun moedertaal is, kan een gebrekkige definitie van "fraude" leiden tot indirect onderscheid.

Wanneer een doelvariabele gedefinieerd is op basis van (historische) menselijke beoordelingen, bestaat het risico dat het algoritme (onbewuste) vooroordelen reproduceert. Zo heeft onderzoek veelvuldig aangetoond dat discriminatie op de arbeidsmarkt een hardnekkig probleem is.[21] Een algoritme getraind op bevooroordeelde wervings- en selectiebeslissingen uit het verleden zal deze vooroordelen reproduceren, met (in)direct onderscheid als gevolg. Selectiealgoritmes in de publieke sector worden soms getraind op beoordelingen van zaakbehandelaars, bijvoorbeeld op basis van dossier- of veldonderzoek. Dit kan leiden tot onderscheid. Zo bleek uit een onderzoek naar de controle uitwonendenbeurs van DUO dat studenten met een niet-Europese migratieachtergrond zes keer vaker handmatig werden geselecteerd voor een huisbezoek ten opzichte van studenten met een Nederlandse of Europese herkomst.[22]

Problemen kunnen ook ontstaan doordat de verzamelde data niet (meer) representatief is voor de context waarin het algoritme wordt toegepast. Een cv-selectiemodel getraind op wervingsbeslissingen in de detailhandel is bijvoorbeeld niet per definitie geschikt voor de gezondheidszorg. In de publieke sector kan een wets- of beleidswijziging ervoor zorgen dat besluiten uit het verleden niet langer van toepassing zijn, bijvoorbeeld omdat de criteria voor het ontvangen van een toeslag zijn aangepast.

Daarnaast bepaalt in veel gevallen bestaand beleid welke datapunten wel of niet zijn verzameld. Dit kan als gevolg hebben dat sommige groepen worden onder- of oververtegenwoordigd in de data. Een bekend voorbeeld betreft risicomodellen bedoeld voor het voorspellen van criminaliteit.[23] Dergelijke modellen worden vaak getraind op criminaliteitsstatistieken. Deze statistieken zijn echter sterk afhankelijk van bestaande politiepraktijken, zoals de verdeling van agenten over buurten en de inzet van proactieve controles. Wanneer politieagenten bijvoorbeeld vaker worden ingezet in buurten met een hoog percentage etnische minderheden of zich schuldig maken aan etnisch profileren, kan dit leiden tot overrepresentatie van etnische minderheden in de criminaliteitscijfers. Een risicomodel getraind op dergelijke data kan daardoor leiden tot (in)direct onderscheid op basis van afkomst.

Het is belangrijk om te benadrukken dat validiteit niet kan worden beoordeeld op basis van de dataset zelf: het gaat over de *interpretatie* van wat er wordt gemeten. Wanneer er een associatie bestaat tussen een persoonskenmerk en de doelvariabele, kan daar dus niet direct uit geconcludeerd worden of er sprake is van gebrekkige validiteit of bestaande ongelijkheid. Daar is altijd extra contextuele informatie voor nodig en moet dus per situatie worden beoordeeld.

---

[19] Een *proxy* is een variabele die wordt gebruikt om een andere variabele mee te nemen, ofwel omdat er een sterk statistisch verband bestaat tussen de twee variabelen ofwel omdat aangenomen wordt dat de ene variabele de andere goed kan voorspellen.
[20] Z. Obermeyer et al. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. Science, 366(6464), 447-453.
[21] Zie bijvoorbeeld Sociaal en Cultureel Planbureau (2020. *Ervaren discriminatie in Nederland II*.
[22] Algorithm Audit (2024). *Addendum vooringenomenheid voorkomen*.
[23] Bao et al. (2021) *It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks*. NeurIPS Datasets and Benchmarks.

3. Het meten van onderscheid

3.1. Kwantitatieve *fairness metrics*

In de computerwetenschappelijke literatuur worden verschillende *fairness metrics* (“eerlijkheidsmetrieken”) voorgesteld waarmee mogelijke discriminatie kan worden gekwantificeerd. Dergelijke statistieken kwantificeren de mate waarin de uitkomsten van het algoritme verschillen tussen (groepen) personen. In de wetenschappelijke literatuur bestaat een grote verscheidenheid aan *fairness metrics*. De belangrijkste verschillen tussen metrieken zitten in het niveau van de vergelijking (vergelijken we individuen of groepen?) en de onderliggende (normatieve) aannames (wát zou gelijk moeten zijn tussen (groepen) personen?).

Group fairness* en *individual fairness

Verreweg het populairste type *fairness metrics* in zowel de wetenschappelijke literatuur als de praktijk zijn zogenoemde *group fairness metrics*. *Group fairness metrics* kwantificeren verschillen in de output van het algoritme tussen (sub)groepen van mensen. In de wetenschappelijke literatuur bestaan hiernaast ook metrieken die eerlijkheid pogen te benaderen vanuit het perspectief van het individu. Zo meten *individual* (“individuele”) *fairness metrics* in hoeverre vergelijkbare datapunten een vergelijkbare voorspelling krijgen[24] en *counterfactual* (“contrafeitelijke”) *fairness metrics* in hoeverre de voorspelling zou veranderen wanneer de persoon tot een andere beschermde groep had behoord, op basis van veronderstelde causale verbanden[25]. Dergelijke metrieken bieden in theorie meer ruimte om normatieve assumpties gedetailleerd vast te leggen en kwantificeren ten op zichtte van *group fairness* metrieken. Deze flexibiliteit is echter ook een valkuil: de metrieken zijn vanuit zowel technisch als normatief oogpunt complex en hierdoor in de praktijk moeilijk toepasbaar. Uit de interviews blijkt dat *individual fairness* en *counterfactual fairness* metrieken in de Nederlandse (semi)publieke sector dan ook niet of nauwelijks worden gebruikt. In de rest van dit rapport is daarom gekozen om te focussen op *group fairness*.

In een *group fairness* analyse worden verschillende statistieken vergeleken tussen groepen. Populaire statistieken in zowel de wetenschappelijke literatuur als de praktijk zijn de selectieratio en misclassificatieratio’s.[26] In kleinere mate wordt gebruik gemaakt van andere statistieken gerelateerd aan de prestaties van het algoritme.[27]

Selectieratio’s

Algoritmes worden vaak ingezet als selectie-instrument. Om te beoordelen of er sprake is van mogelijk onderscheid, ligt het voor de hand te onderzoeken of personen met bepaalde persoonskenmerken relatief vaker of minder vaak worden geselecteerd. In een *group fairness* analyse kan dit worden gekwantificeerd door middel van de selectieratio: de proportie van individuen die wordt geselecteerd (**Figuur 1**). Wanneer de selectieratio’s gelijk zijn tussen (demografische) groepen, wordt in de literatuur gesproken van *demografische pariteit*.[28]

[24] *Individual fairness metrics* beogen te meten in hoeverre personen in een vergelijkbare situatie gelijk worden behandeld. Of een persoon zich in een vergelijkbare situatie bevindt als een andere persoon, wordt gekwantificeerd door middel van een *similarity metric*, welke wordt berekend aan de hand van verschillen in (input) variabelen tussen de twee personen. Een populaire *similarity metric* in de literatuur is bijvoorbeeld de Euclidische afstand. In het geval van twee inputvariabelen, bijvoorbeeld leeftijd en inkomen, is de Eucledische afstand tussen twee datapunten gelijk aan de hemelsbrede afstand tussen de twee punten in een grafiek waarin leeftijd en inkomen tegen elkaar worden uitgezet.
[25] Zie bijvoorbeeld Makhlouf et al. (2024). *When causality meets fairness: A survey*. Journal of Logical and Algebraic Methods in Programming, 101000.
[26] Alle 14 geïnterviewden hebben aangegeven te kijken naar over- en ondervertegenwoordiging in de selectie als ook misclassificatie (indien mogelijk).
[27] Voorbeelden van andere classificatiemetrieken zijn accuraatheid (*accuracy*: het aantal accurate voorspellingen gedeeld door het totaal aantal voorspellingen) en precisie (*precision:* de proportie daadwerkelijke positieven van alle voorspelde positieven). Prestatiemetrieken gerelateerd aan de risicoscore zijn bijvoorbeeld AUC (*Area under the Receiver Operating Curve*: een metriek die meet in hoeverre de scores het mogelijk maken om positieve en negatieve voorbeelden van elkaar te onderscheiden) en kalibratie (*calibration*: de mate waarin de scores van het algoritme overeenkomt met de daadwerkelijke kans op een positief).
[28] Deze *metric* staat ook bekend onder de noemer *statistical parity* en *disparate impact*. Zie bijv. Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). *Building classifiers with independency constraints*. In 2009 IEEE international conference on data mining workshops (pp. 13-18). IEEE.

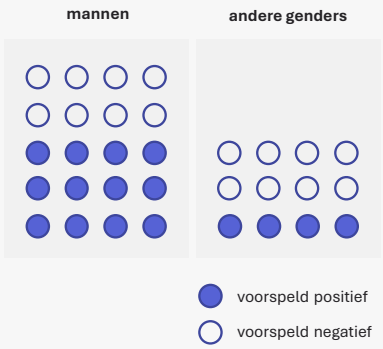
Voorbeeld: demografische pariteit in cv-selectie

Een organisatie gebruikt een algoritme voor geautomatiseerde cv-selectie. De organisatie wil meten in hoeverre de inzet van het algoritme leidt tot onderscheid op basis van gender. Om een vermoeden van onderscheid te meten in de cv-selectie, wordt de groep sollicitanten opgedeeld op basis van gender. Vervolgens berekenen zij voor iedere gendergroep welke proportie van sollicitanten met dat gender is geselecteerd.

Tabel 1 Voorbeeld van een uitgesplitste analyse.

Groep	**Selectieratio** kans dat een sollicitant wordt geselecteerd
Man	0.6
Vrouw	0.33
X	0.33

In dit geval blijkt de selectieratio voor mannen 0.6 te zijn en de selectieratio voor andere genders 0.33. Er is dus geen sprake van demografische pariteit: de cv's van mannen worden immers bijna twee keer zo vaak geselecteerd als dat van vrouwen.



Figuur 1 *De selectieratio is gelijk aan het aantal voorspeld positieven (i.e. geselecteerden) gedeeld door het totaal aantal personen in een groep. In dit voorbeeld is de selectieratio 0.6 voor mannen en 0.33 voor andere genders.*

In de praktijk zijn groepsstatistieken vrijwel nooit exact gelijk. Technisch gezien, is er dan dus geen sprake van pariteit. In de wetenschappelijke literatuur kwantificeren *fairness metrics* daarom over het algemeen in hoeverre pariteit wordt geschonden. Verschillen worden vaak geaggregeerd tot één fairness score die ongelijkheid tussen groepen samenvat, zoals bijvoorbeeld het maximumverschil (wat is het grootste verschil in *false positive rates* tussen mannen, vrouwen, en andere genders?) of minimumratio (wat is het kleinste ratio van selectiepercentages tussen de verschillende leeftijdsgroepen?). Op deze manier kunnen verschillende algoritmes snel met elkaar worden vergeleken. Zulke geaggregeerde scores zijn vaak moeilijk te interpreteren. In de praktijk wordt daarom vaker gebruik gemaakt van (visualisaties van) ruwe groepsstatistieken.

Voorbeeld: mate van demografische dispariteit in cv-selectie

De organisatie wil graag kwantificeren hoe ver het cv-selectie algoritme is verwijderd van volledige demografische pariteit. Hiervoor worden de selectieratio's vergeleken tussen alle mogelijke combinaties van groepen.

Maximale verschil

Voor iedere combinatie van groepen in **Tabel 1**, berekenen we het (absolute) verschil in selectieratio's.

- **Man / Vrouw**: | 0.6 – 0.33 | = 0.27
- **Man / X**: | 0.6 – 0.33 | = 0.27
- **Vrouw / X**: | 0.33 – 0.33 | = 0

Het maximale verschil tussen groepen is in dit geval dus 0.27, namelijk tussen mannen en de andere genders.

