1. SAMENVATTING

Dit rapport bouwt voort op meer dan tien jaar onderzoek door Amnesty International over het gebruik van digitale systemen, artificiële intelligentie (AI) en risicoprofileringssystemen in de overheidssector van Europese landen, en op onderzoek van andere organisaties uit het maatschappelijk middenveld in andere delen van de wereld. Dit geheel van studies toont aan hoe het gebruik van risicoprofileringssystemen in uiteenlopende contexten stelselmatig heeft geleid tot mensenrechtenschendingen, in het bijzonder in maatschappelijke domeinen waar het risico op schendingen hoog is, namelijk rechtshandhaving, sociale zekerheid en migratiebeleid (hierna: hoogrisicodomeinen). Amnesty International definieert risicoprofilering als een inschatting, evaluatie of berekening (soms 'voorspelling' genoemd) van de waarschijnlijkheid dat individuen of groepen een wet of regel zullen overtreden.

Dit rapport biedt een overzicht van systemische problemen bij gebruik van risicoprofilering en reikt een denkkader en argumentatie aan ter ondersteuning van verder onderzoek naar cases uit de praktijk. Daarnaast biedt het rechthebbenden, mensenrechtenverdedigers, ambtenaren, toezichthoudende autoriteiten, advocaten en rechters betrouwbare wetenschappelijke en juridische argumenten om het gebruik van risicoprofilering door staten en andere entiteiten in hoogrisicodomeinen aan te vechten.

Voor een volledige analyse en een goed begrip van de problemen die algoritmen voor risicoprofilering teisteren en die vaak resulteren in mensenrechtenschendingen, zijn inzichten uit verschillende wetenschappelijke disciplines vereist, en moet de technologie in haar historische en maatschappelijke context worden geplaatst. Voorts verkent het rapport de vaak onderbelichte structurele en intersectionele effecten van discriminerende risicoprofilering, en toont het aan hoe technologie discriminatie en ongelijkheid in stand kan houden door staten te voorzien van een valse schijn van objectiviteit. Daarnaast bespreekt het rapport enkele van de meest voorgestelde maatregelen voor het probleem van discriminerende profilering en wijst het op hun beperkingen. Tot slot brengt het rapport alle bevindingen samen in een juridische analyse en slaat het een brug tussen de wetenschappelijke literatuur en internationale mensenrechtennormen en jurisprudentie.

ONDERZOEKSVRAAG

Overheden zijn maar al te graag bereid om risicoprofilering te implementeren, ondanks tal van voorbeelden waaruit blijkt dat dergelijke systemen discriminatie in stand houden, en zonder na te gaan of het gebruik van risicoprofilering noodzakelijk is en in overeenstemming met internationale mensenrechtenstandaarden. Dat is deels te verklaren door de veronderstelling dat technologie, waaronder risicoprofilering, 'objectief' en 'neutraal' zou zijn. Deze schijn van technologische objectiviteit roept vragen op over de verantwoordingsplicht van de overheid wanneer schade ontstaat, en maakt het voor getroffen personen moeilijker om genoegdoening en rechtsherstel te krijgen.

Voorstanders van risicoprofilering stellen dat het organisaties in staat stelt objectievere en consistentere beslissingen te nemen, de doeltreffendheid van de besluitvorming te verhogen, de dienstverlening te stroomlijnen en de kostenefficiëntie te verbeteren door schaarse middelen efficiënter in te zetten voor toezicht. Deze beweringen zijn uitgegroeid tot het dominante narratief ter rechtvaardiging van het gebruik van risicoprofileringsalgoritmen.

Ondanks deze vermeende voordelen van risicoprofilering, zijn er talloze voorbeelden van mislukkingen, terwijl er opvallend weinig gepubliceerd bewijs bestaat van hun effectiviteit in de praktijk. Dit roept de vraag op of risicoprofilering ooit op een veilige manier kan worden ingezet in hoogrisicodomeinen. Om deze vraag te beantwoorden heeft Amnesty International de snelgroeiende interdisciplinaire wetenschappelijke literatuur over de impact van risicoprofilering onderzocht, evenals de voorgestelde technische en beleidsmatige oplossingen voor de aangetoonde tekortkomingen van risicoprofilering.

