Algorithm Audit

Publieke standaard – Betekenisvolle menselijke tussenkomst bij risicoprofileringsalgoritmes

Voorkomen van uitsluitend op profilering gebaseerde besluitvorming

Mei 2025

Samenvatting

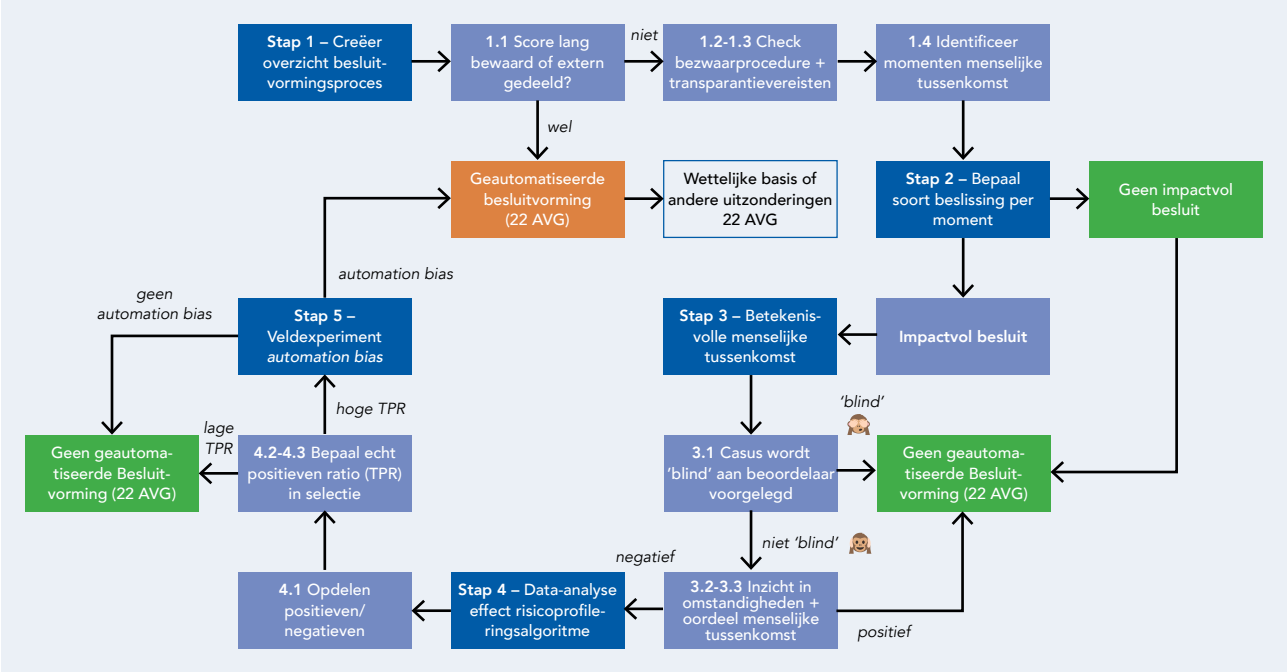
Een volledig geautomatiseerd besluit is volgens artikel 22 lid 1 Algemene Verordening Gegevensbescherming (AVG) verboden. Dit document biedt een pragmatisch stappenplan hoe met dit verbod kan worden omgegaan voor risicoprofileringsalgoritmes. De beschreven stappen bundelen adviezen uit eerder gepubliceerde documenten en zijn aangevuld met praktijkervaring van Algorithm Audit met mens-algoritme interactie in de publieke en private sector. 'Blinde' beoordeling – waarbij beoordelaar niet weten of een casus willekeurig of door een risicoprofileringsalgoritme is geselecteerd – staat in de standaard centraal. Naast dergelijke kwalitatieve waarborgen wordt toegelicht hoe data-analyse ondersteuning kan bieden om verboden geautomatiseerde besluitvorming te voorkomen. Deze publieke standaard verbindt recente jurisprudentie, in het bijzonder de Schufa-uitspraak van het Hof van Justitie van de Europese Unie, met de uitvoeringspraktijk. Dit stappenplan is ook relevant voor naleving van artikel 14 van de AI-verordening dat toeziet op menselijk toezicht bij de inzet van AI-systemen. De standaard richt zich uitsluitend op toepassing van risicoprofileringsalgoritmes in zowel het publieke als private domein. Andere vormen van geautomatiseerde besluitvorming en profilering vallen buiten de reikwijdte van deze publieke standaard.