Minimale ratio

Relatieve verschillen kunnen ook worden gekwantificeerd door middel van het (minimum) ratio van selectieratio's:

- **Man / Vrouw** : 0.6 / 0.33 = 1.81 en **Vrouw / Man** : 0.33 / 0.6 = 0.55
- **Man / X** : 0.6 / 0.33 = 1.81 en **Vrouw / X** : 0.33 / 0.6 = 0.55
- **Vrouw / X** : 0.33 / 0.33 = 1 en **Vrouw / Man** : 0.33 / 0.33 = 1

De minimale ratio tussen groepen is 0.55, tussen mannen en de andere genders.

Bij de beoordeling van een geobserveerd verschil, zijn zowel statistische zekerheid als praktische en normatieve relevantie belangrijk. Wanneer het aantal observaties per groep zeer klein is, zijn berekende groepsstatistieken minder betrouwbaar. Wanneer een dataset bijvoorbeeld slechts 10 cv's van vrouwen bevat, heeft één vrouw meer of minder in de selectie grote invloed op het berekende selectiepercentage, in vergelijking met een dataset welke 100 cv's van vrouwen omvat. Om te beoordelen in hoeverre het geobserveerd verschil mogelijk is te wijten aan toeval, kan gebruik worden gemaakt van statistische hypothesetoetsen. Statistische significantie alleen is echter nooit voldoende om te spreken van een problematisch verschil. Als een dataset duizenden voorbeelden bevat voor alle genders, kan zelfs een zeer klein verschil in selectieratio's van 0.001 statistisch significant zijn. De vraag wordt dan dus in hoeverre het geobserveerde verschil praktisch en normatief relevant is.

Bij een vermoeden van onderscheid is het nuttig om te kijken naar mogelijke redenen. Zo zou een werkgever kunnen aangeven dat er inderdaad relatief meer mannen zijn geselecteerd, maar dat dit kan worden verklaard door het feit dat mannen die bij het bedrijf solliciteerden doorgaans meer werkervaring hadden. Dergelijke factoren kunnen worden meegenomen in een *group fairness* analyse door te conditioneren op de verklarende factor. Er wordt dan gekeken of het verschil te verklaren is door het verschil in jaren werkervaring of dat er nog andere factoren meespelen. Deze variant van *group fairness* wordt ook wel *conditional group fairness* ("conditionele groepseerlijkheid") genoemd. Mogelijke verklaringen kunnen ook een rol spelen bij de objectieve rechtvaardiging, verder toegelicht in 3.2.

Voorbeeld: conditionele demografische pariteit in cv-selectie

De organisatie wil onderzoeken in hoeverre de verschillen tussen genders kunnen worden verklaard door de werkervaring van sollicitanten. De groep sollicitanten wordt opgedeeld op basis van jaren werkervaring (*weinig* of *veel*) en vervolgens worden de selectiepercentages tussen mannen, vrouwen, en overige genders vergeleken binnen elk van de condities. Personen met meer werkervaring hebben een hogere kans om geselecteerd te worden (**Tabel 2**).

Dit verschil verklaart voor een groot deel het verschil in selectieratio tussen mannen en andere genders: binnen groepen met vergelijkbare werkervaring is het verschil in selectiekans tussen genders veel kleiner. Maar ook na correctie voor werkervaring, zien we verschillen tussen genders die de moeite waard zijn om verder te onderzoeken, bijvoorbeeld een verschil van 10 procentpunt tussen vrouwen en andere genders met weinig werkervaring.

Tabel 2 Voorbeeld van de standaardformulering van een uitgesplitste analyse.

Conditie	**Groep**	**Selectieratio** kans dat een sollicitant wordt geselecteerd
Weinig werkervaring	**Man**	0.4
	Vrouw	0.3
	X	0.4
Veel werkervaring	**Man**	0.8
	Vrouw	0.8
	X	0.7

De standaardmethode in de computerwetenschappelijke literatuur vergelijkt uitkomsten tussen groepen, mogelijk geconditioneerd op relevante categorieën (**Tabel 2**). Deze berekeningen kunnen enkel worden gemaakt wanneer voor ieder individu bekend is of deze is geselecteerd of niet. In een alternatieve formulering wordt de distributie van beschermde groepen vergeleken tussen verschillende populaties (**Tabel 3**).[29]

Deze alternatieve formulering maakt het mogelijk om de distributie van beschermde groepen in de selectie te vergelijken met een referentiegroep. Zo kunnen genderproporties van werknemers in een bepaalde organisatie worden vergeleken met de genderproporties in de gehele samenleving of in een bepaalde sector. Dergelijke vergelijkingen kunnen handig zijn wanneer er geen toegang is tot de negatieve labels (zoals in het voorbeeld). Ook vanuit normatief oogpunt is de groep waaruit in de praktijk wordt geselecteerd niet altijd een geschikte referentiepopulatie. Bijvoorbeeld wanneer de genderverhoudingen binnen een groep sollicitanten niet representatief is voor de genderverhoudingen binnen de sector.

Wanneer een uitgesplitste analyse wordt uitgevoerd binnen één populatie, is de formulering zoals in **Tabel 3** wiskundig gezien equivalent aan de vergelijking van selectiepercentages zoals in **Tabel 1**.

[29] De mogelijkheid tot het definiëren van een specifieke referentiegroep maakt dat de alternatieve formulering nauwer overeenkomt met de discriminatietoets. Een nadeel van de alternatieve formulering is dat deze lastiger is om te interpreteren, vanwege de onderlinge afhankelijkheid van proporties. Zie bijvoorbeeld Wachter et al. (2021) *Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI*. Computer Law & Security Review, 41, 105567 en B. Hekkelman, M. Kattenberg, B. Scheer. *Rechtvaardige Algoritmes* (2023). Centraal Planbureau.

Voorbeeld: alternatieve formulering conditionele demografische pariteit in *numerus fixus*

Een journalist wil onderzoeken of er mogelijk sprake is van onderscheid bij de selectie van studenten bij universitaire opleidingen met een beperkte hoeveelheid studentplekken (numerus fixus). De journalist weet niet welke personen zich hebben aangemeld voor de opleidingen informatica en rechten, maar heeft wel achterhaald welke personen zijn begonnen aan de opleiding.

Het is voor de journalist niet mogelijk de selectiepercentages in **Tabel 1** te berekenen: ze heeft wel toegang tot de positieve labels (wie is geselecteerd?), maar niet de negatieve labels (wie is afgewezen?). Ze kan wel de genderproporties in de groep "geselecteerde studenten" vergelijken met de genderproporties in een andere (relevante) populatie: vwo-leerlingen met een relevant profiel. Wanneer de distributie van genders verschilt tussen de geselecteerde studenten en de poule van potentiële studenten (vwo-studenten met een relevant profiel), wijst dit op mogelijk onderscheid op basis van gender.

Tabel 3 Voorbeeld van een alternatieve formulering van conditionele demografische pariteit.

Conditie	**Geselecteerde studenten**			**VWO-studenten met relevant profiel**		
	Man	**Vrouw**	**X**	**Man**	**Vrouw**	**X**
Informatica	60%	35%	5%	50%	45%	5%
Rechten	35%	65%	0%	40%	55%	5%

Misclassificatieratio's

Een factor die vaak wordt aangehaald als mogelijke rechtvaardiging voor verschillen in selectieratio's tussen groepen is de doelvariabele zelf: een groep is vaker geselecteerd, maar dat was ook vaker terecht. Zo zou een werkgever kunnen stellen dat mannen vaker worden geselecteerd dan vrouwen, maar dat dit verschil samenhangt met het feit dat mannen ook vaker gekwalificeerd zijn. In zo'n geval kan worden gekeken naar conditionele demografische pariteit, met als condities "gekwalificeerd" en "niet-gekwalificeerd".

In de praktijk wordt in plaats van conditionele demografische pariteit meestal direct gebruik gemaakt van misclassificatieratio's gebaseerd op de *confusion matrix* (**Tabel 4**). Misclassificatieratio's geven weer wat de kans is op misclassificatie. Hiervoor wordt voornamelijk gekeken naar *valspositieven* (voorbeelden waarvoor het algoritme een positief label voorspelt, maar het daadwerkelijke label is negatief) en *valsnegatieven* (voorbeelden waarvoor het algoritme een negatief label voorspelt, maar het daadwerkelijke label is positief).

Tabel 4 De *confusion matrix* (verwarringsmatrix) laat zien in hoeverre het algoritme in staat is ieder mogelijk label correct te classificeren.

		voorspelde label	
		positief	**negatief**
daadwerkelijke label	**positief**	echt positief	valsnegatief
	negatief	valspositief	echt negatief

Een veelgebruikte metriek is de *false positive rate*: de proportie van echte negatieven die worden gemisclassificeerd als positief. In het geval van risicoprofilering, geeft de *false positive* rate weer wat de kans is dat een niet-frauduleuze

aanvraag wordt geselecteerd voor controle.[30] Een andere metriek is de *false negative rate*: de proportie van daadwerkelijke positieven die worden gemisclassificeerd als negatief. In het geval van een algoritme dat wordt gebruikt om bepaalde hulpmiddelen te verdelen, kwantificeert de *false negative rate* bijvoorbeeld de kans dat een hulpbehoevende burger geen hulp wordt aangeboden.[31]

Voorbeeld: conditionele demografische pariteit in risicoprofilering

Een organisatie maakt gebruik van risicoprofilering voor normovertredingen. Het blijkt dat de selectieratio voor mannen hoger ligt dan de selectieratio voor vrouwen.

Tabel 5 Selectieratio voor mannen en vrouwen.

Groep	**Selectieratio** kans dat een persoon wordt geselecteerd voor controle
Man	0.4
Vrouw	0.2

De organisatie vermoedt dat het verschil in selectieratio deels te verklaren is doordat mannen vaker een normovertreding begaan dan vrouwen. Daarom herhalen ze de analyse, nu geconditioneerd op het al dan niet begaan van een normovertreding. Hieruit blijkt dat mannen die een normovertreding hebben begaan iets vaker (correct) worden geselecteerd dan vrouwen die een normovertreding hebben begaan.[32] Voor alle personen die geen normovertreding hebben begaan, is de kans op (incorrecte) selectie 0.1 voor zowel mannen als vrouwen.[33]

Tabel 6 Selectieratio geconditioneerd op label.

Conditie	**Groep**	**Selectieratio** kans dat een persoon wordt geselecteerd voor controle
Normovertreding begaan	**Man**	0.6
	Vrouw	0.5
Geen normovertreding begaan	**Man**	0.1
	Vrouw	0.1

[30] De *false negative* rate is direct gerelateerd aan de *true negative rate*: de proportie van alle daadwerkelijke negatieven die correct als negatief wordt geclassificeerd.
[31] De *false positive rate* is direct gerelateerd aan de *true positive rate*: de proportie van daadwerkelijke positieven die correct worden geclassificeerd als positief.
[32] Dit is equivalent aan de *true positive rate*.
[33] Dit is equivalent aan de *false positive rate*.

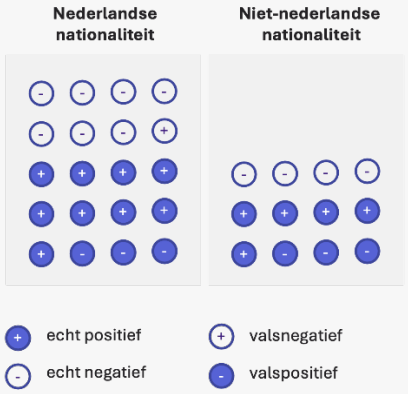
Voorbeeld: gelijke misclassificatieratio's in toeslagen

Een overheidsorganisatie gebruikt een algoritme om personen te identificeren die geen toeslag ontvangen, terwijl zij daar mogelijk wel recht op hebben, zodat de organisatie hen hierover kan informeren. Wanneer de kans dat iemand onterecht geen toeslag ontvangt boven een bepaalde grenswaarde komt, beoordeelt een medewerker of deze persoon inderdaad onterecht geen toeslag ontvangen.

