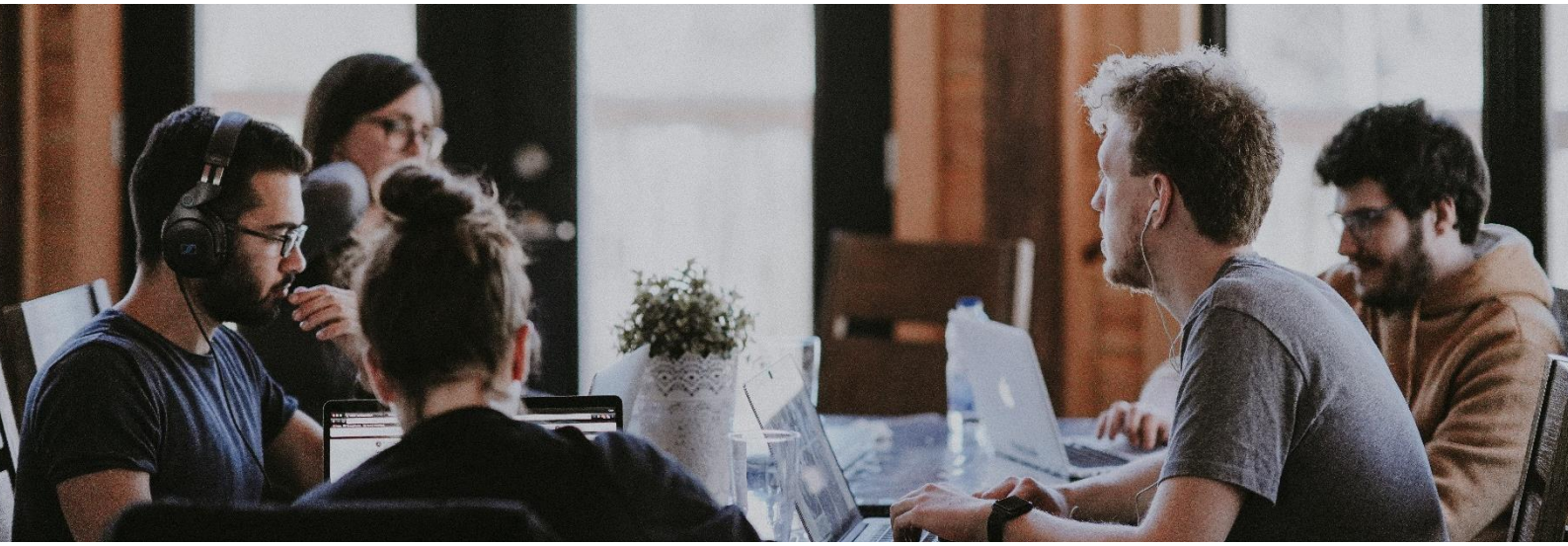
Non-discriminatie bij algoritmes



Bericht aan het parlement

Op 14 november vindt het wetgevingsoverleg over begrotingsonderdelen rondom digitalisering plaats. De kabinetsreactie op het rapport Algoritmes afwegen staat hiervoor op de agenda. Dit rapport gaat in op het gebruik van profilerende algoritmes bij uitvoeringsorganisaties. De politiek verwacht van hen nul procent discriminatie. Maar uitvoeringsorganisaties kunnen deze garantie niet geven. Wat kan wel? Het Rathenau Instituut geeft enkele opties.

De wens: geen discriminerende systemen.

Non-discriminatie staat hoog op de politieke agenda. Premier Rutte heeft erkend dat er bij de overheid, en de Belastingdienst in het bijzonder, sprake is geweest van institutioneel racisme, en gaf aan 'praktische stappen te willen zetten naar nul racisme'[1]. Ook in de Tweede Kamer klinkt de roep om discriminatie tegen te gaan steeds luider. Dit speelt in het bijzonder in de context van profilerende algoritmische systemen, die bijvoorbeeld fraudegevallen kunnen detecteren. Zo riep Kamerlid Koekkoek in een motie op om systemen stop te zetten als niet gegarandeerd kan worden dat discriminatie niet voorkomt. Met andere woorden: de overheid zou alleen systemen mogen gebruiken die niet discrimineren.

Waarom is dat lastig?