Deze publieke standaard is een verdieping op de publieke standaard Profileringsalgoritmes van Algorithm Audit.[1] Onderstaand stappenplan wordt geïntegreerd in Q7 van de open-source AI-verordening implementatie tool.[2]

Overzicht stappenplan

Doorloop onderstaande stappen om geautomatiseerde besluitvorming bij risicoprofilering te voorkomen. Stap 1-5 worden in het vervolg van dit document toegelicht.

[1] Publieke standaard Profileringsalgoritmes, Algorithm Audit (2024).
[2] AI-verordening implementatie tool, Algorithm Audit (2025); Implementatie van de AI-verordening – Definitie van een AI-systeem, Algorithm Audit (2025).



Toelichting stappenplan

Met het Schufa-arrest licht het Hof van Justitie van de Europese Unie (HvJEU) toe hoe het verbod op geautomatiseerde besluitvorming, zoals belegd in artikel 22 Algemene Verordening Gegevensbescherming (AVG), bij gebruik van risicoprofilering geïnterpreteerd moet worden. De hoogste rechter van de EU oordeelt dat er sprake is van uitsluitend op profilering gebaseerde besluitvorming als: 1) er sprake is van een beslissing, 2) dat uitsluitend op profilering is gebaseerd en 3) dat voor de betrokkene rechtsgevolgen heeft of op een andere manier aanmerkelijke gevolgen heeft.[3] Stap 1-5 gaat na of aan deze cumulatieve eis wordt voldaan.

Beoordelen of er sprake is van uitsluitend op profilering gebaseerde besluitvorming is niet alleen een kwalitatieve exercitie. Het effect van profileringsalgoritmes op het besluitvormingsproces kent ook een empirische dimensie. In de Schufa-uitspraak overweegt het HvJEU dat er sprake is van een *"geautomatiseerde vaststelling van een [score] die is gebaseerd op persoonsgegevens"* en dat *"een ontoereikende [score] er in bijna alle gevallen toe [leidt] dat [de bank] weigert het gevraagde krediet te verlenen"*.[4] Aan de hand van onderzoek naar praktijktoepassingen van algoritme-gedreven besluitvorming, in het bijzonder in het Controle Uitwonendenbeurs (CUB)-proces van de Dienst Uitvoering Onderwijs (DUO)[5] en een machine learning-gedreven risicotaxatie algoritme ingezet door een commercieel autodeelplatform[6], wordt in Stap 4 een empirische werkwijze toegelicht hoe kan worden vastgesteld in welke mate de uitkomst van een risicoprofileringsalgoritme wordt gevolgd door beoordelaars. Dit empirische inzicht kan de afweging of in *"bijna alle gevallen"* het advies van het algoritme door een beslismedewerker wordt gevolgd informeren. Stap 5 beschrijft een empirische methode om na te gaan of er sprake is van automation bias in het besluitvormingsproces.

Voor deze publieke standaard is het Schufa-arrest[7], het Consultatiedocument Betekenisvolle menselijke tussenkomt van de Autoriteit Persoonsgegevens (AP)[8], het Advies artikel 22 AVG en geautomatiseerde selectie-instrumenten van de AP[9], het Advies over geautomatiseerde selectietechniek van Pels Rijcken,[10] richtsnoeren van de European Data Protection Board (EDPB)[11] en juridisch-wetenschappelijke literatuur,[12] [13] geraadpleegd.