De organisatie vindt het met name belangrijk dat de kans om niet te worden geselecteerd terwijl iemand wel recht heeft op toeslag (de *false negative rate*) gelijk is voor personen met een Nederlandse nationaliteit en personen met een andere nationaliteit. Daarnaast wil ze graag weten of de kans om wél te worden geselecteerd voor dossieronderzoek, terwijl iemand geen recht heeft op toeslag, vergelijkbaar is (de *false positive rate*). Een uitgesplitste analyse (**Tabel 7**) laat zien in hoeverre de misclassificatieratio's verschillen tussen de twee groepen.

Tabel 7 Voorbeeld van een uitgesplitste analyse met misclassificatieratio's.

Groep	**False negative rate** kans dat een persoon *niet* wordt geselecteerd voor dossieronderzoek, terwijl deze *wel* recht had op een toeslag	**False positive rate** kans dat een persoon *wel* wordt geselecteerd voor dossieronderzoek, terwijl deze *geen* recht had op een toeslag
Nederlandse nationaliteit	0.4	0.1
Niet-Nederlandse nationaliteit	0.3	0.0



Figuur 2 De *false positive rate* is in dit voorbeeld 3/(3+7)=0.30 voor personen met een Nederlandse nationaliteit en 3/(3+4)=0.43 voor personen met een niet-Nederlandse nationaliteit. De *false negative rate* is respectievelijk 1/(1+9)=0.10 en 0/(0+5)=0.00 voor personen met een Nederlandse en niet-Nederlandse nationaliteit.

Een uitdaging bij het berekenen van *group fairness* metrics is statistische betrouwbaarheid. Sommige organisaties hebben slechts toegang tot kleine datasets of hebben te maken met discriminatiegronden met een groot aantal mogelijke categorieën (bijvoorbeeld nationaliteit). In dergelijke gevallen is er niet altijd genoeg data beschikbaar om een statistisch betrouwbare inschatting te maken van de verschillen tussen groepen.[34] Sommige geïnterviewden geven daarom aan behoefte te hebben aan richtlijnen hoe om te gaan met statistische toetsen bij zulke kleine datasets of een groot aantal vergelijkingen.

[34] De benodigde steekproefgrootte kan bijvoorbeeld worden bepaald door middel van een statistische *power* analyse.

3.2. Overlap tussen *group fairness* en de discriminatietoets

In Nederland is zowel direct als indirect onderscheid op grond van beschermde persoonskenmerken verboden, tenzij er een objectieve rechtvaardiging bestaat. In de context van algoritmes en kunstmatige intelligentie, kan sprake zijn van direct onderscheid als een criterium dat valt onder een discriminatiegrond expliciet wordt gebruikt als criterium in een algoritme. Denk bijvoorbeeld aan een wervingsalgoritme dat een variabele voor "gender" gebruikt als inputvariabele. Het uitsluiten van beschermde persoonskenmerken als inputvariabele wordt in grote lijnen gezien als voldoende om directe discriminatie te voorkomen.[35] Met betrekking tot indirect onderscheid, is het verhaal complexer. Over het algemeen wordt aangenomen dat *conditional group fairness* het meest overeenkomt met de indirecte discriminatietoets van het Hof van Justitie van de Europese Unie (HvJ-EU)[36] – mits deze juist is ingericht. In de rest van dit hoofdstuk wordt dieper ingegaan op de overeenkomsten tussen *group fairness* en de discriminatietoets.

Group Fairness Analyse

Een *group fairness* analyse, zoals toegelicht in Sectie 3.1, heeft als doel de vraag te beantwoorden: hoe verschillen de uitkomsten van het model per groep? Om de analyse uit te voeren, moet in ieder geval een keuze gemaakt worden voor een bepaalde statistiek, de persoonskenmerk(en) die worden vergeleken, de populatie waarover de statistieken worden berekend, en voor welke verklarende variabelen wordt geconditioneerd.

Zoals uitgelegd in Sectie 2.1, is er sprake van indirecte discriminatie wanneer een bepaling, maatstaf, of praktijk (bijvoorbeeld algoritmisch-gestuurd beleid) personen die behoren tot een beschermde groep in het bijzonder benadeelt ten op zichtte van personen in een vergelijkbare situatie die niet behoren tot een beschermde groep. Ook de indirecte discriminatietoets is dus, in de kern, een vergelijking van uitkomsten tussen groepen. In hoeverre een *group fairness* analyse overeenkomt met de discriminatietoets, hangt af van hoe de vergelijking is ingericht.

In ieder geval de volgende vier vragen zijn van belang:

1) Welke **uitkomsten** worden vergeleken en in hoeverre representeren deze een juridisch relevant een nadeel?
2) Welke **persoonskenmerken** worden meegenomen en zijn deze ook beschermd onder de relevante wetgeving?
3) Binnen welke **populatie** worden de uitkomsten vergeleken en is dit een juridisch relevante referentiepopulatie?
4) Bij welke **grenswaarde** wordt (niet) ingegrepen en in hoeverre voldoet deze keuze aan een juridische objectieve rechtvaardiging?

Dit zijn stuk voor stuk contextuele en normatieve vraagstukken. De manier waarop ze worden beantwoord bepaalt in grote mate in hoeverre de analyse strookt met de discriminatietoets.

Welke uitkomsten moeten worden vergeleken?

Eén van de sleutelvragen bij de vraag of sprake is van indirecte discriminatie, is of een selectie-instrument personen met een bepaald beschermd persoonskenmerk 'in het bijzonder' benadeelt. Een statistiek, zoals de selectieratio, kan dienen als basis voor de toets als de statistiek de nadelige gevolgen van het algoritme voor een individu kwantificeert.

[35] Recent wetenschappelijk onderzoek beargumenteert dat ook sommige vormen van proxydiscriminatie onder het verbod op directe discriminatie kunnen vallen, zie bijvoorbeeld *Directly discriminatory algorithms*. Adams-Prassl et al., 2023, The Modern Law Review, 86(1), 144-175; Weerts et al. (2024). *Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms*. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1850-1860).

[36] Zie bijvoorbeeld Wachter et al. (2021) *Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI*. Computer Law & Security Review, 41, 105567 en Weerts et al. (2023). *Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree*. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (pp. 805-816).

Voorbeelden van uitkomsten

- Een CV-matchingtool selecteert relatief meer mannelijke dan vrouwelijke sollicitanten.
- Een fraudedetectiemodel geeft een hogere kans op valspositief voor klanten met een niet-Nederlandse nationaliteit die geen overtreding hebben begaan dan voor klanten met een Nederlandse nationaliteit die geen overtreding hebben begaan.
- De gezichtsdetectiemodule in antispieksoftware is minder accuraat voor personen met een donkere huidskleur, waardoor het langer duurt voordat zij aan hun toets kunnen beginnen.

Benadeling kan in de context van de discriminatietoets breed worden geïnterpreteerd. Zo is een selectieve controle door middel van risicoprofilering per definitie 'nadelig', omdat het leidt tot een grote kans om een sanctie te krijgen – ook als de controle niet daadwerkelijk tot een sanctie leidt.[37] Ook een verhoogde (of verlaagde) kans om geselecteerd te worden kan in de juridische zin worden gezien als onderscheid – ook als dit niet (altijd) leidt tot selectie. In fraudedetectie kan bijvoorbeeld worden besloten om alle personen met een risicoscore van 0.6 of hoger te controleren. Wanneer klanten met een niet-Nederlandse nationaliteit gemiddeld een hogere risicoscore hebben ten op zichtte van klanten met een Nederlandse nationaliteit, is dit juridisch gezien een vorm van onderscheid; zelfs als de drempelwaarde voor controle niet (altijd) gehaald wordt.[38]

Niet alle statistieken zijn even gemakkelijk in verband te brengen met een nadelig gevolg voor een individu. Sommige populaire prestatie-indicatoren, zoals kalibratie[39] en de *Area Under the ROC-Curve* (AUC)[40], meten een bepaalde dimensie van de kwaliteit van een risicoscore, maar geven geen (direct) inzicht in de (kans op) selectie.[41] Dit maakt deze metrieken over het algemeen minder bruikbaar voor de discriminatietoets dan statistieken directer gerelateerd aan classificatie, zoals selectieratio's, misclassificatieratios, of gemiddelden van risicoscores.

Uit de interviews blijkt dat sommige organisaties bij het kiezen van een metriek gebruik maken van een hulpmiddel, zoals de Aequitas *fairness tree*.[42] Hierbij is voorzichtigheid geboden: het is essentieel dat een organisatie kan verantwoorden waaróm een bepaalde metriek passend is voor dit specifieke algoritme. In andere organisaties geven respondenten aan ervoor te kiezen om de nadelen eerst te definiëren in "Jip-en-Janneke-taal" en dit daarna pas te vertalen naar een kwantitatieve metriek. In de praktijk wordt vrijwel altijd gekozen voor een variant van selectieratio's of misclassificatieratio's, aldus de geïnterviewden. Veel geïnterviewde organisaties berekenen zowel representatie- als foutgebaseerde statistieken.

Een belangrijke uitdaging bij het berekenen van foutgebaseerde statistieken is een gebrek aan (volledig) gelabelde data. Om misclassificatieratio's accuraat te kunnen schatten, is de trainingsdata idealiter gelabeld door middel van een aselecte steekproef, waarin ieder datapunt een gelijke kans heeft om gelabeld te worden. Dit is in de praktijk niet altijd het

[37] Zie ook Hoofdstuk 3 *Toetsingskader risicoprofilering,* College voor de Rechten van de Mens, 2025.
[38] Zie bijvoorbeeld Weerts et al. (2024). *Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms*. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1850-1860).
[39] Kalibratie meet in hoeverre de risicoscores overeenkomen met daadwerkelijke frequenties van de uitkomst. Voor een gekalibreerd algoritme geldt bijvoorbeeld dat van alle personen die een risicoscore van 0.3 worden toegewezen, de proportie personen met een positief label precies 0.3 is (en niet hoger of lager).
[40] De Area Under the ROC Curve (AUC) meet het onderscheidend vermogen van de risicoscores: in hoeverre is het model in staat onderscheid te maken tussen de verschillende labels van de *doelvariabele*. Een ROC-curve zet de *false positive rate* van een algoritme af tegen de *true positive rate* voor alle mogelijke drempelwaarden.
[41] Neem bijvoorbeeld een fraudedetectiealgoritme. Wanneer het algoritme een lagere AUC heeft voor personen met Nederlandse nationaliteit, ten op zichtte van personen met een andere nationaliteit, betekent dit dat het algoritme over het algemeen minder goed in staat is om onderscheid te maken tussen de labels "fraude gedetecteerd" en "geen fraude gedetecteerd" voor Nederlanders ten opzichte van niet-Nederlanders. Dit zegt echter niet dat personen met een niet-Nederlandse nationaliteit gemiddeld een hogere of juist lagere risicoscore krijgen. Ook laat AUC niet zien of met de gekozen drempelwaarde een grotere kans is op een valspositief of op een valsnegatief: de AUC wordt namelijk berekend op basis van alle mogelijke drempelwaardes. Dit betekent overigens niet dat uitgesplitste analyses met metrics zoals AUC en kalibratie niet nuttig zijn. Zo kunnen dergelijke analyses cruciaal zijn om de oorzaak van onderscheid in (mis)classificaties te achterhalen. Hier zal in Hoofdstuk 4 uitgebreider bij worden stilgestaan.
[42] De *fairness tree* is een beslisboom waar, op basis van enkele vragen, een fairness metriek kan worden geselecteerd (bijvoorbeeld: zijn interventies straffend of assisterend? Zijn de labels betrouwbaar?) Saleiro et al. (2018). *Aequitas: A bias and fairness audit toolkit*. arXiv preprint arXiv:1811.05577.

geval. Zoals eerder genoemd, beïnvloedt bestaand beleid vaak welke datapunten worden gelabeld. In fraudedetectie zijn bijvoorbeeld enkel labels bekend van personen die in het verleden zijn gecontroleerd. Zonder verdere aannames, is het dan niet mogelijk om betrouwbare schattingen te geven van de *false positive rate* en *false negative rate*. Daarnaast geven de geïnterviewde data professionals aan dat het niet altijd wenselijk of haalbaar is om gebruik te maken van een aselecte steekproef, bijvoorbeeld omdat de kosten van een valspositief of valsnegatief erg hoog zijn. Het is bijvoorbeeld niet ethisch verantwoord om een lening te verstrekken aan een persoon die deze lening zeer waarschijnlijk niet kan terugbetalen. Ook bedrijfsorganisatorisch vinden respondenten een aselecte steekproef soms moeilijk te verantwoorden, bijvoorbeeld omdat controleurs het idee hebben dat hun inzet weinig zinvol is wanneer het bij voorbaat al bijna zeker is dat er geen sprake is van een overtreding. Daarnaast is in sommige gevallen het aantal echtpositieven zo laag, dat de steekproef zeer groot moet zijn om überhaupt positieven te identificeren. Hier is niet altijd voldoende capaciteit voor beschikbaar.