WIJDVERSPREIDE SCHADE EN MENSENRECHTENSCHENDINGEN

Risicoprofilering heeft geleid tot discriminatie op grond van onder meer ras en etniciteit, geslacht, socio-economische status en handicap. De risico's op discriminatie die verbonden zijn aan risicoprofilering worden bijzonder duidelijk wanneer ze worden bekeken door de lens van intersectionele discriminatie. Amnesty International onderzocht in verschillende landen diverse gevallen van schade door risicoprofilering, waarbij mensen werden gediscrimineerd op grond van een of meerdere intersecties van hun ras of etniciteit, nationale afkomst, gender, beperking of chronische ziekte, leeftijd en sociaaleconomische status.

Risicoprofilering veroorzaakt discriminerende effecten die verder reiken dan enkel discriminerende uitkomsten of dan de schending van individuele rechten. Het werkt namelijk ook meer structurele vormen van discriminatie in de hand. Risicoprofilering is geworteld in systemen die historisch zijn gebruikt voor het categoriseren, exploiteren en instrumentaliseren van gegevens over mensen, met als doel om maatschappelijke en raciale hiërarchieën tot stand te brengen en in stand te houden. Risicoprofilering kan dus als een verlengstuk van deze bestaande machtssystemen worden gezien.

In hoogrisicodomeinen brengt blootstelling aan risicoprofilering ernstige schade met zich mee, zowel van materiële als van immateriële aard. Slachtoffers ondervinden ernstig psychologisch leed, worden gestigmatiseerd en gaan gebukt onder valse beschuldigingen van fraude of criminaliteit, wat kan leiden tot uithuiszetting of zelfs gevangenisstraf. Vertraagde of geweigerde uitkeringen kunnen grote financiële gevolgen hebben, zoals problematische schulden. Mensen op de vlucht worden geconfronteerd met de dreiging van onterechte detentie en uitzetting. Gemeenschappen die al gemarginalisering worden, lijden onder een verlies van autonomie en angstaanjagende effecten van constante surveillance. Hierdoor verliezen mensen hun vertrouwen in instellingen, wat ernstige gevolgen heeft voor de legitimiteit van de overheid en voor de samenleving als geheel. Deze misstanden worden verergerd door een wijdverspreid gebrek aan transparantie, waardoor individuen machteloos staan om deze systemen aan te vechten of hun mensenrechten te verdedigen.

Risicoprofileringspraktijken brengen risico's met zich mee voor tal van andere mensenrechten. Deze schade kan eveneens een discriminerend karakter hebben, omdat het recht op non-discriminatie zowel een op zichzelf staand recht is als een recht dat op andere rechten van invloed is. Onderzoek door Amnesty International en andere mensenrechtenorganisaties heeft consequent aangetoond dat risicoprofilering een negatieve impact heeft op het recht op een eerlijk proces en de onschuldpresumptie, het recht op rechtsherstel en genoegdoening, het recht op privacy en gegevensbescherming, het recht op sociale zekerheid en op een behoorlijke levensstandaard, en op de volledige verwezenlijking van de menselijke waardigheid. Vanwege ongelijke behandeling is deze schade bovendien ongelijk verdeeld over maatschappelijke groepen, wat waarschijnlijk leidt tot specifieke nadelen voor mensen die deel uitmaken van groepen die al gemarginaliseerde worden.

