4. Profilerende en selecterende AI-systemen: risico's en de aselecte steekproef

SNEL NAAR DIT ONDERDEEL

Veel organisaties gebruiken algoritmes voor profilering of soortgelijke processen waarbij onderscheid tussen mensen wordt gemaakt. Dit hoofdstuk verkent dit onderwerp aan de hand van voorbeelden op het gebied van fraudedetectie. Belangrijk is deze algoritmes altijd als AI-systeem en daarmee onderdeel van een breder proces te zien. Het risico op discriminatie is een terugkerend thema in deze context. Op verschillende plekken in het proces rondom een profilerend AI-systeem kan discriminatie ontstaan. Bijvoorbeeld door niet-representatieve data en overmatig vertrouwen op algoritmische uitkomsten.
Een aselecte steekproef kan worden ingezet als beheersingsinstrument. Voordelen van deze techniek zijn dat algoritmes beter gemonitord kunnen worden en dat deze techniek waarborgt dat er een menselijke beslissing is in het proces. Het aanvullen van fraude-algoritmeprocessen met een aselecte steekproef in is in veel gevallen dan ook aan te bevelen.

4.1 Risicoprofilering en selectie

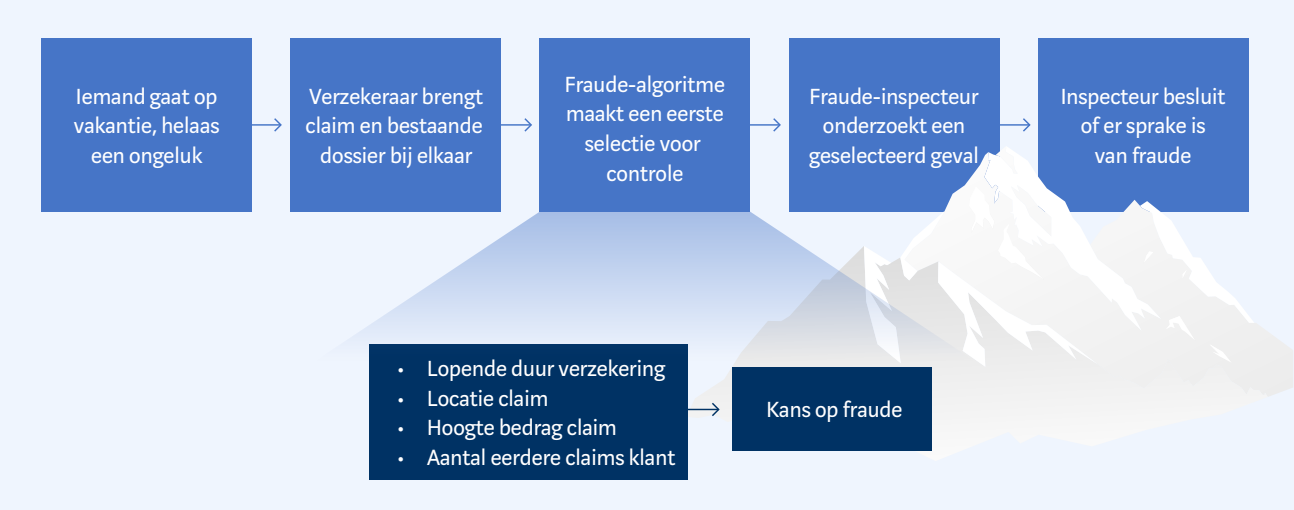
Veel organisaties zetten algoritmes in om te profileren en te selecteren. Op basis van gevallen uit het verleden maken organisaties een inschatting. Hiermee kunnen zij actie ondernemen, zoals een gericht onderzoek naar bepaalde personen. De inschatting dient dan als selectiemiddel om personen te selecteren die een inspecteur gaat onderzoeken.

Een voorbeeld van profilerende en selecterende systemen zijn fraudealgoritmes, waar vrijwel iedereen mee in aanraking komt. Verzekeraars zoeken met algoritmes naar verzekeringsfraude,[116] banken gebruiken fraudemodellen om transacties te controleren[117] en online platforms zoeken naar fraude binnen nieuwe gebruikersaccounts.[118] Vaak weet een burger, klant of gebruiker niet dat er een controle op fraude plaatsvindt. Zolang er geen verdenking is van fraude, blijft het fraudealgoritme een onzichtbare processtap.