---

3 ECLI:EU:C:2023:220, Zaak C-634/21, Hof van Justitie van de Europese Unie (2023).
4 Supra noot 3, randnummer 47 en 48.
5 Addendum Vooringenomenheid voorkomen, Algorithm Audit (2024).
6 Nader te publiceren algoprudentiële casus.
7 Supra noot 3.
8 Consultatiedocument Betekenisvolle menselijke tussenkomst, Autoriteit Persoonsgegevens (2025).
9 Advies artikel 22 AVG en geautomatiseerde selectie-instrumenten, Autoriteit Persoonsgegevens (2024).
10 Advies over geautomatiseerde selectietechniek, Pels Rijcken (2024).
11 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01), European Data Protection Board.
12 Rechtsbescherming tegen risicoprofilering op basis van de AVG, het EVRM en het Handvest, F. Çapkurt, Nederlands Juristenblad (2025).
13 The Right to an Explanation in Practice: Insights from Case Law for the GDPR and the AI Act, L. Metikos en J. Ausloos, Law, Innovation and Technology (2025).

Box 1

Advies Autoriteit Persoonsgegevens over artikel 22 AVG en geautomatiseerde selectie-instrumenten

In licht van de Schufa-uitspraak en het daaruit volgende advies van de landsadvocaat[10] heeft de AP een advies uitgebracht hoe artikel 22 AVG in de context van risicoprofileringsalgoritmes geïnterpreteerd dienen te worden.[9] De AP stelt in dit advies dat risicoprofilering zonder specifieke wettelijke voorzieningen toegepast kan worden, wanneer aan de volgende vijf voorwaarden is voldaan:[14]

i. onderzoek discriminatoire verwerkingen en stel, indien nodig, mitigerende maatregelen in;
ii. onderzoek periodiek of discriminatie zich voordoet;
iii. garandeer dat gevolgen van risicoprofilering pas intreden na betekenisvolle menselijke tussenkomst, zie Stap 3-5 van deze standaard;
iv. voorkom dat risicoprofilering niet andere aanmerkelijke gevolgen heeft, zie Stap 1.1 van deze standaard;
v. maak toepassing van risicoprofilering kenbaar aan betrokkenen.

Deze publieke standaard geeft een praktische invulling aan voorwaardes iii-iv. van het AP advies. Voor stappen i-ii. kan gebruik worden gemaakt van de publieke standaard Profileringsalgoritmes van Algorithm Audit.[15]

Bovenstaand advies van de AP is niet onomstreden. Het wijkt op cruciale punten af van de juridische analyse van de landsadvocaat, onder meer in de interpretatie van wat geldt als een besluit met aanmerkelijke gevolgen. Andere deskundigen wijzen op een eenzijdige lezing van de AVG, in het bijzonder in relatie tot andere Europese wetgeving, zoals het Europees Verdrag van de Rechten van de Mens (EVRM) en het Handvest voor de Grondrechten van de Europese Unie (Handvest).[12] Desalniettemin heeft het kabinet het advies van de AP integraal overgenomen.[16] Algorithm Audit erkent zowel kritiek op het AP-advies als het initiatief om concreet handelingsperspectief te bieden bij algoritme-gedreven besluitvormingsprocessen. Door het uitbrengen van de publieke standaarden Profileringsalgoritmes en Betekenisvolle menselijke tussenkomst hoopt Algorithm Audit een bijdrage te leveren hoe in de praktijk risicoprofileringsalgoritmes verantwoord kunnen worden ingezet.