Het is belangrijk dat de politiek het debat over discriminatie voert, en de verbinding legt met digitalisering. Maar de eis dat algoritmes simpelweg niet discrimineren, is om twee redenen lastig waar te maken:

1. De wet kan op verschillende manieren geïnterpreteerd worden: er kan verschil van opvatting bestaan over wanneer er eigenlijk sprake is van een ongerechtvaardigd onderscheid. Hoewel onderscheid maken op basis van beschermde gronden als etniciteit of geslacht in principe niet mag, kan het volgens de wet in sommige situaties gerechtvaardigd zijn. Dit probleem zie je terug bij de vertaling van discriminatiewetgeving naar algoritmische systemen.
2. Discriminatie kan worden opgespoord, maar dit biedt geen 100% zekerheid: het is niet mogelijk om discriminatie bij lerende algoritmische systemen volledig uit te sluiten. Dit heeft te maken met indirecte discriminatie, waarbij een algoritme niet direct onderscheid maakt op basis van een beschermde grond zoals etniciteit of geslacht, maar met verloop van tijd kenmerken gaat gebruiken die met deze gronden correleren: de zogeheten proxy's.

Het is zaak dat deze nuances in het politieke debat erkend en besproken worden, zodat kabinet, parlement en uitvoering gezamenlijk de verantwoordelijkheid voor algoritmische systemen kunnen nemen. Dit bericht licht de twee nuances toe en schetst voorliggende handelingsopties.

Reden 1: het non-discriminatierecht is niet absoluut. Discriminatie verwijst naar het maken van een ongerechtvaardigd onderscheid, maar wanneer is daar precies sprake van? Het non-discriminatierecht is niet altijd absoluut – het draait regelmatig om het maken van de juiste afweging. Zo bestaan er bij directe discriminatie wettelijke uitzonderingsgronden. Het is bijvoorbeeld bij sollicitaties toegestaan om bij gelijke geschiktheid onderscheid te maken op grond van geslacht, als in een organisatie vrouwen of mannen zijn ondervertegenwoordigd. Ook bij indirecte discriminatie is er een 'objectieve rechtvaardiging' mogelijk, bijvoorbeeld als voor een functie een taaleis wordt gesteld. Hier zullen mensen met een andere nationaliteit dan de Nederlandse minder aan kunnen voldoen, maar het kan voor de functie gerechtvaardigd worden geacht. Met andere woorden: discriminatievraagstukken vereisen altijd een discussie over de precieze context waarbinnen een onderscheid wordt gemaakt, en het doel van het onderscheid.

---

[1] https://www.trouw.nl/binnenland/rutte-wil-naar-een-land-met-nul-racisme-na-gesprek-met-black-lives-matter-en-kick-out-zwarte-piet~bca46e65/

Uiteindelijk moet er volgens de wet sprake zijn van proportionaliteit: weegt het onderscheid op tegen de baten die er tegenover staan? Wordt er onevenredig afbreuk gedaan aan het non-discriminatierecht?

Bovendien hoeft datgene wat juridisch mag, niet altijd maatschappelijk of politiek wenselijk te zijn. Zo besloot de rechter toe te staan dat de Koninklijke Marechaussee etniciteit mee mag nemen tijdens het, deels geautomatiseerd, vaststellen van de identiteit, nationaliteit en verblijfsstatus van burgers. De rechter oordeelde namelijk dat etniciteit een objectieve aanwijzing kan zijn voor iemands vermeende nationaliteit, en het nu eenmaal de taak van de Marechaussee is om, mede gegeven iemands nationaliteit, te controleren of de wet overtreden wordt. Voor het gebruik van etniciteit bood de wet dus de nodige ruimte. Toch heeft de Koninklijke Marechaussee naderhand besloten, onder andere vanwege het maatschappelijke vertrouwen, 'om etniciteit geen indicator in profielen of selectiebeslissingen meer te laten zijn'[2]. De ruimte die de wet biedt, betekent dat uitvoeringsorganisaties die algoritmes inzetten, afwegingen moeten maken over hoe zij rechtsregels het beste interpreteren en in code omzetten.

Want in de kern worden algoritmes ingezet om onderscheid te maken: welke groepen hebben als eerste hulp nodig, of bij wie is nader onderzoek nodig? Dat onderscheid mag in principe niet op beschermde gronden worden gemaakt. Maar hoe meet je of dat inderdaad niet gebeurt?

De wetenschap heeft daarvoor verschillende methodes ontwikkeld, die *fairness metrics* genoemd worden. De methodes verschillen in hun opvatting van *fairness,* eerlijkheid. Zo berekent de ene meetmethode of bepaalde groepen oververtegenwoordigd zijn in de uitkomsten, wat kan wijzen op historische bias in de datasets. Een andere methode houdt rekening met verschillen tussen groepen, om de kans op selectie gelijk te maken. Het lastige is echter dat de verschillende meetmethodes elkaar direct kunnen uitsluiten. Dat wil zeggen: als het algoritme aan de eisen van de ene meetmethode voldoet, kan het niet altijd tegelijkertijd ook voldoen aan de eisen van de andere methode. Een bepaald algoritme kan dus 'fair' zijn volgens de ene methode, en 'unfair' volgens de andere. Het punt is niet dat de ene *metric* correcter is dan de andere. Het punt is dat de keuze tussen *metrics* afhangt van je idee van eerlijkheid, en daarmee inherent politiek is. De keuze welke *metric* gebruikt wordt, moet helder uitgelegd worden aan de politiek, zodat er debat over gevoerd kan worden.