“Het ontbreken van waarborgen in de uitvoeringspraktijk van risicogericht toezicht en de schending van wet- en regelgeving, hebben als gevolg dat sommige mensen meer kans hebben dan andere mensen om in beeld te komen bij de overheid en te worden gecontroleerd in het kader van fraudebestrijding.”
Parlementaire enquêtecommissie: Blind voor mens en recht, 7.1

Fouten bij fraudealgoritmes kunnen grote gevolgen hebben voor personen. Naast rechtmatigheidsvraag-stukken– mag een dergelijk algoritme in bepaalde situatie ingezet worden en mogen bepaalde indicatoren gebruikt worden – is het een essentieel aandachtspunt dat fouten in deze algoritmes grote gevolgen hebben. Dit is te zien in de toeslagenaffaire. Uit onderzoek van de Autoriteit Persoonsgegevens (AP) bleek dat in fraudeopsporing bij de kinderopvangtoeslag onrechtmatig gebruik is gemaakt van gegevens over dubbele nationaliteit en dat het een discri-minatoire verwerking van persoonsgegevens betrof.[119] De parlementaire enquêtecommissie bevestigde de schending van grondrechten en besloot dat hier sprake was van een schending van de eerbiediging van de persoonlijke levens-sfeer en op gelijke behandeling.[120] Een tweede voorbeeld is het Systeem Risico Indicatie (SyRI). SyRI werd ingezet om socialezekerheidsfraude te detecteren. De rechter oordeelde dat de methode in strijd is met het recht op respect voor privé- en familieleven. De rechtbank woog hierbij mee dat het systeem onvoldoende inzichtelijk en controleerbaar was, terwijl de inzet van SyRI (onbedoeld) wel discriminerende effecten met zich kon meebrengen.[121] Een derde en recenter

FIGUUR 4.1: EEN FRAUDEALGORITME IS ALTIJD ONDERDEEL VAN EEN BREDER PROCES. IN DEZE WEERGAVE IS HET FRAUDE-ALGORITME EEN VAN DE STAPPEN IN DE VERWERKING BIJ EEN SCHADECLAIM VOOR EEN REISVERZEKERING



voorbeeld is de controle op fraude met de uitwonende beurs voor studenten, waarbij DUO een selectiealgoritme gebruikte. In een onderzoeksrapport werd geconcludeerd dat er sprake was van discriminatie, waar het kabinet en DUO vervolgens hun excuses voor aanboden.[122] Dit komt in het geval van deze casus bovenop de discriminatie die samenhangt met het gebruik van indicatoren als opleidingsniveau als onderscheidende factor voor frauderisico.[123]

4.2 Discriminatie en overmatig vertrouwen

Risico's bij de inzet van fraudealgoritmes ontstaan vaak in de processtappen rondom het algoritme. De uitkomsten van een algoritme hangen af van de processtappen ervoor en de uiteindelijke impact hangt af van hoe er met de uitkomsten wordt omgegaan. Om dit te illustreren is een fictief voorbeeld van een fraudecontroleproces van een reisverzekeraar schematisch weergegeven.

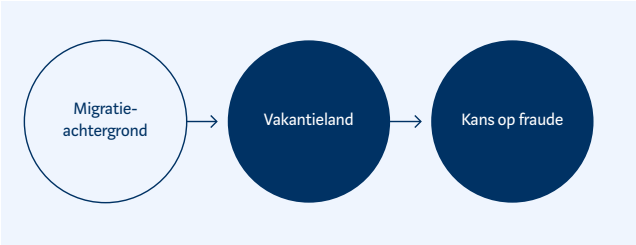
Het algoritme is hier afhankelijk van de ingediende claim en bestaande databronnen. Vervolgens dient het algoritme hier als een voorselectie in het fraudecontroleproces: een inspecteur onderzoekt vervolgens de gevallen met een hoge risico-indicatie (zie figuur 4.1).