Een bijkomende uitdaging bij het bepalen van foutgebaseerde metrieken is dat er soms vertraging zit in het beschikbaar komen van het daadwerkelijke ("*ground-truth*") label. Denk bijvoorbeeld aan een algoritme dat personen selecteert die mogelijk extra hulp nodig hebben bij het vinden van een nieuwe baan: het wordt pas na verloop van tijd duidelijk of de geselecteerde personen binnen bepaalde tijd een baan hebben gevonden.

Tot slot is het van belang alert te zijn dat een algoritme vaak slechts één stap is binnen een meerstaps besluitvormingsproces. In publieke organisaties worden selectiealgoritmes bijvoorbeeld vaak ingezet ter ondersteuning van beslismedewerkers. Wanneer een cv-matchingtool wordt gebruikt ter ondersteuning van een recruiter, is het bijvoorbeeld niet enkel van belang wat de output van het algoritme is, maar ook hoe de recruiter deze output gebruikt om tot een uiteindelijke selectie te komen. Cognitieve *biases* zoals automatiseringsbias (*automation bias*) en contextuele factoren zoals een hoge werkdruk, kunnen ervoor zorgen dat zaakmedewerkers te vaak meegaan in het besluit van de uitkomst van het algoritme. In andere gevallen kunnen zaakbehandelaars juist vooringenomenheid in het proces introduceren. Sommige respondenten geven daarom aan graag breder te willen kijken dan het algoritme en ook andere stappen in het besluitvormingsproces te toetsen op onderscheid.

Welke (beschermde) persoonskenmerken zijn relevant?

Een tweede element in de analyse is het persoonskenmerk dat bepaalt tussen welke groepen een vergelijking wordt gemaakt. Het Nederlandse non-discriminatierecht verbiedt discriminatie op basis van verscheidene discriminatiegronden. De Awgb verbiedt bijvoorbeeld onderscheid op grond van op grond van godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat. Er kan alleen een beroep op deze wetgeving worden gedaan als er sprake is van discriminatie op een specifiek terrein dat onder de wetgeving valt, zoals arbeid, goederen en diensten, het vrije beroep, en sociale bescherming. Naast de Awgb, zijn er ook enkele wetten die van toepassing zijn in specifieke contexten, zoals de Wet gelijke behandeling op grond van leeftijd bij arbeid, welke bescherming biedt aan mensen die gediscrimineerd worden vanwege hun leeftijd bij arbeid.

Niet alle mogelijk relevante maatschappelijk factoren zijn discriminatiegronden in alle gebieden in de Awgb. Zo is bijvoorbeeld opleidingsniveau (nog) geen discriminatiegrond in de Awgb[43], maar kan een organisatie het toch belangrijk vinden om te zorgen dat de inzet van een algoritme niet leidt tot onderscheid op basis van dit persoonskenmerk. De non-discriminatiebepalingen van de Nederlandse Grondwet en internationale verdragsbepalingen, zoals het Europees Verdrag van de Rechten van de Mens (EVRM) kennen een bredere reikwijdte waarbij geen sprake van een limitatieve grondenlijst. Op basis van deze wetgeving kunnen ook andere gronden kritisch worden getoetst, met name als deze raken aan iemands identiteit.

Sommige dataprofessionals geven aan dat het lastig is om te bepalen op welke persoonskenmerken moet worden getoetst. Een veelvoorkomende uitdaging is een gebrek aan toegang tot (geschikte) variabelen die beschermde persoonskenmerken meten. Data over discriminatiegronden zijn persoonsgegevens onder de Algemene Verordening

[43] Zie College voor de Rechten van de Mens (2024). *Wetgevingsadvies mogelijkheid tot opname van de grond 'opleidingsniveau' in de Algemene wet gelijke behandeling*.

Gegevensbescherming (AVG). Voor de verwerking van persoonsgegevens is altijd een wettelijke grondslag nodig. Sommige kenmerken, zoals geslacht en leeftijd, zijn redelijk vrij beschikbaar en hier wordt dan ook vaak op getoetst. Veel data over discriminatiegronden, zoals bijvoorbeeld migratieachtergrond en seksuele geaardheid, vallen echter onder de categorie bijzondere persoonsgegevens. Het verwerken van deze gegevens is in principe verboden, tenzij een specifieke uitzondering geldt.[44] Ook wanneer organisaties persoonsgegevens al verzamelen voor andere doeleinden, kan het soms jaren duren voordat het duidelijk wordt of die data ook gebruikt mag worden voor een discriminatietoets.

Uitzonderingen in de AVG en de discriminatietoets

De meest algemeen toepasbare uitzondering in de AVG betreft uitdrukkelijke toestemming van de betrokkene. Uitdrukkelijke toestemming moet onder strikte omstandigheden moet worden verzameld, zo moet deze bijvoorbeeld vrijelijk gegeven, specifiek, en ondubbelzinnig zijn. Om van deze uitzondering gebruik te maken, moeten betrokkenen dus expliciet toestemming geven voor het gebruik van een bijzonder persoonsgegeven in de context van een discriminatietoets.

Een tweede mogelijk relevante uitzondering, beschreven in Artikel 9(1)(g), is noodzakelijkheid voor een zwaarwegend algemeen belang. In de wetenschappelijke literatuur hebben sommige onderzoekers gesuggereerd dat het meten en voorkomen van discriminatie als een dergelijk zwaarwegend algemeen belang zou kunnen worden geïnterpreteerd.[45] Echter, deze uitzondering geldt alleen als hiervoor in de (nationale) wet een rechtsbasis is gecreëerd. In Nederland bestaat op dit moment geen wet die een dergelijke algemeen toepasbare uitzondering verzorgt in het kader van de discriminatietoets. Wel bestaan specifieke uitzonderingen voor organisaties met een wettelijke bevoegdheid voor het verwerken van bijzondere persoonsgegevens in het uitvoeren van hun wettelijke taak. In Nederland regelt bijvoorbeeld de Wet Centraal Bureau voor de Statistiek (CBS) onder bepaalde omstandigheden een uitzondering op het verbod op het verwerken van bijzondere persoonsgegevens in het kader van onderzoek. Samenwerking met het CBS zou daarom, onder strikte voorwaarden, overheidsorganisaties in staat kunnen stellen statistisch onderzoek uit te voeren ten behoeve van het toetsen van mogelijk onderscheid met behulp van bijzondere persoonsgegevens als deze niet (met de juiste doelbinding en uitzonderingsgrond) voorhanden zijn binnen de organisatie zelf.

Tot slot heeft de Europese Commissie in de Artificiële Intelligentie Verordening (AIV) een bepaling opgenomen die organisaties onder strikte omstandigheden in staat zou stellen bijzondere persoonsgegevens te gebruiken om discriminatie te toetsen.[46] Deze uitzondering geldt alleen voor AI-systemen die onder de AIV vallen, zogenoemde “hoog-risico” systemen. Jurisprudentie zal moeten uitwijzen welke AI-systemen precies onder de wet vallen en onder welke omstandigheden de uitzondering zal gelden.

Uit de interviews blijkt dat sommige organisaties hebben geëxperimenteerd met proxy’s voor beschermde kenmerken, bijvoorbeeld op basis van publiekelijk beschikbare data van het CBS. Denk bijvoorbeeld aan het gebruik van postcode als proxy voor migratieachtergrond: wanneer selectie sterk verschilt per postcodegebied, kan dat een teken zijn van mogelijk onderscheid. Daarnaast noemen sommige professionals als mogelijke oplossing het verkrijgen van toegang tot microdata van het CBS. Hier zijn bepaalde variabelen gerelateerd aan bijvoorbeeld migratieachtergrond en sociaaleconomische status geregistreerd op persoonsniveau. Deze data zijn bijvoorbeeld gebruikt in de audit van het risicoprofileringsalgoritme van controle uitwonendenbeurs van DUO.[47] Sommige data professionals ervaren hierbij echter wel een zekere terughoudendheid: hoewel zij het belangrijk vinden om te toetsen op *bias*, is het verzamelen van dergelijke gevoelige gegevens ook vatbaar voor datalekken en misbruik. Een mogelijke middenweg wordt geboden door de Selectiviteitsscan, ontwikkeld door onderzoekers van het Centraal Planbureau. Deze tool stelt overheidsorganisaties in

[44] Artikel 9(1) van de AVG.
[45] M. van Bekkum & F.Z. Borgesius (2023). *Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?*. Computer Law & Security Review, 48, 105770.
[46] Artikel 10(5) van de AI-verordening
[47] Algorithm Audit (2024). *Addendum vooringenomenheid voorkomen*.

staat om algoritmes, via een derde partij, te toetsen zonder toegang te hebben tot gevoelige persoonsgegevens.[48] Daarnaast geven sommige respondenten aan dat ook CBS-microdata niet alle mogelijk relevante discriminatiegronden bevatten. Zo zijn bijvoorbeeld religie en seksuele geaardheid niet beschikbaar. Voor het toetsen op deze gronden biedt CBS-data dus geen oplossing.

Wat is de juiste referentiepopulatie?

Een belangrijke dimensie bij het bepalen van indirect onderscheid is de referentiegroep: personen die niet behoren tot de beschermde groep die mogelijk wordt benadeeld, maar wel in een vergelijkbare situatie verkeren. De vergelijking tussen de beschermde groep en de referentiegroep bepaalt of er juridisch gezien sprake is van onderscheid. De identificatie van de juiste referentiepopulatie is sterk contextafhankelijk en daarom mede normatief. Het bepalen van deze groep vereist dus een grote mate van zorgvuldigheid. Met betrekking tot mogelijke discriminatie in fraudecontroles in een bepaalde sector van het midden- en kleinbedrijf zijn bijvoorbeeld de (etnische) verhoudingen binnen die sector van belang en niet de etnische verhoudingen in de Nederlandse bevolking als geheel. Een praktische uitdaging is dat een organisatie niet altijd beschikt over alle relevante data. Zo kan een bedrijf bijvoorbeeld toegang hebben tot een dataset van personen die hebben gesolliciteerd, maar heeft het wellicht geen zicht op de verhoudingen binnen de populatie geschikte kandidaten die *niet* hebben gesolliciteerd. Uit de interviews blijkt niet dat geïnterviewden uitdagingen ondervinden bij het bepalen van een referentiepopulatie.

Bij welke grenswaarde moet (niet) worden ingegrepen?

In de praktijk komt het vrijwel nooit voor dat selectieratio's of misclassificatieratio's exact gelijk zijn tussen verschillende groepen. Juridisch gesproken is er dan sprake van onderscheid. Zoals besproken in Hoofdstuk 2 is onderscheid in principe verboden, tenzij er sprake is van een objectieve rechtvaardiging. In EU-wetgeving wordt de objectieve rechtvaardigingstoets ingevuld door het proportionaliteitsprincipe: het beleid moet een legitiem doel nastreven en hiervoor een geschikt en noodzakelijk middel zijn.