Reden 2: Discriminatie kan niet altijd opgespoord of voorkomen worden. Als eenmaal besloten is hoe oneerlijk of ongerechtvaardigd onderscheid wordt opgespoord, dan is het niet gegarandeerd dat een organisatie daarin slaagt. Dit probleem speelt vooral bij lerende algoritmische systemen, die in een bepaalde mate zelf op zoek gaan naar verbanden om voorspellingen te doen. Deze systemen kunnen immers onverhoopt proxyvariabelen hanteren die indirect blijken te discrimineren. Ze kunnen bijvoorbeeld een, tot dan toe, niet bekend verband ontdekken tussen creditcardgegevens en geslacht. Het systeem neemt de beschermde grond – geslacht – niet direct mee, maar kan via creditcardgegevens (de proxy) toch indirect onderscheid maken. Omdat lerende systemen nieuwe verbanden, en daarmee nieuwe

[2] https://www.volkskrant.nl/nieuws-achtergrond/marechaussee-stopt-met-etnisch-profileren-aan-de-grens-hoewel-het-van-de-rechter-wel-mag~bc543b9c/

proxyvariabelen, kunnen ontdekken, kan dit soort indirecte discriminatie er gaandeweg insluipen. De *fairness metrics* stellen een organisatie in staat om dit soort indirecte discriminatie op te sporen. Alleen, het waarnemen van discriminatie stelt een programmeur niet automatisch in staat het algoritme zo aan te passen dat voortaan alleen nog valide onderscheidingen worden gemaakt. In de toekomst kan het algoritme weer andere proxyvariabelen gaan gebruiken. Dat betekent dat niet alleen tijdens het ontwerp of in de pilotfase op bias moet worden gecontroleerd, maar dat dit continue moet worden gedaan – ook nadat het systeem *live* gaat. En dat betekent weer, dat er geen garantie gegeven kan worden dat het systeem tijdens het functioneren nooit zal discrimineren. Bovendien kan het nodig zijn om te beschikken over gegevens gerelateerd aan de beschermde gronden, juist om te controleren dat er geen discriminatie plaatsvindt. Maar deze gegevens, bijvoorbeeld over etniciteit, zijn lang niet altijd bij overheidsorganisaties beschikbaar. Het zijn gevoelige persoonsgegevens, die volgens de Algemene Verordening Gegevensbescherming niet zomaar verwerkt mogen worden. Overheidsorganisaties mogen dus niet altijd over deze gegevens beschikken.

De conclusie is helder: er kan vanwege het risico op indirecte discriminatie vooraf geen garantie gegeven worden dat lerende algoritmische systemen niet zullen discrimineren. De aard van de technologie brengt een discriminatierisico met zich mee. Beleid en politiek zetten nu in op goede ontwerprichtlijnen en het professionaliseren van het risicobeheer. Maar daarmee zijn de risico's te verkleinen, niet uit te bannen. De vraag is uiteindelijk welk risico politiek en maatschappelijk acceptabel is. Wanneer is de opbrengst van een algoritmisch systeem de risico's waard?

Handelingsopties voor beleid en politiek

Vanwege de onzekerheid over de precieze betekenis van discriminatie en de beperkte mogelijkheid om discriminatie te voorkomen, dreigt er een impasse: aan de ene kant verwacht de politiek van uitvoeringsorganisaties nul procent discriminatie. Aan de andere kant kunnen uitvoeringsorganisaties deze garantie niet geven. Wij geven beleid en politiek daarom vier handelingsopties mee.

1. Welk risico acht de politiek acceptabel? Allereerst is de vraag of en welk risico de politiek acceptabel vindt bij de inzet van lerende algoritmes. De politicus die garanties wil dat geen enkele discriminatie voorkomt, zal moeten afzien van lerende algoritmes. Er zijn alternatieven. Uitvoeringsorganisaties kunnen bij fraudedetectie bijvoorbeeld terugvallen op een aselecte steekproef. Dat betekent wellicht een lagere pakkans, of de inzet van meer menskracht. En dus wordt deze optie al snel afgedaan als duur en inefficiënt. Maar vaak is niet duidelijk hoeveel effectiever een profilerend systeem is, én wordt vergeten dat de vele beheersmaatregelen van een profilerend systeem ook met een prijskaartje komen. En soms is de kans op discriminatie op geen enkele manier uit te sluiten, omdat gekozen moet worden tussen menselijke en een algoritmische beoordeling, en beide kunnen zijn blootgesteld aan vooroordelen. Politiek en uitvoering dienen in dialoog de proportionaliteit van systemen te bespreken, waarbij expliciet aandacht is voor de kosten die risicobeheersing met zich meebrengt.