BELANGRIJKSTE CONCLUSIES UIT DE WETENSCHAPPELIJKE LITERATUUR

De doelstellingen en processen die overheden volgen bij de ontwikkeling van risicoprofileringssystemen, vertonen gelijkenissen met die van het wetenschappelijk onderzoek: overheden proberen kennis te vergaren door middel van kwantitatieve analyse (inferentie). Bij risicoprofilering neemt deze kennis de vorm aan van een voorspelling of inschatting dat een persoon een regel zal overtreden. Deze inschatting of voorspelling is gebaseerd is op de mate waarin die persoon 'lijkt' op personen die in het verleden dergelijke overtredingen hebben begaan. Wetenschappelijk onderzoek hanteert echter strikte methoden om ongeldige praktijken en onbetrouwbare resultaten te voorkomen. Volgens de wetenschappelijke literatuur die Amnesty International heeft onderzocht en op basis van gesprekken met vooraanstaande wetenschappelijke experts, ontbreken dergelijke strikte methoden en praktijken bij risicoprofilering, zowel in de private sector als binnen overheidsdiensten. Dit ondermijnt de validiteit, robuustheid en betrouwbaarheid van risicoprofilering. Risicoprofilering wordt daardoor in een groeiende hoeveelheid wetenschappelijke publicaties door vooraanstaande AI-onderzoekers als pseudowetenschap bestempelt.

MEERDERE VORMEN VAN BIAS ZIJN INHERENT AAN RISICOPROFILERING

Enerzijds gebruiken overheden vaak reeds bestaande, ongeschikte administratieve data om voorspellende modellen te trainen (zogenaamde convenience samples), in plaats van nieuwe gegevens te verzamelen die specifiek of relevant zijn voor het voorspellen van het 'risicovolle' gedrag. Dit wijkt af van de methodologie die standaardpraktijk is binnen de kwantitatieve sociale wetenschappen.

Anderzijds is het zelfs met 'foutloze' data onmogelijk om een 'objectief' of 'neutraal' risicoprofileringsalgoritme te ontwerpen. Gegevens over mensen en de interpretatie daarvan zijn nooit 'objectief', omdat ze worden gevormd door de sociale, historische en institutionele context. Wanneer overheden sociale data uit het verleden gebruiken om te voorspellen wie een misdrijf zoals fraude zal plegen, richten zij hun aandacht onvermijdelijk op mensen die behoren tot groepen die historisch gezien zijn onderdrukt of gemarginaliseerd. Daarmee reproduceren ze en versterken ze onrecht uit het verleden.

Deze data-bias kan nooit volledig worden weggewerkt met technische middelen of door meer data toe te voegen. Dit komt doordat de bias voortvloeit uit maatschappelijke fenomenen die ten grondslag liggen aan de processen van het genereren en verzamelen van data. Het idee dat alle bias kan worden vermeden, is dus een misvatting. Dit leidt tot gebrekkige systemen die systemische discriminatie en ongelijkheid versterken onder het mom van technologische neutraliteit. Risicoprofilering verhult daarmee impliciete normen en bestaande ongelijkheden en leidt naar alle waarschijnlijkheid tot het in stand houden en versterken van stereotypen.

Bias is bovendien niet beperkt tot het type data of de selectie daarvan. Wanneer overheidsinstanties een voorspellingsdoel definiëren – zoals het risico op het plegen van socialezekerheidsfraude – verankeren zij institutionele prioriteiten en structurele vooroordelen rechtstreeks in de logica van het systeem. Deze optimalisatiedoelen dicteren hoe het algoritme werkt, wat betekent dat zelfs met 'foutloze' data het systeem nog steeds discriminatie zal veroorzaken als zijn kerndoelstelling leidt tot disproportionele controles van gemarginaliseerde bevolkingsgroepen.

RISICOPROFILERING VERTOONT METHODOLOGISCHE GEBREKEN

De literatuur over de wetenschappelijke validiteit van voorspellingsmodellen geeft een ernstige waarschuwing af over een gebrek aan legitimiteit.

Constructen zoals het individuele risico op het plegen van criminaliteit of socialezekerheidsfraude zijn extreem lastig om te voorspellen, omdat zij niet op betrouwbare wijze kunnen worden geoperationaliseerd en gemeten. Daarom wordt gebruikgemaakt van onnauwkeurige en bevooroordeelde proxy-indicatoren, zoals arrestaties als proxy voor recidive, of onopzettelijke fouten in uitkeringsaanvragen als proxy voor frauduleuze aanvragen. Er is daarnaast een fundamenteel gebrek aan betrouwbare referentiedata ('ground truth') over criminaliteit of fraude. Daardoor is de wetenschappelijke validiteit van risicoprofileringsmodellen fundamenteel zwak of zelfs onbestaande, met potentieel desastreuze gevolgen voor de betrokken personen. Dergelijke voorspellingsmodellen hebben een grote kans om vooringenomen en foutieve voorspellingen te maken, waardoor mensen worden blootgesteld aan willekeurige beslissingen.