Discriminatie is een belangrijk risico bij de inzet van fraudealgoritmes. Het logische doel van een algoritme is om onderscheid te maken, daarbij moet wel altijd bekeken worden of het maken van onderscheid juridisch wel te rechtvaardigen is. De selectie moet fraudeurs bevatten en juist géén mensen die niet frauderen. De term 'discriminatie' refereert hier aan artikel 21 van het Handvest van de Grondrechten van de Europese Unie, waarin staat: "Elke discriminatie, met name op grond van geslacht, ras, kleur, etnische of sociale afkomst, genetische kenmerken, taal, godsdienst of overtuigingen, politieke of andere denkbeelden, het behoren tot een nationale minderheid, vermogen, geboorte, een handicap, leeftijd of seksuele geaardheid, is verboden."[124] Een duidelijk voorbeeld van discriminatie is het benadelen van een kandidaat in een sollicitatieprocedure op basis van geslacht of nationaliteit.[125] Discriminatie kan zowel direct als indirect plaatsvinden. Bij indirecte discriminatie wordt er geen rechtstreeks onderscheid gemaakt op basis van een verboden grond. Bijvoorbeeld een postcode als selectiecriterium. Er zijn postcodegebieden waar veel mensen met een migratieachtergrond wonen. Als postcodegebieden met een hoog risico in een algoritme hiermee samenvallen, kan er sprake zijn van indirecte discriminatie op basis van migratieachtergrond.[126] Naast discriminatie zijn er ook andere grondrechtenrisico's bij de inzet van AI-systemen. Vaak zijn deze verbonden met het toepassingsgebied van het systeem. Zie ook box 4.1.

Niet-representatieve data zorgen voor oneerlijke uitkomsten. Een statistisch model is afhankelijk van de data waarmee het getraind wordt. Als een groep mensen weinig voorkomt in de data, zal het algoritme meer foute voorspellingen doen voor die groep. Dit leidt tot discriminerende uitkomsten wanneer een toepassing nadelig uitpakt.[127]

Voorbeeld: Stel dat de reisverzekeraar uit het eerdergenoemde schema geen goed systeem heeft voor claims die in een andere taal binnenkomen. Deze claims gaan eerst door het fraudealgoritme, maar de data komen zonder dat iemand dit doorheeft in een andere vorm in de database terecht. Het algoritme kan dan niet goed leren om met

FIGUUR 4.2: WANNEER EEN VERBODEN GROND (ONGEOBSERVEERD) EEN RELATIE HEEFT MET EEN INDICATOR IN HET ALGORITME, BESTAAT ER EEN VERHOOGD RISICO OP DISCRIMINATIE



claims in een andere taal om te gaan en zal hier meer fouten in maken. Dit heeft mogelijk discriminerende gevolgen.

Wanneer een beschermde groep wel voorkomt in de data maar niet op een representatieve manier, kan dit ook leiden tot discriminatie.

Voorbeeld: Stel dat het proces bij de verzekeraar enige tijd onbewust vooringenomen verliep. Hierdoor zijn mensen in een beschermde groep vaker aangeduid als fraudeur. Het algoritme zal dit patroon overnemen en deze groep in de toekomst een te hoog risico toewijzen.

Een negatieve feedbackloop verergert discriminatie. Het algoritme leert van eerdere fraudegevallen om nieuwe gevallen te vinden. Wanneer deze nieuwe gevallen na verloop van tijd gebruikt worden om weer van te leren, ontstaat er een feedbackloop. Deze feedbackloop kan discriminatie verergeren door selectiebias.

Voorbeeld: Stel dat volgens de data claims uit een bepaald land vaker frauduleus zijn. Claims uit dat land krijgen dan een hogere risico-indicatie en zullen vaker worden onderzocht. Intensiever zoeken naar fraude, is op zichzelf al een manier om daar ook meer fraude te vinden. Wanneer het algoritme opnieuw getraind wordt, zal dit land dan ook een nog sterkere focus krijgen. Dit herhaalt zich: er is een zelfversterkende feedbackloop ontstaan.