2. Keuzes bij de inzet van algoritmes. Als politiek en beleid bereid zijn enig risico te accepteren, en willen inzetten op lerende algoritmes, dan hoort daar ook een gesprek bij over het ‘sturen’ van algoritmes: bijvoorbeeld door fraudesystemen te laten focussen op grote overtredingen, of door juist in te zetten op algoritmes die mensen helpen hun recht te halen. Dit kan de risico’s voor burgers verkleinen. Het is zaak dat politiek en uitvoering spreken over welke *fairness metric* in een bepaalde maatschappelijke context juist wordt geacht, en uitvoeringsorganisaties transparant zijn over de gekozen *metrics*. De complexiteit van deze *metrics* moet voor de politiek toegankelijk worden gemaakt, zodat daarover debat plaats kan vinden. Vanwege de inherente onzekerheid, is het ook belangrijk dat politiek en uitvoering periodiek systemen evalueren: maken systemen hun beloftes waar, of verwezenlijken enkele risico’s zich toch? Dan is het zaak opnieuw afwegingen te maken – en het niet louter tot een uitvoeringsprobleem te bestempelen.
3. Professionaliseren en aanscherpen van het risicobeheer. De afgelopen jaren zijn er diverse richtlijnen en toetsingskaders ontwikkeld om risico’s tijdig te identificeren en te verminderen. Om die goed te gebruiken, is een breed scala aan maatregelen bij uitvoeringsorganisaties nodig, waaronder de invoering van een helder normenkader, dat duidelijk maakt wie bij de ontwikkeling van algoritmes voor welk aspect van mensenrechtenbescherming verantwoordelijk is, en het versterken van de ethische en juridische expertise van medewerkers, onder meer door trainingen en het opzetten van een ethische commissie. Het Rathenau rapport Algoritmes afwegen geeft een overzicht van deze maatregelen, en geeft aan hoe ze verder verscherpt kunnen worden.
4. Aanscherping wetgeving en toezicht. Een Kamermotie van lid Marijnissen roept het kabinet op om databases op te schonen en de nationaliteit van burgers niet mee te nemen bij risico-analyses. Tegelijkertijd zijn deze gegevens nodig om *bias* in algoritmes te detecteren. De aankomende EU Wet op de Artificiële Intelligentie (AI Act) voorziet daarom in deze mogelijkheid, mits de verzameling van gevoelige gegevens met de juiste waarborgen is omgeven. In het verleden heeft de Autoriteit Persoonsgegevens diverse keren geconstateerd dat de verwerking van persoonsgegevens van burgers door overheidsinstanties niet op orde was. Dat de juiste waarborgen daadwerkelijk worden gerealiseerd is dus niet vanzelfsprekend. Het is belangrijk dat politiek en uitvoering in gesprek gaan over de noodzakelijke waarborgen en waar nodig het toezicht aanscherpen..

Een andere optie is om de rechtspositie van individuen te versterken: het is nu voor burgers erg lastig om erachter te komen en aan te tonen dat ze door een algoritme zijn gediscrimineerd. Door meer inzicht te krijgen in achterliggende profielen, en de bewijslast te versimpelen en deels om te draaien, kan dat makkelijker worden gemaakt.

De laatste optie is om te kijken of handreikingen verplicht kunnen worden. De overheid heeft inmiddels nuttige handreikingen gepubliceerd ten aanzien van het ethisch en mensenrechtelijk gebruiken van algoritmes, zoals het Impact Assessment Mensenrechten en Algoritmen en de handreiking non-discriminatie. Een aangenomen motie roept het kabinet op het IAMA te verplichten. Dat geldt niet voor de handreiking non-discriminatie.

Tot slot – meer beschermde gronden nodig in de toekomst?

Het gebruik van lerende systemen brengt complexe risico-afwegingen met zich mee. De verwachting is dat deze afwegingen in de toekomst alleen maar complexer worden. Want lerende algoritmes zullen mensen op manieren kunnen onderscheiden en opdelen, die we nu niet kunnen voorspellen of zelfs kunnen bedenken. Dat kunnen gevoelige gronden zijn, zoals het inkomen van mensen. En dus is de politieke vraag die op tafel komt of er meer beschermde gronden, zoals etniciteit en geslacht, in de wet moeten worden opgenomen.

Dit bericht vormt een nadere verdieping van het rapport Algoritmes afwegen. Op deze pagina vindt u tevens een apart overzicht (download) van al onze aanbevelingen in één tabel. Met name aanbevelingen zes en zeven kunnen nog nader vormgegeven worden door het kabinet. Deze aanbevelingen gaan over de eisen die de wetgever kan stellen aan lerende algoritmes. Tevens wordt opgemerkt dat er nog weinig bekend is over de *effectiviteit* van lerende systemen.