Daarnaast dient opgemerkt te worden dat het centraal stellen van 'blinde' beoordeling door beslismedewerkers in Stap 3 van deze standaard afwijkt van het AP advies. De AP stelt op p.6 van haar advies dat: "*Om betekenisvol te zijn en daadwerkelijk gevolgen zo nodig te kunnen voorkomen ofwijzigen, moet degene die 'tussenkomt' voldoende kunnen beoordelen of selectie in een bepaald geval terecht is. De behandelend ambtenaar zal daarom moeten weten hoe het geautomatiseerde proces (selectieregel/algoritme/technologie) werkt en moeten begrijpen hoe en op welke wijze deze (de totstandkoming) van het uiteindelijke besluit vormt en beinvloedt*". Algorithm Audit weegt het risico van *automation bias* bij deze werkwijze (de neiging van beoordelaars om het advies van het risicoprofileringsalgoritme over te nemen zonder daarbij kritisch na te denken) zwaarder dan het belang dat beoordelaars begrijpen hoe het risicoprofilerings werkt, aangezien dit niet de primaire verantwoordelijkheid van een eerstelijns-, maar van tweedelijns algoritmedeskundigen zou moeten zijn.

[14] Supra noot 9, p.14-17.
[15] Supra noot 1.
[16] Kamerstukken II 2024/25 2024D47487.

Stappenplan

Het uitvoeren van de Stap 1-5 helpt verboden geautomatiseerde besluitvorming bij inzet van risicoprofileringsalgoritmes te voorkomen maar biedt hiervoor geen garantie, omdat het afhankelijk is van de wijze waarop de stappen worden uitgevoerd en van de keuzes die hierin worden gemaakt.

Stap 1 – Creëer overzicht besluitvormingsproces

1.1 Stel een schematisch overzicht op van het gehele besluitvormingsproces. Ga na of de uitkomsten van het risicoprofileringsalgoritme voor langere termijn bewaard of intern of extern gedeeld worden waaruit 'aanmerkelijke gevolgen'[17] kunnen voortvloeien.

1.2 Stel vast dat belanghebbenden in het besluitvormingsproces tegen het oordeel in bezwaar kunnen gaan.

1.3 Ga na dat toepassing van een risicoprofileringsalgoritme voldoende kenbaar is voor betrokkenen, bijv. middels vermelding van verwerkte gegevens door algoritme middels brief of pop-up notificatie in app van platform of dienst.

1.4 Stel vast op welke momenten in het besluitvormingsproces mogelijk sprake zou moeten zijn van betekenisvolle menselijke tussenkomst.

LET OP: wanneer in Stap 1.1 is vastgesteld dat er sprake is van het voor langere termijn opslaan of delen van de uitkomsten van een risicoprofileringsalgoritme dan kan er gemakkelijk sprake zijn van aanmerkelijke gevolgen voor betrokkenen. Dit is zonder specifieke wettelijke voorziening en bijbehorende waarborgen verboden.[18] Wanneer dit het geval is kan niet meer door het opvolgen van de volgende stappen in deze standaard voorkomen worden dat sprake is van verboden volledig geautomatiseerde besluitvorming volgens artikel 22 lid 1 AVG. De huidige werkwijze kan alleen worden vervolgd wanneer een van de uitzonderingen in artikel 22(2) AVG geldt en wanneer voldoende waarborgen zijn ingeregeld.

Stap 2 – Bepaal soort beslissing

Of een risicoprofileringsalgoritme valt onder het verbod in artikel 22 AVG hangt af van het effect van de beslissing die, geïnformeerd door de uitkomst van het algoritme, genomen wordt. Ga na of er aan de hand van het algoritme een beslissing wordt genomen die een 'rechtsgevolg' heeft voor betrokkenen of hen op een andere manier 'in aanmerkelijke mate treft'. Dit is bijvoorbeeld het geval als een van de volgende type beslissingen wordt genomen[19]:

i. Een formeel besluit, bijvoorbeeld opleggen van een belastingaanslag, het toekennen of afwijzen van een uitkering of toeslag, een beslissing volgend op bezwaar, het toekennen of afwijzen van een vergunning of subsidie;

ii. Een beslissing met financiële gevolgen, zoals de mogelijkheid tot het krijgen van een betalingsregeling of in aanmerking komen voor krediet;

iii. Een overeenkomst sluiten, bijvoorbeeld een arbeidsovereenkomst of koopovereenkomst;

iv. Selectie voor controle, wanneer die controle ingrijpend is voor de betrokkene zoals een huisbezoek;