Verder worden 'theorievrije' voorspellende toepassingen in de wetenschappelijke literatuur beschreven als methodologisch onverantwoord en potentieel gevaarlijk. De aantrekkingskracht van 'datagedreven' technieken, waaronder risicoprofilering, schuilt deels juist in hun vermogen om rechtstreeks uit grote datasets correlaties af te leiden zonder gebruik te maken van vooraf vastgelegde theoretische aannames. Echter, om te waarborgen dat voorspellingen gegenereerd door machine learning degelijk onderbouwd zijn, moeten de methodologische normen van de betrokken disciplines worden nageleefd. Dat houdt onder meer in dat een causaal plausibele theorie wordt geformuleerd, dat ze wordt ingebed in bestaande theoretische kaders en vervolgens aan de hand van empirische gegevens wordt getoetst aan de hand van gouden standaarden voor causale inferentie. De keuze van het op te lossen probleem, de optimalisatiedoelstellingen, de gegevens die worden verzameld en de wijze waarop dat gebeurt, de categorisering van die gegevens, de selectie en constructie van variabelen en de interpretatie van de modelresultaten worden stuk voor stuk gestuurd vanuit een bepaalde, vaak niet expliciet gemaakte theoretische invalshoek. Het doen van observaties of metingen is dus onlosmakelijk verbonden met een onderliggende theorie, ook al blijft deze op de achtergrond. Als deze theorie onbenoemd blijft, kan ze niet kritisch onderzocht worden. Dergelijke strenge methodes ontbreken flagrant in de overheidspraktijk van risicoprofilering.

Een andere essentiële praktijk van wetenschappers bestaat erin betekenisloze of schijncorrelaties uit te sluiten en de onderliggende causale mechanismen te onderzoeken. Bij risicoprofilering door overheden gebeurt dit niet, en wordt er daarnaast geen theorie gespecificeerd. Hierdoor is risicoprofilering niet robuust of deugdelijk onderbouwd, en geeft het een vertekend beeld van de menselijke realiteit waarin het wordt ingezet. Er is geen betrouwbare manier om te weten wanneer, waar of hoe de voorspellingen van het model zullen falen, wat tot schade voor mensen leidt. Arbitraire of willekeurige correlaties zullen ten onrechte worden geïnterpreteerd als causale verbanden en worden behandeld als empirische waarheden, waardoor impliciete ideeën en normen verhuld worden. Patronen in de sociale wereld weerspiegelen maatschappelijke normen, conventies en sociale structuren. Aan individuen of sociale groepen is niets inherent 'natuurlijk' of 'neutraal'.

GRENZEN VAN VOORSPELLING IN COMPLEXE SOCIALE SYSTEMEN

Mensen en hun bewust gedrag kunnen worden beschouwd als complexe adaptieve fenomenen, waardoor zij inherent onberekenbaar zijn. Deze conclusie wordt empirisch ondersteund door grootschalige studies uit de computationele sociale wetenschappen. Hieruit volgt dat bepaalde voorspellingen van menselijk gedrag simpelweg niet door machine learning kunnen worden gemaakt. Er bestaan situaties waarin geen enkel AI-systeem ooit kan werken. In sommige van deze gevallen kan er geen plausibel verband bestaan tussen waarneembare gegevens en het gedrag dat men wil voorspellen, zoals tussen ras of etniciteit en criminaliteit (etnisch profileren). In andere gevallen kunnen er, ongeacht de hoeveelheid gegevens, geen indicatoren of proxy-indicatoren bestaan die goed genoeg of objectief genoeg zijn om het onderliggende fenomeen adequaat te modelleren. Dit laatste geldt onder meer voor risicoprofileringssystemen die criminaliteit, levensloop of socialezekerheidsfraude proberen te voorspellen op het niveau van een individu of een specifieke locatie. Zulke voorspellingstoepassingen zijn weerlegd en worden afgedaan als ondeugdelijke wetenschappelijke praktijken.