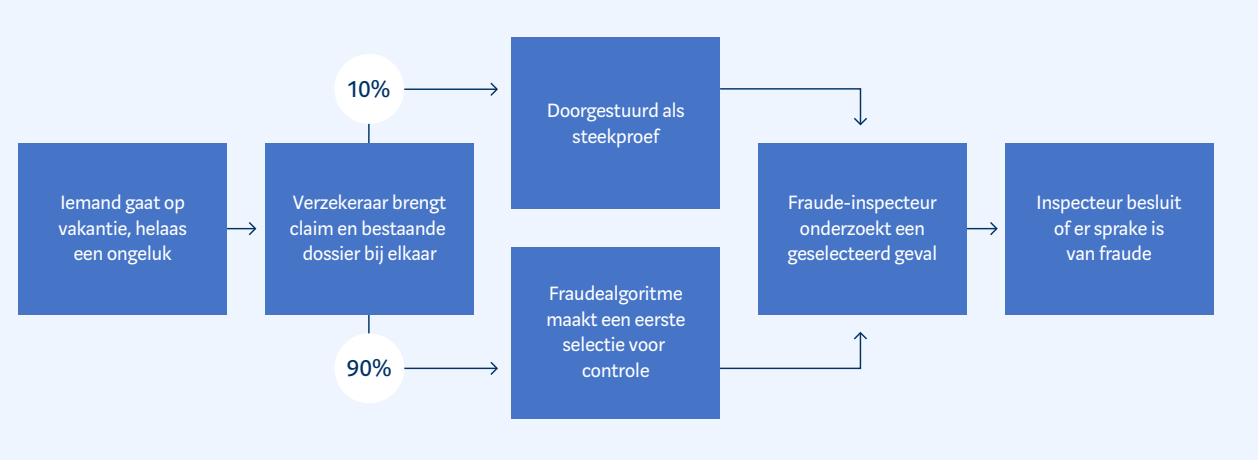
Voldoende en correcte data hebben is geen garantie voor non-discriminatie. Kenmerken voor het beoogde onderscheid zijn namelijk vaak door proxies gerelateerd aan de verboden gronden voor onderscheid.

Voorbeeld: Stel dat migratieachtergrond invloed heeft op het land waar mensen op vakantie gaan. En daarmee indirect ook de inschatting van het algoritme van de kans op fraude. Migratieachtergrond wordt niet geobserveerd (zie figuur 4.2). Ervan uitgaande dat migratieachtergrond een verboden grond is om onderscheid op te maken, zal dit algoritme mogelijk ongewenst discrimineren. Zelfs wanneer de data die gebruikt zijn voor de training van het algoritme een perfecte afspiegeling zijn van de werkelijkheid.

Overmatig vertrouwen op algoritmes is een belangrijk risico. Er is interactie tussen de menselijke beoordelaar en de uitkomst van een algoritme. In plaats van de uitkomst te controleren, hebben mensen de neiging om de uitkomst snel voor waarheid aan te nemen: "de computer zal het wel weten." Dit fenomeen wordt ook wel automation bias genoemd. Daardoor lijkt het alsof er een waarborg is met menselijke tussenkomst, terwijl deze in werkelijkheid beperkt effectief is.

Overmatig vertrouwen op algoritmes geeft ruimte aan discriminatie. Zoals besproken, kan er op verschillende manieren discriminatie ontstaan in de voorspellingen van fraudealgoritmes. De menselijke tussenkomst dient onder andere als waarborg tegen discriminatie. Als menselijke beoordelaars overmatig vertrouwen op algoritmes, kunnen discriminerende voorspellingen worden overgenomen. Het is essentieel dat de menselijke beoordelaar kritisch blijft kijken naar de uitkomst van een algoritme.

FIGUUR 4.3: EEN VERZEKERAAR KAN EEN STEEKPROEF INZETTEN DOOR EEN PERCENTAGE VAN DE SCHADECLAIMS WILLEKEURIG TE SELECTEREN, EN DEZE ONAFHANKELIJK VAN HET ALGORITME TE STUREN VOOR ONDERZOEK.



4.3 Aselecte steekproef

Een mogelijke maatregel om de genoemde risico's te verminderen is de aselecte steekproef. Wanneer een aselecte steekproef wordt toegepast, wordt een deel van de gevallen willekeurig geselecteerd om onderzocht te worden op fraude. In de afbeelding is een aselecte steekproef van 10% weergegeven. (zie figuur 4.3).

Wat het optimale percentage gevallen is dat de steekproef moet selecteren, verschilt per fraudealgoritme. Een belangrijke overweging hierbij is dat de steekproef pas kan dienen als referentie als er voldoende gevallen onderdeel van uitmaken.