---

[17] Supra noot 9, p.6-8 en p.16. Wanneer de uitkomsten van een algoritme bewaard of gedeeld wordt voor andere toepassingen volgen al gauw aanmerkelijke gevolgen voor betrokkenen. Zo kunnen bewaarde of gedeelde uitkomsten van een risicoprofileringsalgoritme een langdurig of blijvend effect op de betrokkene hebben. Bijvoorbeeld omdat een opgeslagen risicoselectie of risicoscore herhaaldelijk tot onderzoek leidt, of door er een zichzelf versterkend effect van opeenvolgende onderzoeken. Door uitkomsten te delen kan een risicoselectie of risicoscore een eigen leven gaan leiden, bijvoorbeeld doordat derde partijen de uitkomst gebruiken op een manier die niet voorzien is, die niet de juiste waarborgen kent of door het opstellen van een 'zwarte lijst'. Zo gebruikten verschillende gemeentes de uitkomsten van het algoritme 'Preselect Recidive' om lijsten op te stellen om jongeren te monitoren. Zie Follow The Money-artikel.

[18] Supra noot 9, p.16.

[19] Supra noot 8, 10, 11 en 13.

v. Een beslissing die iemands toegang tot onderwijs raakt, bijvoorbeeld toelating tot een universiteit en toewijzing van scholen;
vi. Beslissingen die iemands kans op werkgelegenheid beïnvloedt, bijvoorbeeld verwerking van sollicitaties en het toewijzen van opdrachten aan zelfstandigen;
vii. Betrokkene op andere wijze in aanmerkelijke mate treft.

Het bovenstaande is geen uitputtende lijst. Er zal altijd een context afhankelijke afweging moeten worden gemaakt of een beslissing aanmerkelijke gevolgen heeft voor een betrokkene. Bij deze afweging moeten potentiële gevolgen (kansen/risico's) worden meegewogen. Gevolgen hoeven dus niet al hun intreden te hebben gedaan en gevolgen hoeven ook niet altijd hetzelfde te zijn voor alle betrokkenen.[20]

Voorbeelden van beslissingen zonder rechtsgevolg of aanmerkelijke gevolgen zijn:

i. Uitdelen van een waarschuwing;[21]
ii. Prioritering van aanvragen, verzoeken, klachten, zonder invloed op de afhandeling hiervan;
iii. Selectie voor controle, wanneer die controle niet ingrijpend is voor de betrokkene. De AP stelt dat het moeten verstrekken van nadere informatie ten behoeve van een controle geen aanmerkelijk gevolg is voor een betrokkene.[22]

Er is geen sprake van verboden geautomatiseerde besluitvorming als uit het algoritme-gedreven besluitvormingsproces geen aanmerkelijke gevolgen voortvloeien voor betrokkenen. Algorithm Audit raadt aan om in dit geval alsnog maatregelen voor betekenisvolle menselijke tussenkomst in te richten (zie Stap 3), ondanks dat dit geen wettelijke verplichting meer is.

LET OP: indien het algoritme-gedreven besluitvormingsproces beperkt onderbouwd en/of gedocumenteerd wordt ingezet kan dit indirect – via inbreuk op fundamentele rechten (zoals eerbiediging van de persoonlijke levenssfeer of gelijke behandeling) – betrokkenen alsnog in aanmerkelijke mate treffen en daarmee binnen de reikwijdte van artikel 22 AVG vallen. Zie ook publieke standaard Profileringsalgoritmes van Algorithm Audit.[23]

Stap 3 – Betekenisvolle menselijke tussenkomst

3.1 Stel vast of alle casussen 'blind' met beoordelaars worden gedeeld. Bij 'blinde' selectie is het voor de beoordelaar niet bekend hoe (door een risicoprofileringsalgoritme, willekeurig of andere vorm) de casus is geselecteerd. De beoordelaar kan dit ook niet uit de context afleiden. Bij blinde beoordeling wordt de beoordelaar niet beïnvloed door de uitkomst van het algoritme.