Legitiem doel

De inzet van een algoritme in een besluitvormingsproces moet een legitiem doel nastreven. Wat geldt als een legitiem doel? Ten eerste moet het doel voorzien in een daadwerkelijke behoefte en nooit op zichzelf genomen inherent discriminerend zijn. Ook moet het doel niet in strijd zijn met bestaande wet- en regelgeving. Een legitieme doelstelling kan redelijk breed worden geïnterpreteerd, denk bijvoorbeeld aan het verbeteren van operationele efficiëntie of het uitvoeren van een wettelijke taak. De legitimiteit van de doelstelling zal daarom in de meeste gevallen geen struikelblok vormen bij de inzet van algoritmes. Wel is het van belang dat organisaties expliciet documenteren welk doel precies wordt nagestreefd.

Geschiktheidstoets

In de geschiktheidstoets moet een organisatie aantonen dat de selectiecriteria die gebruikt worden in het algoritme, hetzij op zichzelf, hetzij samen met andere criteria, op effectieve wijze bijdragen aan de verwezenlijking van het nagestreefde legitieme doel. De geschiktheidstoets is in feite een empirische toets. Neem bijvoorbeeld de inzet van risicoprofilering in handhaving. Een legitiem doel zou hier kunnen zijn 'fraude opsporen'. Als de inzet van een risicoprofileringsalgoritme leidt tot een hoger aantal veroordeelde verdachten ten op zichtte van bestaand beleid, zou een organisatie kunnen betogen dat het algoritme effectief bijdraagt aan het legitieme doel.

Respondenten geven aan effectiviteit van algoritmes vaak te meten met klassieke metrieken voor voorspelkracht, zoals accuraatheid, *true positive rate*, en *true negative rate.* Ook de *hit rate*[49] wordt regelmatig genoemd, met name in de context van controles.

[48] Hekkelman et al. (2025). *Selectiviteitsscan: Veilige Toetsing van Algoritmes*. Centraal Planbureau.

[49] In de computerwetenschappelijke literatuur refereert *hit rate* normaal gesproken aan *recall* of de *true positive rate:* het aantal daadwerkelijke positieven dat als zodanig is geclassificeerd door het algoritme, i.e. het aantal echtpositieven gedeeld door het totaal aantal daadwerkelijke positieven. Sommige geïnterviewden gebruiken *hit rate* om te refereren naar de precisie van het algoritme: het aantal *daadwerkelijke* positieven van alle voorbeelden die als positief zijn geclassificeerd, i.e. het aantal echtpositieven gedeeld door het totaal aantal *voorspelde* positieven.

Soms is het lastig om metrieken van voorspelkracht te verbinden aan effectiviteit ten op zichtte van het beoogde doel, geven enkele geïnterviewden aan. Hoe link je bijvoorbeeld een bepaalde *false positive rate* aan bespaarde tijd? Hoe bepaal je hoeveel geld is bespaard door de inzet van een algoritme? Neem bijvoorbeeld *predictive policing* algoritmes, welke als doel hebben om politieagenten effectiever te verspreiden over verschillende buurten en zo criminaliteit beter op te sporen of zelfs te voorkomen. Er bestaan twijfels over het meten van de effectiviteit van dergelijke algoritmes.[50] Een stijging in arrestatiecijfers betekent bijvoorbeeld niet per definitie dat het algoritme effectiever was dan bestaand beleid, het kan ook betekenen dat er simpelweg meer criminaliteit plaatsvond. Aan de andere kant kan een daling in arrestatiecijfers betekenen dat de inzet van agenten een preventieve werking had, maar deze trend had misschien ook wel plaatsgevonden zonder extra blauw op straat. Dergelijke interacties tussen een algoritme en de *status quo* maken het ingewikkeld om de effectiviteit van bestaand beleid te vergelijken met algoritmisch beleid.

De doelvariabele speelt een belangrijke rol in het beoordelen van effectiviteit van een algoritme. Een doelvariabele met gebrekkige validiteit kan er immers voor zorgen dat het algoritme niet of minder effectief is om het beoogde doel te bereiken. Een algoritme dat “geschiktheid” van sollicitanten voorspelt op basis van vooringenomen wervingsbeslissingen in het verleden zal bijvoorbeeld niet leiden tot de selectie van de beste kandidaten. Een algoritme dat “fraude in toeslagenaanvragen” beoogt te detecteren maar feitelijk “onbedoelde fouten in toeslagenaanvragen” voorspelt, leidt niet tot het detecteren van meer fraude – ook niet bij een hoge voorspelkracht.

Geschiktheid moet in principe worden getoetst voor *alle* selectiecriteria die worden toegepast in het algoritme. In de praktijk wordt vaak (in eerste instantie) het algoritme als geheel getoetst. Wanneer een organisatie als doel heeft gesteld om de efficiëntie van controles te verbeteren, kan bijvoorbeeld worden getoetst of een bepaalde selectieregel een hogere precisie heeft dan een aselecte steekproef, wat zou betekenen dat de selectieregel met een gelijk aantal controles meer normovertredingen detecteert.

Indien er sprake is van onderscheid in uitworp van het algoritme, moet in ieder geval worden achterhaald waar dit onderscheid vandaan komt. Sommige respondenten toetsen bijvoorbeeld of er een specifieke inputvariabele is die grote verschillen veroorzaakt in risicoscores. Ook geven sommige geïnterviewden aan te toetsen in hoeverre individuele inputvariabelen bijdragen aan de voorspelkracht van het algoritme. Dit kan worden getoetst door de voorspelkracht van twee varianten van het algoritme te vergelijken: een versie met en een versie zonder de specifieke inputvariabele.

Met betrekking tot een mogelijke rechtvaardiging, moet ook worden bepaald of de input variabelen op een inhoudelijk relevante manier worden ingezet als selectiecriterium. Complexe, zelflerende algoritmes kunnen selectiecriteria toepassen die niet inhoudelijk zijn geëvalueerd. Dit risico is met name evident bij algoritmes die worden getraind op ongestructureerde data, zoals afbeeldingen of tekst. Het is bijvoorbeeld niet triviaal om onomstotelijk aan te tonen op basis van welk kenmerk op een foto een gezichtsherkenningsalgoritme tot een classificatie komt. Maar ook in het geval van tabulaire data kunnen variabelen op onverwachte manieren worden gecombineerd of gebruikt die inhoudelijk niet relevant zijn. Denk bijvoorbeeld aan een cv-selectie algoritme dat, als gevolg van leeftijdsdiscriminatie in het verleden, minder hoge scores toekent aan sollicitanten omdat zij méér jaren werkervaring hebben. Een objectieve rechtvaardiging van een complex algoritme dat onderscheid maakt is hierdoor moeilijk (zo niet onmogelijk) te geven: het is niet uit te leggen waarom het onderscheid-veroorzakende deel onmisbaar en inhoudelijk relevant is voor het beoogde doel.

Een selectiecriterium of algoritme dat leidt tot indirect onderscheid en niet effectief is, levert discriminatie op en is verboden.

Noodzakelijkheidstoets

Een derde criterium voor een objectieve rechtvaardiging is dat de inzet van het algoritme noodzakelijk is om het legitieme doel te bereiken. Eén van de vereisten voor noodzakelijkheid is subsidiariteit: als er een minder ingrijpend alternatief bestaat, moet dit worden gebruikt. Een alternatief kan zowel algoritmisch als niet-algoritmisch zijn. Om te spreken van een redelijk alternatief, moet worden beoordeeld of het effectief, betaalbaar, en uitvoerbaar is. Een alternatief geldt ook

[50]Zie bijvoorbeeld C. Lum et al. (2017). *Understanding the limits of technology’s impact on police effectiveness*. Police quarterly, 20(2), 135-163.; L. Waardenburg, A.V. Sergeeva & M. Huysman (2020). *Predictive policing ontcijferd: Een etnografie van het 'Criminaliteits Anticipatie Systeem 'in de praktijk*. Cahiers Politiestudies, 1(54), 69-88.

als redelijk alternatief wanneer het minder onderscheid maakt maar wel marginaal meer kost of (iets) minder effectief is dan het aanvankelijke algoritme. In het kader van de subsidiariteitsvraag, zou een rechter de beklaagde kunnen vragen om aan te tonen dat zij redelijke stappen hebben ondernomen om een minder discriminerend alternatief te vinden.

In de noodzakelijkheidstoets spelen verklarende variabelen, zoals toegepast in *conditionele demografische pariteit* (Sectie 3.1), een mogelijke rol. Neem een cv-selectie algoritme dat leidt tot onderscheid op basis van gender, doordat mannen in een bepaalde sector gemiddeld gezien meer werkervaring hebben dan vrouwen. Deze uitspraak kan worden ondersteund door de selectieratio's geconditioneerd op werkervaring: als na conditionering geen sprake meer is van onderscheid, kan het onderscheid dus worden verklaard door verschillen in werkervaring. Wanneer een werkgever kan beargumenteren dat werkervaring essentieel is om de functie te vervullen, volgt dat alternatieve selectiealgoritmes die niet leiden tot onderscheid op basis van gender minder effectief zijn.

In het geval van een simpel regelgebaseerd selectiealgoritme kan, als minder discriminerend alternatief bijvoorbeeld een specifiek selectiecriterium dat onderscheid veroorzaakt, worden weggelaten uit het algoritme. Bij complexe, zelflerende algoritmes, kan het lastig zijn een algoritmisch alternatief te vinden dat geen onderscheid maakt. Immers, wanneer de doelvariabele gecorreleerd is aan een beschermde grond, zal een zelflerend algoritme dit reproduceren - ook wanneer specifieke inputvariabelen worden verwijderd en het model opnieuw wordt getraind. Indien er geen objectieve rechtvaardiging is voor het onderscheid, is in dergelijke gevallen het enige algoritmisch alternatief een volledig nieuw algoritme met een nieuwe doelvariabele die niet (of minder) is gecorreleerd aan een beschermde grond. Een allocatiealgoritme dat onderscheid veroorzaakt doordat "zorgkosten" wordt gebruikt als proxy voor zorgbehoeftes, kan bijvoorbeeld worden vervangen door een algoritme dat gerichtere gezondheidsindicatoren voorspelt.

Proportionaliteitstoets

Tot slot vindt dan de daadwerkelijke proportionaliteitstoets plaats (in juridisch jargon, de proportionalitetistoets *stricto sensu*). Hier wordt een afweging gemaakt tussen de behaalde doelen van het algoritme en negatieve effecten. Het algoritme mag nooit meer nadeel opleveren dan nodig om het legitieme doel te bereiken. Ook moet er een redelijk evenwicht zijn tussen alle belangen. Hoe groot is de effectiviteitswinst door inzet van het algoritme? Hoe onevenredig is het effect van indirect onderscheid? Hoe indringend of belastend is de nadelige behandeling? Hoe zwaarwegend is het legitieme doel? Wat zijn de gevolgen als het algoritme niet wordt ingezet? Gezien de contextuele aard van de discriminatietoets in Nederlandse en Europese wetgeving, is hier vooraf geen eenduidig antwoord op te geven. Er zal altijd per geval moeten worden afgewogen of sprake is van discriminatie of niet.

De vier-vijfde (80%) regel is niet van toepassing in Nederland

In de computerwetenschappelijke literatuur wordt regelmatig de zogenoemde vier-vijfde-regel (*four-fifths rule*) aangehaald als grenswaarde voor "eerlijkheid". Deze vuistregel, vastgelegd in specifieke richtlijnen voor de interpretatie van arbeidswetgeving in de Verenigde Staten, stelt dat de ratio tussen het selectiepercentage van de beschermde groep ten minste vier-vijfde (80%) is van het selectiepercentage van de referentiegroep. Gezien Nederland niet tot de Amerikaanse jurisdictie behoort, is deze grenswaarde echter niet relevant in de Nederlandse context.[51]

De proportionaliteitsafweging wordt door de geïnterviewde dataprofessionals ervaren als één van de grootste uitdagingen in het toetsen van discriminatie van algoritmes. Hoewel meerdere geïnterviewden expliciet aangeven dat zij vinden dat deze keuzes niet aan hen moet worden overgelaten, hebben zij de indruk dat de verantwoordelijkheid soms te veel bij ontwikkelaars wordt gelaten. In meerdere geïnterviewde organisaties wordt de normatieve discussie aangegaan in een multidisciplinaire setting. Sommige data professionals ervaren moeite met het presenteren van technische details en statistische metrieken. Hoe maak je bijvoorbeeld een misclassificatieratio concreet en invoelbaar? Zelfs met dezelfde

[51] Ook in de Amerikaanse context is de vier-vijfde-regel op zichzelf niet voldoende om discriminatie aan te tonen (of weerleggen). Zie ook E.A. Watkins & J. Chen (2024). *The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness*. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 764-775).

onderliggende data kan de keuze voor een specifieke vorm mogelijk effect hebben op hoe de data wordt geïnterpreteerd, stelt een geïnterviewde. Twee selectieratio's, 0.25 en 0.30, kunnen bijvoorbeeld worden gepresenteerd als een verschil van 5 procentpunt of een verschil van 20%.