Risicoprofileringssystemen roepen specifieke en ernstige normatieve bezwaren op. Als gevolg hiervan schieten deze systemen tekort volgens hun eigen maatstaven, omdat ze simpelweg geen nauwkeurige voorspellingen opleveren.

HET DOMINANTE NARRATIEF MIST ONDERBOUWING

Systemen voor risicoprofilering worden vaak genoemd als een methode waarmee staten de dienstverlening kunnen stroomlijnen, de kostenefficiëntie kunnen verbeteren, misdaden, inclusief fraude, kunnen voorkomen en migratie kunnen beheersen. Deze beweringen, die uitgaan van de premisse van schaarse middelen, zijn inmiddels weerlegd als rechtvaardiging voor beleid. Immers, ze worden empirisch niet ondersteund en zijn louter politiek bruikbare aannames die armoede en andere maatschappelijke problemen veranderen van politieke problemen tot een probleem van 'efficiëntie' dat vervolgens moet worden opgelost via automatisering en toezicht. In plaats van de beloofde voordelen blijkt een veel vaker voorkomend gevolg de bestraffing van de meest gemarginaliseerde groepen in de samenleving wanneer zij proberen toegang te krijgen tot hun rechten en/of essentiële overheidsdiensten. Amnesty International en andere organisaties hebben in talrijke casestudy's aangetoond dat risicoprofileringssystemen personen die al een of meerdere vormen van discriminatie of marginalisatie ervaren, onevenredig vaak geassocieerd worden met een hoger crimineel of financieel 'risico'.

VERONTRUSTENDE PARALLELLEN MET EUGENETICA EN WETENSCHAPPELIJK RACISME

In sommige omgevingen waarin machine learning wordt toegepast – zoals bij risicoprofilering door de overheid – vertonen algemeen aanvaarde normen opvallende gelijkenissen met die van vroege eugenetici. Eugenetici benadrukten de objectiviteit van de cijfers en methoden waarmee zij werkten en verdedigden het principe "laat de cijfers voor zichzelf spreken". Het beschouwen van correlatie als inherent voorspellend en het vermijden van theoretische onderbouwing en causaliteit waren kenmerkende praktijken van eugenetica en wetenschappelijk racisme. Daardoor verdoezelden zij bewust hun racistische ideeën en presenteerden zij cijfers als zijnde vrij van politieke waarden. Het is verontrustend dat dergelijke methoden om mensen te beoordelen en te rangschikken opnieuw hun intrede hebben gedaan op het beleidstoneel, zij het onder andere verklaarde ambities: om individuen te selecteren voor handhaving en controle.

CONCLUSIE UIT DE ACADEMISCHE LITERATUUR

Het opstellen van een risicoprofiel voor socialezekerheidsfraude of criminaliteit is geen realistische technische onderneming, en evenmin een geloofwaardige vorm van evidencebased beleid. Het is een poging om verdenking te operationaliseren in afwezigheid van betrouwbare referentiedata (*ground truth*). Dergelijke systemen zullen onvermijdelijk discrimineren, vanwege de bias die inherent zijn aan de gegevens en de sociale fenomenen die ze kwantificeren.

TOETSEN VAN RISICOPROFILERING AAN INTERNATIONALE MENSENRECHTENRECHTENNORMEN

Naast een overzicht van de wetenschappelijke literatuur, biedt dit rapport een juridische analyse en antwoord op de vraag of risicoprofilering leidt tot onderscheid op grond van verdachte gronden, en op de vraag of er voor dit onderscheid een redelijke en objectieve rechtvaardiging bestaat.