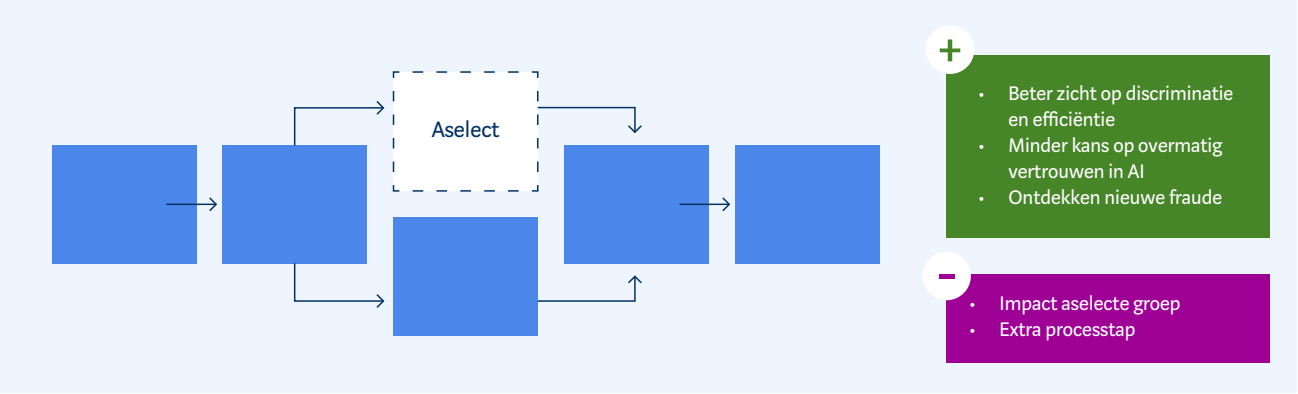
Met de aselecte steekproef als referentie kan een deel van de discriminatierisico's worden gemonitord. Door de steekproef als referentie te gebruiken, kan bijvoorbeeld worden gevolgd of een groep niet onevenredig naar voren komt door het algoritme. Het belang van willekeurige selectie voor het verkrijgen van representatieve data is eerder bijvoorbeeld beschreven door de EU Agency for Fundamental Rights (FRA).[128] Ter illustratie wederom het voorbeeld van de reisverzekering.

Voorbeeld: stel dat het fraude-algoritme van de reisverzekeraar voor 95% gevallen selecteert met nationaliteit Y. Er wordt een steekproef ingezet, en van de fraude die binnen de steekproef wordt gevonden blijkt slechts 15% van de mensen nationaliteit Y te hebben. Het lijkt er hier op dat het algoritme onevenredig veel nadruk legt op het controleren van mensen met nationaliteit Y.

Door de aselecte steekproef kan het risico op overmatig vertrouwen worden verminderd. Wanneer een steekproef een deel van de selectie voor fraudeonderzoek verzorgt, zijn niet alle onderzoeken meer op basis van een hoogrisicosignaal. Hiermee is het proces zo in te richten dat de inspecteur niet weet of een te onderzoeken geval aselect geselecteerd is. Daardoor kan de inspecteur niet meer blind uitgaan van de algoritme-uitkomsten, maar wordt de inspecteur aangemoedigd kritisch te zijn.

Op welke groep de aselecte steekproef van toepassing is, verschilt per context. In sommige toepassingen zal een fraudeindicatie geen selectie betekenen, maar juist een uitsluiting. Het blokkeren van een online bestelling is hier een voorbeeld van. Bij zo'n geval hoort een ander procesoverzicht. Een steekproef kan hier worden ingezet door aselect hoogrisicogevallen door te laten. Er zijn ook toepassingen waarbij de impact op de willekeurig geselecteerde personen significant is. Denk aan huisbezoeken als onderdeel van een fraudeonderzoek. Hierbij staat een belangenafweging centraal: een steekproef inzetten betekent niet dat je zomaar bij iemand mag binnenstappen.

FIGUUR 4.4: DE BESLISSING OM EEN ASELECTE STEEKPROEF IN TE ZETTEN HANGT AF VAN DE GEVOLGEN VOOR HET HELE PROCES



Naast risicovermindering draagt een aselecte steekproef bij aan het meten van efficiëntie en verkennen van nieuwe soorten fraude. De referentie die een steekproef verzorgt, kan worden gezien als basis om de prestaties van het algoritme mee te vergelijken. Daarnaast zal het element van willekeur ervoor zorgen dat onbekende en nieuwe vormen van fraude in de loop van de tijd ook voorkomen in de data. Een nieuwe versie van het algoritme kan hier vervolgens van leren.

4.4 Het overwegen waard

De beslissing om een aselecte steekproef in te zetten, hangt af van de gevolgen voor het hele proces. De techniek raakt aan verschillende onderdelen in het proces en kan dus niet worden afgewogen tegen een op zichzelf staand risico. Om de beslissing inzichtelijk te maken, kan een systeem zonder steekproef worden vergeleken met een systeem met steekproef (zie figuur 4.4).

De aselecte steekproef draagt in veel gevallen bij aan een verantwoordere inzet van profilerende en selecterende AI-systemen. Bij de inzet van dit soort AI-systemen is de aselecte steekproef dan ook het overwegen waard.