Geïnterviewden geven aan behoefte te hebben aan handvatten bij het voeren van de proportionaliteitsafweging. Welke informatie is nodig om tot een besluit te komen? Hoe ziet een goede onderbouwing eruit? Hoe zorg je ervoor dat alle relevante partijen deel (kunnen) nemen aan het gesprek? Casussen vinden veelal plaats achter gesloten deuren, waardoor ook niet geleerd kan worden van andere organisaties, terwijl hier wel behoefte aan is.

4. Het voorkómen van discriminatie

Discriminatie is verboden. Voor handhavende overheidsinstanties volgt uit het recht daarnaast een voorzorgplicht om discriminatie zoveel mogelijk te voorkomen.[52] Voldoen instanties niet aan hun voorzorgsverplichting, dan kan dit een schending van het discriminatieverbod opleveren. Maar hoe voorkom je discriminatie door de inzet van algoritmes?

In dit hoofdstuk worden een aantal methodes voor het ondervangen voor risico's op onbedoeld indirect onderscheid door het gebruik van algoritmes verder toegelicht. De nadruk ligt op "technische" interventies en aandachtspunten bij de ontwikkeling van een algoritme. Gezien de complexiteit en contextafhankelijkheid van discriminatie in de context van algoritmes, is dit geen volledig overzicht van alle mogelijke voorzorgsmaatregelen.

4.1. De klassieke benadering: eerlijkheid als optimaliseringsprobleem

In de computerwetenschappelijke literatuur zijn het afgelopen decennium honderden methodes voorgesteld om "*bias*" te mitigeren en "eerlijkheid" te verbeteren.[53] De klassieke benadering in de technische literatuur is om (indirect) onderscheid te behandelen als een optimaliseringsprobleem: een zo hoog mogelijke voorspelkracht behalen onder een randvoorwaarde *(constraint*) die (mogelijk) onderscheid kwantificeert.

Voorbeelden van "fairness" methodes

- De dataset wordt gemanipuleerd tot er geen meetbare (of weinig) correlatie is tussen "gender" en "selectie". Een model getraind op deze dataset zal het gebrek aan correlatie reproduceren, met als gevolg dat de voorspellingen niet of weinig correleren met "gender".
- Het zelflerende algoritme wordt "bestraft" tijdens de *training* wanneer voorspellingen niet accuraat zijn voor datapunten gerelateerd aan mensen met een migratieachtergrond. Op deze manier wordt gepoogd een model te trainen dat even goede voorspellingen maakt, ongeacht migratieachtergrond.
- Een selectiealgoritme gebruikt een lagere beslisdrempel voor vijftigplussers, zodat zij vaker worden geselecteerd.

Kwantitatieve randvoorwaardes worden vaak gebaseerd op *group fairness* metrieken. Denk bijvoorbeeld aan een maximaal verschil in *true positive rates* tussen groepen of een minimum ratio in selectieratio. De meeste methodes leggen verder geen (of weinig) restricties op de manier waarop aan de randvoorwaarde wordt voldaan.

[52] Zie ook Hoofdstuk 4 van College voor de Rechten van de Mens (2025). *Toetsingskader risicoprofilering*.
[53] Zie voor een recent overzicht M. Hort et al. (2024) *Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey*. ACM J. Responsib. Comput. 1, 2, Article 11 (June 2024), 52 pages.

Voorbeeld: groep-specifieke beslisdrempels

Eén van de simpelste technieken in de computerwetenschappelijke literatuur is het gebruik van groep-specifieke beslisdrempels (*decision threshold)*. In binaire classificatieproblemen geeft de beslisdrempel aan vanaf welke risicoscore een datapunt als "positief" wordt bestempeld. De beslisdrempel speelt een belangrijke rol in de afweging tussen valspositieven en valsnegatieven. Een hogere drempel leidt over het algemeen tot minder (vals)positieven, terwijl een lagere drempel minder (vals)negatieven geeft.

Door iedere groep een unieke beslisdrempel toe te kennen, kunnen statistieken zoals selectieratio's en misclasssificatieratio's worden gemanipuleerd. Er kan bijvoorbeeld gekozen worden voor een lagere beslisdrempel voor vrouwen in een cv-selectiemodel, zodat het selectiepercentage stijgt (alsook het percentage valspositieven).

In software-implementaties van deze techniek is de selectie van beslisdrempels vaak geautomatiseerd: er wordt gekozen voor de drempels welke voldoen aan een vooraf gedefinieerde randvoorwaarde (zoals een maximaal verschil in selectiepercentages) en zo min mogelijk wordt ingeboet op voorspelkracht (e.g. *accuracy*).

Een groeiend aantal wetenschappelijke publicaties waarschuwt dat simplistische optimalisatie kan leiden tot ineffectieve en zelfs schadelijke interventies. Een belangrijk argument hiervoor is dat in de klassieke benadering veelal geen rekening wordt gehouden met de onderliggende oorzaak van ongewenst onderscheid door de inzet van het algoritme.

De klassieke benadering houdt veelal geen rekening met de onderliggende oorzaak van onderscheid

Neem bijvoorbeeld een verschil in misclassificatiepercentages (i.e. *false positive rates* en *false negative rates*) tussen twee beschermde groepen, groep *A* en groep *B*. Hiervoor kunnen verschillende oorzaken bestaan.

Ten eerste kan het verschil veroorzaakt worden door een verschil in voorspelkracht: het model is simpelweg minder goed in staat om "positieven" en "negatieven" van elkaar te onderscheiden voor de ene groep ten opzichte van de andere groep. Een tweede mogelijke oorzaak is een verschil in de verdeling van echte "positieven" en "negatieven" tussen groepen. Wanneer de kans op een "echte positief" in de beschermde groep *A* hoger is dan in de groep *B* (i.e. er bestaat een associatie tussen de doelvariabele en het beschermde persoonskenmerk), is bij een gelijke beslisdrempel de kans op een valspositief in groep *A* hoger (en de kans op een vals negatief kleiner) ten op zichtte van groep *B*. De associatie kan vervolgens weer te wijten zijn aan problemen met datavaliditeit, bestaande ongelijkheid in de samenleving, of een combinatie hiervan (zie Hoofdstuk 2).

Een technische interventie gericht op het minimaliseren van het verschil in misclassificatiepercentages, maakt echter geen onderscheid tussen deze situaties. Dit heeft directe gevolgen voor de effectiviteit en wenselijkheid van een interventie. Als verschillen tussen groepen te wijten zijn aan te weinig (voorspellende) data, zijn interventies tijdens het *trainen* van het model bijvoorbeeld slechts beperkt effectief. Wanneer verschillen veroorzaakt worden door een (ongewenste) associatie tussen de doelvariabele en de beschermde grond, hangt de effectiviteit en wenselijkheid af van de specifieke oorzaak van deze associatie. Als er twijfels bestaan over de validiteit van *labels* in de *training* dataset, is het bijvoorbeeld maar de vraag of het gelijktrekken van misclassificatiepercentages de juiste mensen bereikt. Aan de data zelf kan dan immers niet worden afgelezen of een datapunt een "echte" positief of negatief betreft. Denk bijvoorbeeld aan een cv-selectiealgoritme: zelfs als we vermoeden dat in het verleden wervingsbeslissingen zijn beïnvloed door vooroordelen, weten we niet om welke specifieke cv's het gaat. Het gelijktrekken van misclassificatiepercentages kan, per toeval, leiden tot selectie van de "juiste" cv's, maar dat is geen garantie.

Bij het mitigeren van de gevolgen van bestaande ongelijkheid in de samenleving, komen vaak ingewikkelde normatieve afwegingen kijken, die niet zomaar kunnen worden gevat in een kwantitatieve randvoorwaarde. Denk bijvoorbeeld aan de beoordeling van de inhoudelijke relevantie van de gebruikte inputvariabelen. Waar sommigen het gebruik van contant geld als risicofactor in een risicoprofileringsalgoritme proportioneel achten, vinden anderen het verband tussen contant geld en fraude te zwak in relatie tot de mogelijke discriminatoire gevolgen. Er zijn immers veel legitieme redenen om contant geld te gebruiken. Zulke redeneringen zijn niet kwantificeerbaar. Ook kan de inzet van een technische interventie ongewenste bijwerkingen hebben. Zo is het bijvoorbeeld mogelijk om een verschil in misclassificatiepercentages te minimaliseren door het misclassificatiepercentage van de "bevoordeelde" groep te verhogen, zonder de situatie van de benadeelde groep in absolute zin te verbeteren.[54]

4.2. Discriminatie voorkomen

Klassieke "*debiasing*" methodes uit de wetenschappelijke literatuur zullen veelal niet effectief zijn in het voorkomen van ongewenst onderscheid. In de Nederlandse praktijk lijken deze interventies weinig tot niet te worden gebruikt: niet één van de geïnterviewden geeft aan dat gebruikt wordt gemaakt van deze methodes. Meerdere respondenten ervaren terughoudendheid of geven aan geen voorstander te zijn in het toepassen van deze technieken. Een alternatieve benadering, waarbij de oorzaken van onderscheid worden geïdentificeerd en gericht worden aangepakt, heeft de voorkeur. In de rest van dit hoofdstuk zal dieper worden ingegaan op alternatieve methodes en hoe deze in verhouding staan tot de discriminatietoets.

Voorspelkracht

Om voorspelkracht gelijk te trekken tussen groepen, kan een data scientist gebruik maken van alle standaardmethodes gericht op het verbeteren van voorspelkracht. Het belangrijkste verschil is de focus op voorspelkracht voor verschillende (sub)groepen, in plaats van gemiddelde voorspelkracht over de gehele populatie.

In veel gevallen zal de beschikbare data een belangrijk knelpunt zijn voor voorspelkracht. Mogelijke gerichte interventies zijn het verzamelen van extra datapunten of het identificeren van extra variabelen met hoge(re) voorspelkracht voor specifieke subgroepen. Uit de interviews blijkt dat het verzamelen van extra datapunten soms voorkomt. Het identificeren van extra variabelen lijkt weinig voor te komen. Het verzamelen van nieuwe variabelen is vaak zeer kostbaar. Vaak is daarom in eerdere stadia van het ontwikkelproces al uitgebreid nagegaan welke data beschikbaar is, aldus sommige geïnterviewden. Sommige respondenten ervaren met name datakwaliteit als een grote uitdaging en geven aan dat zij graag zien dat er binnen hun organisatie meer ruimte en aandacht is voor het verbeteren van datakwaliteit.

Gebrekkige voorspelkracht kan ook te wijten zijn aan specifieke ontwerpkeuzes, van datavoorbereiding tot modellering. Het is daarom raadzaam om bij het maken van ontwerpkeuzes niet enkel af te gaan op *metrieken* berekend over de gehele populatie, maar ook expliciet te evalueren wat het effect is voor (kwetsbare) subgroepen. Recent onderzoek heeft benadrukt dat er vaak meerdere modellen bestaan voor een specifieke voorspeltaak met vergelijkbare voorspelkracht, maar grote verschillen in individuele voorspellingen of eigenschappen zoals uitlegbaarheid en robuustheid.[55] Het meenemen van prestaties op (sub)groepsniveau in de evaluatie van een algoritme is daarom relevant met betrekking tot de subsidiariteitsvereiste besproken in Sectie 3.2: bestaat er een net zo goed model waarbij de voorspelkracht minder scheef verdeeld is? Net als bij het meten van mogelijke discriminatie, vormen normen van privacy en gegevensbescherming hier een aandachtspunt.