INGEBAKKEN ONDERSCHEID

Het gebruik van historische sociale data om criminaliteit of fraude te voorspellen, leidt er onvermijdelijk toe dat personen die behoren tot historisch onderdrukte of gemarginaliseerde groepen het doelwit worden. Dit komt neer op onderscheid op verdachte gronden. Dergelijk onderscheid wordt onder internationale mensenrechtennormen gekwalificeerd als discriminatie, als voor dit onderscheid geen redelijke en objectieve rechtvaardiging bestaat. Mensen worden vanaf hun geboorte ingedeeld in sociaal geconstrueerde hiërarchieën en krijgen ongelijke kansen toebedeeld, wat uiteindelijk leidt tot uiteenlopende levensuitkomsten. Vervolgens komen diezelfde mensen onder toezicht te staan van de instellingen die hen geacht worden te beschermen. Risicoprofilering in hoogrisicodomeinen voorspelt daarmee geen toekomstig gedrag – het reproduceert voornamelijk onrecht uit het verleden.

Veelvoorkomende, maar ontoereikende maatregelen voor bias in algoritmische systemen zijn onder meer het verwijderen van verdachte persoonskenmerken uit datasets, of het verwijderen van hun individuele proxy's – zoals postcode als proxy voor ras. Amnesty International concludeert dat deze maatregelen er niet in slagen de diepere, structurele problemen van data aan te pakken. Verdachte persoonskenmerken zullen impliciet weerspiegeld worden in andere, onderling verbonden variabelen, zoals uitgavenpatroon, woonsituatie of het niet komen opdagen bij medische afspraken. Deze variabelen kunnen dienen als indirecte proxy's, omdat systemische ongelijkheden en discriminatie in alle levensdomeinen doorwerken. Dit brengt een ongemakkelijke waarheid aan het licht: kenmerken die mogelijk relevant zijn voor risicoprofilering en ogenschijnlijk 'neutraal' zijn, zijn ongelijk verdeeld over verschillende groepen. Dergelijke verschillen in basisfrequenties (*base rates*) weerspiegelen vaak historische en voortdurende discriminatie op institutioneel of systemisch niveau.

MISDAAD- EN FRAUDEBESTRIJDING ALS DEKMANTEL VOOR INVASIEF TOEZICHT

Hoewel het terugdringen van fraude en andere vormen van criminaliteit een legitieme doelstelling van overheden is, is het risico reëel dat deze doelstellingen misbruikt worden als dekmantel voor invasieve surveillance van gehele gemeenschappen. Als er geen hard bewijs is dat de omvang van bijvoorbeeld uitkeringsfraude ingrijpende surveillance en risicoprofilering rechtvaardigt, is het belangrijk om kritisch te staan tegenover deze doelstellingen. Risicoprofilering wordt het vaakst ingezet in contexten waarin vooral groepen worden geraakt die al gestigmatiseerd of gemarginaliseerd worden, terwijl de meest bevoorrechte mensen in de samenleving grotendeels worden gevrijwaard van zulke controle. Deze selectieve aandacht en het gebruik van invasieve controlemiddelen vloeien vaak voort uit bestaande stereotypen en vooroordelen die met name gemarginaliseerde groepen voorstellen als inherent crimineel of gevaarlijk. Zulke stereotypen worden versterkt door structurele problemen, zoals etnisch profileren, overmatige politiecontrole (*over-policing*) en hogere veroordelingspercentages van geracialiseerde mensen.

IS RISICOPROFILERING NOODZAKELIJK EN EFFECTIEF?

Om een voorspellend systeem te toetsen aan het discriminatieverbod, moeten beleidsmakers

- Op duidelijke en transparante wijze maatstaven vaststellen die direct aansluiten bij de door hen nagestreefde, legitieme doelstellingen;
- De causale verbanden specificeren die inzichtelijk maken hoe voorspellingen tot het doel leiden;
- Bewijzen dat op voorspellingen gebaseerde maatregelen daadwerkelijk de beoogde maatschappelijke voordelen opleveren.

Het mag dus niet alleen maar draaien om het opschroeven van administratieve statistieken of het nastreven van een louter marktgedreven notie van efficiëntie, zoals een verhoogd aantal arrestaties of een grotere pakkans.