3.2 Verkrijg inzicht in de omstandigheden waarin beoordelaars een beslissing moeten nemen. Daarbij kunnen de volgende vragen helpen:

i. Op basis van welke informatie dienen beoordelaars een besluit te beoordelen, of tegen het algoritme in te gaan?
ii. Hoeveel data krijgen beoordelaars te zien tijdens het maken van een besluit?
iii. Welke eisen worden er aan beoordelaars gesteld om een besluit te kunnen nemen?
iv. Hoeveel tijd hebben beoordelaars doorgaans voor het beoordelen van de uitkomst van een algoritme? Hoe verhoudt dit zich tot de aard van het te nemen besluit?

[20] Supra noot 9, p.8.
[21] Let op: wanneer hier een formeel besluit wordt genomen is dit wél een beslissing met een rechtsgevolg.
[22] Supra noot 9, p.8.
[23] Supra noot 1.

3.3 Stel vast of er sprake is van betekenisvolle menselijke tussenkomst door een beoordelaar. Ga na of de volgende vragen positief kunnen worden beantwoord:[24]

- **i.** Indien geen sprake is van blinde beoordeling (zie Stap 3.1): Begrijpen beoordelaars hoe en op basis van welke gegevens het profileringsalgoritme tot een resultaat komt?
- **ii.** Indien geen sprake is van blinde beoordeling (zie Stap 3.1): Zouden beoordelaars het besluit ook zonder profileringsalgoritme kunnen maken?
- **iii.** Hebben de beoordelaars voldoende tijd om een afweging te maken?
- **iv.** Kunnen beoordelaars specifieke omstandigheden meenemen in hun beoordeling die een algoritme niet meeweegt? Voorkom dat profileringskenmerken zowel in het risicoprofileringsalgortime als in de werkinstructies zijn opgenomen. [25]
- **v.** Hebben de beoordelaars de mogelijkheid om elkaar of een leidinggevende om hulp te vragen?
- **vi.** Vinden er kwaliteitscontroles plaats op het werk van beoordelaars?
- **vii**. Wordt het algoritme aangepast na feedback van beoordelaars, betrokkenen, of monitoring?

Wanneer in Stap 3.1 wordt vastgesteld dat casussen blind worden gedeeld met beoordelaars, dan is de beoordeling geheel onafhankelijk van het risicoprofileringsalgoritme. Het besluit is dan niet uitsluitend op profilering gebaseerd. Er is dan geen sprake van verboden geautomatiseerde besluitvorming.

Als de vragen uit Stap 3.3 over het besluitvormingsproces positief kunnen worden beantwoord is het aannemelijk dat beoordelaars rekening houden met andere factoren dan enkel de uitkomst van het risicoprofileringsalgoritme. Het besluit is dan niet uitsluitend op profilering gebaseerd omdat er sprake is van betekenisvolle menselijke tussenkomst.[26]

Stap 4 – Data-analyse effect risicoprofileringsalgoritme

LET OP: als je in Stap 3.1 hebt vastgesteld dat casussen blind aan beoordelaars worden gedeeld, dan hoeft deze stap niet uitgevoegd te worden.

4.1 Deel de uitkomsten van het risicoprofileringsalgoritme, voordat deze aan een beoordelaar ter besluitvorming worden voorgelegd, op in twee categorieën, bijvoorbeeld een 'hoog risico'- en een 'minder hoog risico'-categorie. Naar de categorie 'hoog risico' wordt verwezen als *positieven*, naar de categorie 'minder hoog risico' wordt verwezen als *negatieven*.[27]

4.2 Ga na welk deel van de aan beoordelaars ter besluitvorming voorgelegde cases tot de 'hoog risico'-categorie (positieven) en 'minder hoog risico'-categorie (negatieven) behoort.

4.3 Bepaal de echt positieven ratio: deel het aantal cases waarvan de beoordelaar instemt met vervolgactie zoals aanbevolen door het risicoprofileringsalgoritme (echt positieven) door het aantal door het risicoprofileringsalgoritme als 'hoog risico' aangemerkte cases (positieven).