Datavaliditeit

Wanneer mogelijk indirect onderscheid wordt veroorzaakt door gebrekkige datavaliditeit, is de meest voor de hand liggende interventie het verzamelen van betere data. In het geval van zelflerende algoritmes is de validiteit van de doelvariabele cruciaal. Het doel van een zelflerend algoritme is immers om de doelvariabele accuraat te voorspellen.

---

[54] Dit word ook wel *levelling down* genoemd. Zie ook Weerts et al. (2025) *Are There Exceptions to Goodhart's Law? On the Moral Justification of Fairness-Aware Machine Learning*. ACM Journal on Responsible Computing.

[55] E. Black et al. (2022) *Model multiplicity: Opportunities, concerns, and solutions* (2022). In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency (pp. 850-863).

Wanneer de doelvariabele geen juiste representatie is van het concept dat men beoogt te voorspellen, kan dit zowel gebrekkige effectiviteit als onderscheid opleveren.

Constructvaliditeit is geen binair gegeven, maar een schaal. Of een variabele (of, meer generiek, "meetinstrument") als valide kan worden beschouwd, moet worden ondersteund door kritisch redeneren.[56] Is het, op basis van expertoordelen, aannemelijk dat de variabele meet wat het zou moeten meten? Worden alle relevante aspecten van het fenomeen meegenomen? Correleert de variabele met andere, gevestigde meetinstrumenten? Correleert de variabele *niet* met irrelevante variabelen (bijvoorbeeld een discriminatiegrond)? Wanneer het algemeen bekend of redelijkerwijs aannemelijk is dat een *doelvariabele* gecorreleerd is met een beschermde grond, is een inhoudelijke onderbouwing van de *doelvariabele* een minimale vereiste in de geschiktheidstoets.

Wanneer wordt vastgesteld dat een variabele niet voldoet, zullen interventies gericht moeten zijn op het identificeren van een (meer) valide meetinstrument. In sommige gevallen is het mogelijk een andere, geschiktere en inhoudelijk relevantere variabele, te identificeren. Zo zou in het geval van een allocatiealgoritme in de zorg de *doelvariabele* "zorgkosten" kunnen worden vervangen door een gevalideerde set biomarkers, welke gezondheidstoestand gerichter meten.

In de praktijk ervaren sommige geïnterviewde data professionals weinig speelruimte in hun keuze voor de doelvariabele. Vaak wordt de beslissing voor een specifieke doelvariabele gemaakt op basis van de organisatorische context (i.e. effectiviteit) of is er simpelweg geen betere variabele beschikbaar. Andere respondenten geven aan dat het doel van het algoritme vooraf niet altijd goed genoeg is gedefinieerd door de interne opdrachtgever, wat het bemoeilijkt om een geschikte doelvariabele en prestatie-indicator te selecteren.

Wanneer een dataset is verzameld op basis van bestaand beleid, in plaats van willekeurige selectie, is deze vaak niet representatief – een bedreiging voor constructvaliditeit. Denk bijvoorbeeld aan fraudedetectie, waar vaak enkel data beschikbaar is voor gevallen waarin eerder is besloten over te gaan tot inspectie. Wanneer de dataset niet willekeurig is geselecteerd, maar wel voldoende dekkend is, kan soms gebruik worden gemaakt van herweging van datapunten om de scheve verdeling te mitigeren. Meestal zal het echter nodig zijn om een nieuwe dataset te verzamelen op basis van willekeurige selectie. Dit gebeurt ook in de Nederlandse praktijk steeds vaker, al geven enkele respondenten aan dat zij dit graag structureler zouden willen inbedden in de selectieprocessen in hun organisatie.

Ook in het geval van mogelijk vooringenomen labels zal nieuwe data moeten worden verzameld, waarbij vooringenomenheid van beoordelaars zoveel mogelijk wordt uitgesloten, bijvoorbeeld door gerichte opleiding en verbeterde werkinstructies.[57] Het toetsen naar mogelijk vooringenomen labels is niet eenvoudig. Zo is het soms praktisch niet haalbaar om een controle meerdere malen uit te voeren door verschillende zaakmedewerkers. Hier speelt nog mee dat gevolgen van selecties vaak pas (veel) later zichtbaar worden. Denk bijvoorbeeld aan het aanbieden van extra hulp bij het zoeken naar een baan: pas na geruime tijd wordt duidelijk of een persoon hier daadwerkelijk baat bij heeft gehad. Dit maakt het ingewikkeld om de validiteit van menselijke beoordelingen te toetsen en te monitoren.

Tot slot geven geïnterviewden aan dat het belangrijk is om alert te zijn op verschuivingen in betekenis van de doelvariabele over tijd. Zo kan bijvoorbeeld door een wetswijziging de criteria voor het toewijzen van een middel, zoals een toeslag of uitkering, veranderen. In dat geval zijn historische labels niet langer valide. Enkele respondenten geven daarom aan dat het belangrijk is om, naast een (binair) label, ook rijkere metadata te verzamelen die laten zien waar een bepaald besluit in het verleden op is gebaseerd.

Ongelijkheid in de samenleving

In sommige gevallen weerspiegelt een associatie tussen een doelvariabele en een discriminatiegrond bestaande ongelijkheden in de maatschappij. Hoe hiermee moet worden omgegaan is bij uitstek een normatief vraagstuk.

[56] Zie voor een overzicht van vormen van constructvaliditeit in het kader van discriminatie en algoritmes: A.Z. Jacobs & H. Wallach (2021) *Measurement and fairness* (2021). In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 375-385).
[57] Zie bijvoorbeeld het *Toetsingskader Risocoprofilering* (2025).

Wanneer gebruik wordt gemaakt van algoritmes, is het belangrijk alert te zijn op inputvariabelen die mogelijk een proxy kunnen zijn voor een discriminatiegrond[58]. In het kader van een mogelijk objectieve rechtvaardiging van indirect onderscheid, moeten alle selectiecriteria immers daadwerkelijk inhoudelijk relevant zijn voor het nemen van de beslissing (zie Sectie 3.2). Hoewel enkele respondenten aangeven dat er in sommige organisaties nog weinig bewustzijn is omtrent proxyvariabelen en het belang van inhoudelijke relevantie, is het actief toetsen op mogelijke proxy's vaak een standaardonderdeel van een *bias* toets (als deze wordt uitgevoerd).

Meerdere respondenten noemen dat hun organisatie streeft naar het minimaliseren van het aantal inputvariabelen. Variabelen worden ten eerste soms juridisch getoetst, bijvoorbeeld door middel van een DPIA. Beschermde gronden worden in principe niet meegenomen en ook zeer duidelijke proxy's worden weggelaten uit de trainingsdataset. Daarnaast vindt ook steeds vaker een kwalitatieve discussie plaats in multidisciplinaire teams. Welke variabelen zijn er beschikbaar? Welke variabelen worden door domeinexperts relevant geacht? Kunnen dit mogelijk proxy's zijn voor een beschermde grond? Is de variabele nog steeds relevant (bijvoorbeeld na een wetswijziging)? Indien mogelijk, worden variabelen soms ook kwantitatief getoetst op een correlatie met een beschermde grond. Als er een correlatie is, wordt gekeken of hier een goede verklaring voor is. Ook tijdens het modelleren wordt aandacht besteed aan minimale complexiteit van het algoritme, aldus enkele geïnterviewden. Bij vergelijkbare voorspelkracht wordt dan de voorkeur gegeven aan een minder complex model, bijvoorbeeld met minder variabelen of minder complexe relaties.

In steeds meer organisaties wordt bij het uitvoeren van een *bias* toets dus actief gekeken naar mogelijke proxydiscriminatie. Hierbij moet wel worden benadrukt dat het uitsluiten van discriminatiegronden en mogelijke proxy-variabelen slechts beperkt effectief is om indirect onderscheid te voorkomen door zelflerende algoritmes. Zelflerende algoritmes zijn immers specifiek ontworpen om complexe relaties tussen variabelen te identificeren. Zo lang er een associatie bestaat en het model accuraat genoeg is, zal een bestaande associatie tussen een discriminatiegrond en de doelvariabele worden gereproduceerd en zal de inzet van het algoritme dus indirect onderscheid veroorzaken – óók wanneer de meest voor de hand liggende proxyvariabelen uit de dataset zijn verwijderd.

Het is mogelijk om het algoritme bij te sturen om indirect onderscheid te minimaliseren. In het geval van simpele selectiealgoritmes, kunnen selectiecriteria die onderscheid veroorzaken worden weggelaten. In het geval van complexere zelflerende algoritmes is dit moeilijker te verantwoorden. Zoals ook beschreven in Hoofdstuk 4.1, is dit een complex vraagstuk, onder andere met het oog op mogelijk voorkeursbeleid. Waar interventies gericht op het verbeteren van voorspelkracht en datavaliditeit over het algemeen duidelijk als doel hebben om discriminatie te voorkomen[59], is de grens tussen voorkeursbeleid en het voorkomen van discriminatie veroorzaakt door bestaande ongelijkheid in de samenleving een ingewikkeld normatief vraagstuk.[60]

In sommige gevallen biedt een *bias* analyse ook de mogelijkheid om ongelijkheden direct bij de bron aan te pakken, aldus een geïnterviewde. Het kan bijvoorbeeld voorkomen dat bepaalde groepen burgers vaker geselecteerd worden voor controle, omdat zij bijvoorbeeld vaker niet de juiste gegevens hebben aangeleverd of niet op de hoogte zijn van een betalingsverplichting. In dergelijke gevallen kan gerichte informatievoorziening in de toekomst onderscheid voorkomen.

Organisatorische voorzorgsmaatregelen

Hoewel dit rapport focust op statistische analyses en methodieken, stipten geïnterviewden ook organisatorische uitdagingen en voorzorgsmaatregelen aan ter voorkoming van discriminatie.

[58] Zie bijvoorbeeld College voor de Rechten van de Mens (2025). *Toetsingskader Risicoprofilering*.
[59] Zo vermoedde dating app Breeze in 2023 dat hun matching algoritme minder accurate matching kansen geeft voor gebruikers met een donkere huidskleur. Het College oordeelde dat dating app verplicht is om maatregelen te nemen om discriminatie te voorkomen. Zie College voor de Rechten van de Mens (2023), Oordeel 2023-82.
[60] Weerts et al. (2024). *The neutrality fallacy: when algorithmic fairness interventions are (Not) positive action (2024).* In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2060-2070).

Ten eerste geven meerdere respondenten aan dat zij graag zouden zien dat organisaties hun algoritmes actiever of beter monitoren, zodat signalen van (ongewenst) onderscheid tijdig kunnen worden opgepakt. Ook het verbeteren van klachtenprocedures wordt meerdere malen genoemd.[61]

Daarnaast stippen sommige geïnterviewde data professionals het belang aan van integratie van *bias* toetsing in het gehele ontwikkelproces. Zij zouden idealiter zien dat belangrijke normatieve keuzes, zoals de prestatiemetriek en fairness metriek en keuze voor bepaalde inputvariabelen en doelvariabele, goed worden gedocumenteerd – inclusief een duidelijke verklaring voor deze keuzes.