In plaats daarvan laten overheden zich vaak meeslepen door het narratief over de onvermijdelijkheid van technologische ontwikkelingen, en richten zij zich op misplaatste en beperkte opvattingen van 'effectiviteit' – bijvoorbeeld door effectiviteit gelijk te stellen de voorspellende prestaties van risicoprofileringsalgoritmen in vergelijking met willekeurige selectie. Dit is een oppervlakkige en onbetrouwbare maatstaf voor 'effectiviteit' in de context van toetsing aan het discriminatieverbod.

Onderscheid door risicoprofilering berust grotendeels op statistische correlaties en mist elke theoretische of empirische onderbouwing. Tenzij schijnverbanden rigoureus worden uitgesloten, moet ervan worden uitgegaan dat de correlatie een schijnverband is. Een correlatie zegt op zichzelf niets betekenisvols over de kans dat een individu of een groep een wet zal overtreden of fraude zal plegen. Daarom kunnen statistische generalisaties over verhoogde pakkans niet worden ingeroepen als rechtvaardiging voor disproportionele controles van om gemarginaliseerde groepen of individuen. Dit soort utilitaristische redeneringen is onverenigbaar met de waarden die belichaamd worden door non-discriminatiebepalingen.

De door Amnesty International onderzochte literatuur en casestudy's tonen aan dat risicoprofileringssystemen fundamenteel onnauwkeurig zijn en een zorgwekkende hoeveelheid vals-positieve opleveren. Deze vals-positieven komen bovendien hoofdzakelijk terecht bij geracialiseerde en gemarginaliseerde mensen. Overheden zijn, ondanks het overweldigende empirische bewijs, niet in staat deze ernstige operationele tekortkomingen te onderkennen. Bovendien zouden voorspellingen op basis van risicoprofilering, zelfs als ze accuraat zouden zijn, niet automatisch leiden tot doeltreffende interventies. Er is geen daadwerkelijke en aantoonbare daling van het aantal strafbare feiten na voorspellingen op basis van risicoprofilering. Dit ondergraaft de beoordeling van de effectiviteit van deze systemen onder mensenrechtenstandaarden.

Zelfs in de veronderstelling dat risicoprofilering nauwkeurige voorspellingen oplevert, toont onderzoek aan dat een hoge voorspellende nauwkeurigheid op zichzelf zelden leidt tot de verwezenlijking van het doel waarvoor voorspellingsmodellen wordt ingezet. Het is buitengewoon moeilijk om meetbare resultaten op korte termijn, zoals risicovoorspellingen, te koppelen aan bredere beleidsdoelstellingen. Onderzoek toont bijvoorbeeld aan dat *predictive policing* toepassingen er niet zijn geslaagd is om bredere beleidsdoelen te bereiken. Het is daarom van essentieel belang om de doelen die overheden stellen kritisch te bevragen. Hierbij moet de overheid aantonen dat zowel de beleidsdoelstellingen als de gekozen instrumenten daadwerkelijk bijdragen aan een overkoepelend maatschappelijk doel. Overheden moet niet worden toegestaan zich te baseren op al te brede generalisaties om inperkingen van fundamentele rechten te rechtvaardigen.

Kortom, de effectiviteit van risicoprofilering als beleidsinterventie staat op zijn zachtst gezegd ter discussie.

BEWUSTE STEREOTYPERING EN DE PERFORMATIEVE EFFECTEN VAN RISICOPROFILERING

Zelfs als een risicoprofiel enige, mogelijk niet verklaarde, voorspellende waarde heeft op groepsniveau, blijft het op individueel niveau grotendeels onjuist en bijgevolg onrechtvaardig. Het ontneemt mensen immers de mogelijkheid om af te wijken van wat als norm wordt beschouwd. Door mensen op basis van een risicovoorspelling te selecteren voor controles, *creëren* overheden zelf de causale gevolgen van die voorspelling. Mensen worden onvermijdelijk

behandeld *alsof* zij zich reeds verdacht hebben gedragen of zelfs de wet hebben overtreden. Wanneer mensen op basis van risicoprofilering aan extra controle worden onderworpen, worden zij dus de facto als verdachten behandeld. Correlaties worden daarbij beschouwd als inherent voorspellend, waardoor mensen worden bestraft omdat ze deel uitmaken van een (statistische) groep. Dit komt neer op stereotypering.