4.4 Bepaal of ondernomen vervolgactie na menselijke tussenkomst hoofdzakelijk bestaat uit door het risicoprofileringsalgoritme als 'hoog

[24] 10 vragen uit Stap 3.2 en Stap 3.3 zijn door Algorithm Audit geselecteerd als de meest relevante vragen uit de in totaal 93 opgenomen vragen in het Consultatiedocument Betekenisvolle Menselijke Tussenkomst van de AP. De andere vraag is opgenomen op basis van praktijkervaring van Algorithm Audit; infra noot 25.

[25] In het CUB-proces van DUO werd bijvoorbeeld zowel in het risicoprofileringslalgoritme als in de werkinstructies onderscheid gemaakt op het kenmerk 'afstand tot ouder(s)'.

[26] Artikel 29 Werkgroep, Richtsnoeren inzake geautomatiseerde individuele besluitvorming en profilering voor de toepassing van Verordening (EU) 2016/679, 2017, p.24.

[27] In het CUB-proces van DUO werd de uitkomst van het risicoprofileringsalgoritme (een risicoscore tussen de 0 en 180) opgedeeld in een risicocategorie 'hoog risico' (score tussen 60-180) en een 'minder hoog risico'-categorie (score tussen 0-59).

risico' aangemerkte cases (hoge echt positieven ratio). Bij een te hoge echt positieven ratio moet de betekenis van de menselijke tussenkomst worden bevraagd. Herhaal in dit geval Stap 3 en voer Stap 5 uit.

Aan de hand van Stap 4.1-4.4 kan worden vastgesteld hoe vaak beoordelaars tegen de voorspelling van het algoritme ingaan. Een best-practice is om niet enkel door het risicoprofileringsalgoritme als 'hoog risico' geclassificeerde cases (blind) aan beoordelaars voor te leggen, maar deze aan te vullen met een op voorhand vastgestelde verhouding willekeurig geselecteerde casussen en eventueel aangevuld met mogelijke andere vormen van selectie, zoals signaal-gedreven casussen.[28]

[28] Signaal-gedreven cases worden geselecteerd naar aanleiding van signalen uit andere delen van de organisatie. Bij bedrijven kan dit bijvoorbeeld de financiële afdeling zijn waar achterstallige betalingen worden gemonitord. Bij gemeenten kan dit de afdeling Werk en Inkomen zijn die de afdeling Handhaving en Toezicht informeert over mogelijk verdachte situaties.

Box 2

Voorbeeld Stap 4 – Bepalen van echt positieven ratio

Stap 4 wordt middels een voorbeeld uitgewerkt. In dit voorbeeld wordt ervan uitgegaan dat alle betrokkenen in de doelpopulatie een score door het risicoprofileringsalgoritme krijgen toegekend.

4.1 Van de in totaal 13 betrokkenen worden 5 door het risicoprofileringsalgoritme als 'hoog risico' (positief) geclassificeerd.

4.2 Niet alleen de 5 positieven, ook 2 negatieven worden blind aan een beoordelaar voorgelegd.

4.3 Uit de menselijke tussenkomst volgt dat:

> 3 van de 5 positieven terecht als positieve zijn geclassificeerd (echt positief);
> 2 van de 5 positieven ten onrechte als positieve zijn geclassificeerd (vals positief);
> 1 van de 2 negatieven terecht als negatieve is geclassificeerd (echt negatief);
> 1 van de 2 negatieven ten onrechte als negatieve is geclassificeerd (vals negatief).

Merk op dat deze verhouding bijgesteld moeten worden nadat de uitkomst van een mogelijke bezwaarprocedure van een betrokkenen bekend is.