Tegelijkertijd ervaren de geïnterviewden soms terughoudendheid binnen hun organisatie om te toetsen op onderscheid. In sommige organisaties zien geïnterviewden een gebrek aan prioriteit. Veel organisaties ervaren grote druk om algoritmes en AI toe te passen, bijvoorbeeld om kosten te besparen. Het toetsen op onderscheid heeft dan soms een lagere prioriteit. Ook is het vaak onduidelijk in hoeverre er een wettelijke verplichting bestaat voor het toetsen op onderscheid. In andere gevallen is er geen gevoel van noodzaak door een gebrek aan bewustzijn over de risico's op discriminatie (bijvoorbeeld door het gebruik van proxyvariabelen). Ook zien respondenten dat discriminatie een gevoelig onderwerp kan zijn binnen een organisatie. Organisaties vrezen reputatieschade en besluiten soms om dan maar niet te toetsen. Ook wordt het toetsen op onderscheid vaak ervaren als tijdrovend en complex. De afgelopen jaren is er een grote hoeveelheid kaders en hulpmiddelen beschikbaar geworden. Zo noemen respondenten bijvoorbeeld het IAMA en DPIA, maar ook het Toetsingskader risicoprofilering, het Algoritmekader, en het fairness handboek van de gemeente Amsterdam. Sommige organisaties hebben ook een eigen algoritmebeleid ontwikkeld. Het doorlopen van al deze stappen kost vaak veel tijd. Zo kan het soms jaren duren voordat een DPIA is afgerond. Ook zien sommige geïnterviewden dat kaders niet altijd goed aansluiten op de praktijk: niet alle vragen lijken even relevant lijken voor iedere casus. Dit kan ervoor zorgen dat binnen een organisatie weinig motivatie is om met toetsing aan de slag te gaan. Het uitvoeren van het kader wordt dan een checklist, terwijl veel respondenten aangeven juist het voeren van de discussie over de rechtvaardiging van het algoritme essentieel te vinden.

Respondenten geven aan behoefte te hebben aan concretere handvatten in het voeren en documenteren van normatieve discussies. Hoewel instrumenten zoals het IAMA helpen om mogelijke risico's in kaart te brengen, missen respondenten vaak specifieke handvatten voor het toetsen en voorkomen van *bias*. Hoe voer je een discriminatietoets gedegen uit? Welke stappen moet een organisatie in ieder geval ondernemen? Niet iedere organisatie heeft een ethicus in dienst die een ethische sessie kan begeleiden. Meerdere respondenten geven aan graag meer ervaringen te willen delen met andere organisaties. Sommige respondenten noemen ook transparantie ten op zichtte van de maatschappij als belangrijk aandachtspunt: men moet achter het besluit kunnen staan om een algoritme te gebruiken, vertrouwen opbouwen, en de discussie aangaan met de maatschappij.

[61] Voor het organiseren van een goede klachtprocedure bij discriminatie, zie ook de brochure College voor de Rechten van de Mens (2013). *Zorgvuldig omgaan met discriminatieklachten*.

5. Conclusies

Algoritmes worden steeds vaker ingezet om producten, processen, en beleidsmaatregelen doeltreffender en efficiënter te maken. Het gebruik van algoritmes brengt echter ook significante risico's en heeft in de afgelopen jaren herhaaldelijk tot discriminatie geleid. In de computerwetenschappen is het afgelopen decennium een veelvoud aan kwantitatieve methodes ontwikkeld voor het meten en mitigeren van onderscheid en (mogelijke) discriminatie. Voor zowel juristen als data professionals is het vaak nog onduidelijk in hoeverre deze kwantitatieve methodes overlappen met de juridische discriminatietoets. In dit rapport is daarom ingegaan op de volgende vragen:

1. Wat zijn gangbare kwantitatieve methodes voor het meten en voorkomen van discriminatie in algoritmes?
2. In hoeverre overlappen deze methodes met de discriminatietoets?
3. Hoe gebruiken ontwikkelaars verschillende methodes en welke uitdagingen ondervinden zij hierbij?

In dit hoofdstuk worden de hoofdpunten van dit rapport samengevat.

Discriminatie is geen technisch probleem. Wel zijn er bepaalde technische aspecten die het risico op discriminatie door algoritmes vergroten, zoals verschillen in voorspelkracht en een associatie met discriminatiegronden. De wet verbiedt zowel *direct onderscheid*, waarbij ongelijke behandeling plaatsvindt op basis van (beschermde) persoonskenmerken, als *indirect onderscheid*, waarbij een op het oog neutrale maatregel een groep mensen met een bepaald beschermd kenmerk ongelijk benadeelt. In de context van algoritmes is zowel het meten als voorkomen van indirect onderscheid een uitdaging. Indirect onderscheid als gevolg van de inzet van een algoritme kan onder andere ontstaan door verschillen in voorspelkracht tussen beschermde groepen als een associatie tussen een beschermd persoonskenmerk en de doelvariabele. Een dergelijke associatie kan enerzijds het gevolg zijn van bestaande ongelijkheden in de samenleving en anderzijds worden veroorzaakt door gebrekkige datavaliditeit, bijvoorbeeld door een slecht gekozen doelvariabele, bevooroordeelde historische labels, of over- of ondervertegenwoordiging van beschermde groepen in selecties in het verleden.

Een *group fairness* analyse komt tot op zekere hoogte overeen met de discriminatietoets, maar kwantitatieve gegevens alleen zijn niet voldoende om onderscheid of discriminatie aan te tonen. In het afgelopen decennium hebben onderzoekers veel verschillende *fairness metrics* (eerlijkheidsmetrieken) voorgesteld. *Group fairness* analyses, waarin groepsstatistieken zoals selectieratio's of misclassificatieratio's worden vergeleken tussen groepen, zijn het meest gangbaar in de wetenschappelijke literatuur. Ook de juridische indirecte discriminatietoets vergelijkt, in de kern, uitkomsten tussen (groepen) personen: worden personen die behoren tot een beschermde groep in het bijzonder benadeeld ten op zichtte van anderen in een vergelijkbare situatie? In hoeverre een *group fairness* analyse overeenkomt met de indirecte discriminatietoets, hangt echter af van de gekozen groepsstatistiek, het persoonskenmerk dat groepen definieert, de referentiepopulatie waarover de statistieken worden berekend, en de gestelde grenswaarde voor ongewenst onderscheid.

Een kernvraag in het beoordelen van onderscheid en discriminatie is of het algoritme doet waar het voor is ontworpen. Voorspelkracht alleen is hiervoor niet voldoende om de effectiviteit van het algoritme aan te tonen. Aandachtspunten zijn, onder andere:

- Is de variabele die wordt voorspeld door het algoritme (de *doelvariabele*) een valide instrument om het fenomeen dat wordt voorspeld te meten? Is de verzamelde data representatief, of zijn bepaalde groepen onder- of overgerepresenteerd? Wanneer data is gelabeld op basis van menselijke beoordelingen, is er voldoende rekening gehouden met mogelijke vooringenomenheid? Is de beschikbare data nog actueel?
- Leidt de inzet van het algoritme ook daadwerkelijk tot het behalen van het nagestreefde doel, i.e. is het effectief? Is de *doelvariabele* daadwerkelijk relevant voor het doel dat wordt nagestreefd door de inzet van het algoritme? Zijn de gebruikte criteria effectief?
- Zijn de gebruikte criteria (*inputvariabele*n) inhoudelijk relevant voor de beslissing? Is de data van voldoende kwaliteit?
- Is de voorspelkracht van het model voldoende voor alle groepen? Generaliseren de prestaties naar de praktijk?

Technische methodes die optimaliseren voor kwantitatieve randvoorwaarden, zonder een diepere grondoorzaakanalyse van geobserveerd onderscheid, zijn niet effectief om discriminatie te voorkomen. In de computerwetenschappelijke literatuur is een veelvoud aan methodes voorgesteld om (zelflerende) algoritmes te "*de-biasen*". Deze methodes zijn vaak ingericht om voorspelkracht te maximaliseren onder een kwantitatieve randvoorwaarde, zoals gelijke groepsselectieratio's. Simplistische optimalisaties kunnen leiden tot ineffectieve en zelfs schadelijke interventies, omdat veelal geen rekening wordt gehouden met de onderliggende oorzaak van onderscheid.

Effectieve toetsing en preventie van discriminatie vereist multidisciplinaire samenwerking, documentatie van normatieve discussies, actieve monitoring, en transparantie. Data professionals ervaren hierin obstakels zoals een gebrek aan geschikte data, geen toegang tot beschermde persoonskenmerken, juridische onzekerheid, en soms lage prioriteit voor het toetsen van onderscheid binnen een organisatie. Daarnaast hebben zij behoefte aan richtlijnen voor het uitvoeren van gedegen en statistisch betrouwbare toetsen en het meten van effectiviteit onder verschillende omstandigheden, zoals onvolledig gelabelde data en meerstapse besluitvormingsprocessen. Tot slot geven geïnterviewden aan grote behoefte te hebben aan hulpmiddelen voor het voeren van normatieve discussies en het uitwisselen van ervaringen.

Het meten en voorkómen van discriminatie door de inzet van algoritmes is bovenal een normatieve uitdaging. Hoewel kwantitatieve metrieken mogelijk onderscheid inzichtelijk kunnen maken, valt of staat de analyse met de onderbouwing. De uitdaging ligt hierbij om de verbinding te maken tussen technische ontwerpkeuzes, (juridische) normen, en bovenal de impact die de inzet van het algoritme heeft op burgers, klanten, of andere belanghebbenden. Deze bevinding onderstreept de noodzaak van interdisciplinaire samenwerking, waarbij de kloof tussen data professionals, juristen, en de maatschappij wordt gedicht.

Bijlage A. Interviewstudie

Onderzoeksvragen

Het doel van de interviewstudie is om meer te weten te komen over de huidige *best practices*, uitdagingen, en behoeftes van ontwikkelaars en gebruikers van algoritmes en AI-systemen met betrekking tot het meten van mogelijk onderscheid en voorkomen van discriminatie. Hierbij stonden de volgende onderzoeksvragen centraal:

- In hoeverre gebruiken ontwikkelaars/gebruikers van data/algoritmes (technische) bias/fairness methodes om onderscheid te meten en toetsen of discriminatie te voorkomen?
- Waar liggen op dit moment de grootste uitdagingen in de praktijk om discriminatie of ongewenst onderscheid te detecteren en voorkomen in data en algoritmes?

Deelnemers

De beoogde deelnemers voor deze studie zijn professionals die betrokken zijn bij de ontwikkeling en/of het gebruik van data en algoritmes binnen (semi)overheidsinstellingen of andere organisaties met een belangrijke maatschappelijke functie (banken, verzekeraars, zorginstellingen), waarbij rekening is gehouden met mogelijk ongewenst onderscheid of discriminatie. Deelnemers zijn doelgericht benaderd omdat zij zich in hun werk bezighouden met de verantwoorde inzet van algoritmes, bijvoorbeeld als (interne) consultant of data scientist. De resultaten in dit rapport zijn gebaseerd op interviews met 14 respondenten. Deelnemers hebben tussen de 5 en 25 jaren werkervaring in een data-gerelateerde functie. Van alle deelnemers identificeerden 8 zich als man en 6 als vrouw.

Tabel 8 Functieprofielen respondenten

Functieprofiel	Aantal
Consultant (intern)	4
Consultant (extern)	3
Data-analist of data scientist	5
Programmamanager	2

Interviewopzet

Er is gebruik gemaakt van een semigestructureerd interviewprotocol. Ieder interview duurde ongeveer een uur (met uitschieters naar 45 en 75 minuten) en vond plaats via Microsoft Teams in de periode van Juli tot en met September 2025.

Tijdens het interview is ingegaan op de volgende onderwerpen:

- **Algemeen**: doel, motivatie, prestatie-indicatoren en andere details van het algoritme/AI-systeem.
- **Detecteren en meten**: processen en methodes voor het in kaart brengen en kwantitatief meten van discriminatoire effecten.
- **Wegen**: processen om te besluiten of actie moet worden ondernomen naar aan leiding van gemeten onderscheid.
- **Mitigeren en voorkomen**: processen en methodes om discriminatoire effecten te mitigeren of voorkomen.

Ethiek en privacy

Gezien de gevoeligheid van het onderwerp is uiterst zorgvuldig omgegaan met de vergaarde gegevens. De onderzoeksopzet is goedgekeurd door de ethische toetsingscommissie van Technische Universiteit Eindhoven (TU/e). Voor het delen van gepseudoanonimiseerde interviewtranscripties is een overeenkomst opgesteld tussen het College en de TU/e.