Dit zal onvermijdelijk leiden tot de vaststelling van een hoger aantal overtredingen binnen deze groepen en lagere aantallen binnen andere groepen. Dit fenomeen, dat uitvoerig is bestudeerd, resulteert uiteindelijk in vicieuze cirkels (feedback loops). Deze gegevens kunnen vervolgens opnieuw worden gebruikt als input voor het trainen van toekomstige risicoprofileringsmodellen. Het effect van risicoprofilering is daarom *performatief*: mensen of groepen worden omgevormd van statistische, hypothetische verdachten tot daadwerkelijke verdachten, waardoor bestaande vooroordelen worden bevestigd of nieuwe vooroordelen ontstaan.

Deze dynamiek wordt bovendien nog verergerd door bestaande systemische discriminatie. Aangezien correlatie op zichzelf geen oorzakelijk verband aantoont, en concepten zoals ras of etnische afkomst geen oorzakelijk verband vertonen met wetsovertreding, leggen deze correlaties eigenlijk eerder systemische discriminatie bloot dan 'risico's'.

ONDOELTREFFENDE EN ONTOEREIKENDE WAARBORGEN

Technische maatregelen om discriminerende resultaten te voorkomen of te herstellen, zoals *algorithmic fairness*, blijken in de praktijk ontoereikend te zijn. Het kernprobleem is dat statistische methodologie wordt behandeld als een vervanging voor het aanpakken van onderliggende sociale en structurele problemen. Dergelijke maatregelen zijn er consequent niet in geslaagd de mensenrechtelijke bezwaren weg te nemen en worden steeds vaker bekritiseerd in de wetenschappelijke literatuur. Technische waarborgen staan bovendien doeltreffende maatregelen, zoals wettelijke verboden, in de weg, en hollen daarmee uiteindelijk de verantwoordingsplicht van de overheid uit. Ook Betekenisvolle menselijke tussenkomst voorafgaand aan het definitieve besluit blijft een ontoereikende waarborg. Ten eerste omdat het onderscheid door de risicoselectie al heeft plaatsgevonden. Ten tweede vanwege de hierboven beschreven performatieve effecten. Tenslotte omdat het onvoldoende bescherming biedt tegen de nog altijd onopgeloste problematiek van *automation bias*.

CONCLUSIE: RISICOPROFILERING IS INHERENT DISCRIMINEREND

Gelet op de ernstige schade die risicoprofilering toebrengt aan mensen en gemeenschappen, afgewogen tegen het belang van de doelen, en de overkoepelende twijfels over de effectiviteit en de noodzaak ervan, kan het onderscheid dat inherent is aan risicoprofileringssystemen niet als proportioneel worden beschouwd en dus niet redelijk en objectief worden gerechtvaardigd. Hieruit volgt dat risicoprofilering in hoogrisicodomeinen discriminerend is en dus onverenigbaar is met internationale mensenrechtennormen.

HOOFDAANBEVELING

Amnesty International vindt dat het gebruik van systemen voor risicoprofilering, -voorspelling of -taxatie, zowel datagedreven als op regels gebaseerde, moet worden verboden in de hoogrisicodomeinen rechtshandhaving, sociale zekerheid en migratie, ongeacht de mate van menselijke betrokkenheid bij de uiteindelijke beslissing. Staten moeten AI-regelgeving ontwikkelen of bestaande wetgeving aanpassen om dit verbod te waarborgen. In afwachting van de invoering van dergelijke regelgeving, en ongeacht wijzigingen aan regelgeving, moeten overheidsinstanties dringend stoppen met het gebruik van deze systemen. Zie het laatste gedeelte van dit rapport voor een volledige lijst met aanbevelingen.