De inzet van een aselecte steekproef raakt aan de AI-verordening en de AVG. Een systeem rondom een fraudealgoritme moet (met of zonder steekproef) aan de wetgeving voldoen. Een profilerend of selecterend AI-systeem verwerkt persoonsgegevens en dus is de AVG van toepassing. Daarnaast treedt de komende jaren de AI-verordening in werking. Binnen de AI-verordening is het relevant of een AI-systeem als hoogrisicosysteem classificeert.

Voor bijvoorbeeld een fraudealgoritme voor essentiële overheidsuitkeringen en -diensten zal dit zo zijn. Aanbieders van hoogrisicosystemen zijn verplicht de redelijkerwijs te voorziene risico's vast te stellen en te beperken. Potentiële discriminatie is zo'n risico. De aselecte steekproef kan hierbij worden ingezet als beheersingsinstrument.

Bepaalde methoden om discriminatie tegen te gaan zijn afhankelijk van de verwerking van bijzondere persoonsgegevens. Om bijvoorbeeld te meten of een groep met een bepaalde afkomst anders behandeld wordt, zal informatie over afkomst verwerkt moeten worden. Deze gevoelige gegevens, 'bijzondere persoonsgegevens' genoemd, krijgen extra bescherming in de AVG. De verwerking van bijzondere persoonsgegevens is verboden, tenzij er een uitzondering is. De AI-verordening kan voor AI-systemen met een hoog risico onder strikte voorwaarden voorzien in een uitzondering voor het verwerken van bijzondere persoonsgegevens om discriminatie op te sporen of tegen te gaan.

Box 4.1

Grondrechtenrisico's bij de inzet van AI-systemen

Het gebruik van AI-systemen kan direct en indirect leiden tot schending van grondrechten. Dit risico speelt zowel bij zeer simpele algoritmes zoals beslisbomen als complexe systemen, bijvoorbeeld op basis van neurale netwerken.

Een bekend risico bij AI-systemen is te zien bij het recht op non-discriminatie (artikel 21 van het Handvest), maar er zijn ook risico's voor andere grondrechten. Non-discriminatie vraagstukken springen vaak in het oog omdat AI-systemen in veel gevallen immers juist worden ingezet om onderscheid te maken. Daarbij moet nadrukkelijk bekeken worden of het maken van onderscheid juridisch wel te rechtvaardigen is. Maar er is bijvoorbeeld ook het grondrecht op rechtvaardige arbeidsomstandigheden (artikel 31 van het Handvest). Dit kan onder druk komen te staan door algoritmisch management. Bijvoorbeeld wanneer dit afbreuk doet aan gezonde, veilige of waardige arbeidsomstandigheden. Een ander voorbeeld is de bescherming van persoonsgegevens (artikel 8 van het Handvest), een grondrecht dat waarborgt dat gegevens eerlijk worden verwerkt met toestemming van de betrokkene of op basis van een wettelijke grondslag. AI-systemen werken op basis van grote hoeveelheden data, waaronder vaak ook persoonsgegevens. Deze persoonsgegevens moeten rechtmatig en in lijn met dit grondrecht worden verwerkt, ook wanneer het een AI-systeem betreft. De AP ziet als onafhankelijke autoriteit toe op de naleving van deze regels.

Twee andere relevante grondrechten zijn vrijheid van informatie en het recht op behoorlijk bestuur. Beide grondrechten komen terug in deze rapportage en zijn relevant in de context van AI-systemen. Het grondrecht dat raakt aan de vrijheid en de pluriformiteit van de media is van belang bij de inzet en het gebruik van AI in de online informatievoorziening (zie hoofdstuk 2). Het recht op behoorlijk bestuur waarborgt dat zaken onpartijdig, billijk en binnen een redelijke termijn behandeld moeten worden door overheidsinstellingen en organen. De inzet van simpele algoritmes en AI-systemen in het publieke domein heeft de afgelopen jaren tot risico's en incidenten geleid die moeilijk samen gaan met dit recht.

Grondrechtenrisico's en schendingen bij de inzet van AI-systemen zijn nog te vaak buiten beeld. Ook wanneer specifieke wetgeving niet alle aspecten van nieuwe technologie adresseert, of wanneer er nog geen specifieke wetgeving is, zullen grondrechten altijd moeten worden eerbiedigt en beschermd. Om hier nader op in te gaan komt de AP later dit jaar met een factsheet over grondrechtenrisico's bij de inzet van AI-systemen.