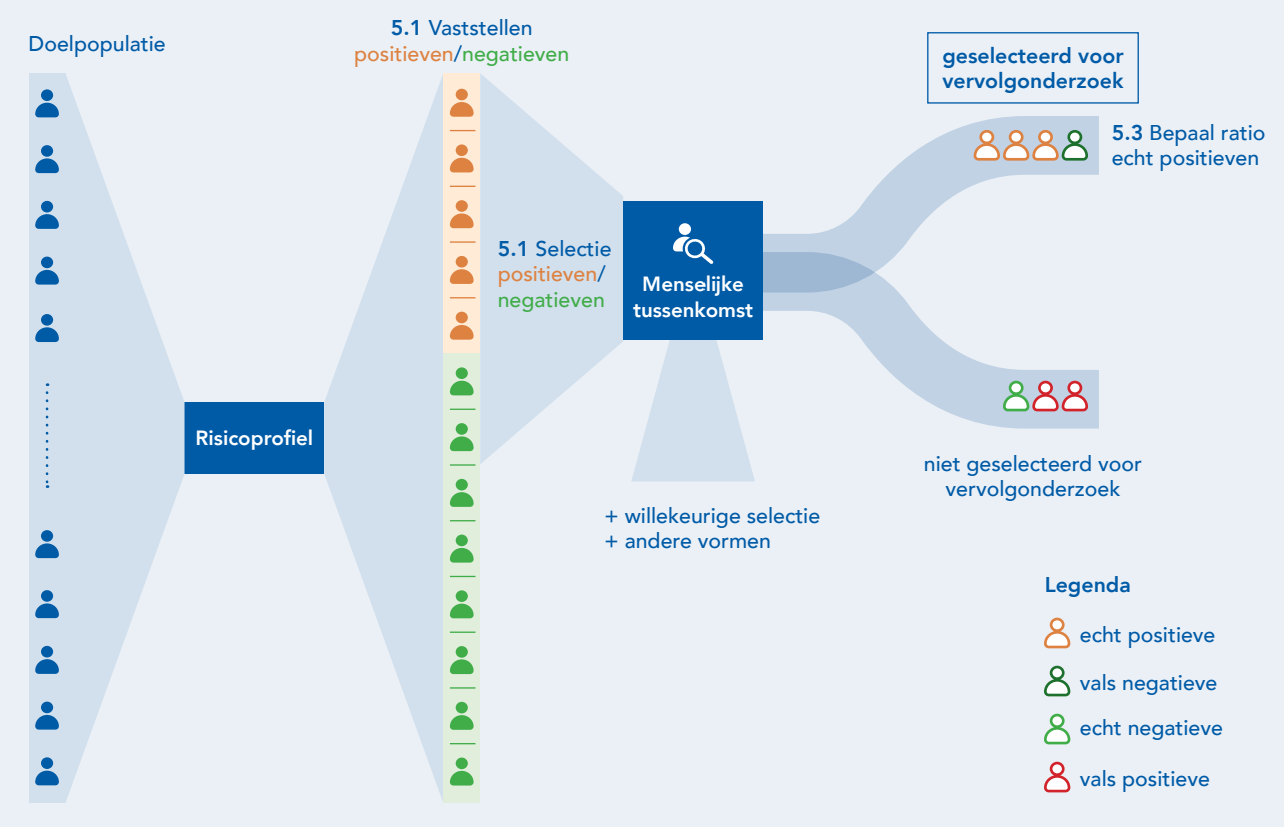
4.4 Het vaststellen van de echt positieven ratio (3/5=60%) kan helpen bij het maken van de afweging of er sprake is van betekenisvolle menselijke tussenkomt. Hoe hoger de echt positieven ratio hoe waarschijnlijker dat er geen sprake is van betekenisvolle menselijke tussenkomst. In dat geval is er mogelijk sprake van (verboden) volledige geautomatiseerde besluitvorming.

Praktijkvoorbeeld met betrekking tot Stap 4.2: In het CUB-proces van DUO in 2014 hadden 2.400 van de 3.179 geselecteerde studenten voor huisbezoek het label 'hoog risico' toegekend gekregen door het risicoprofileringsalgoritme (75%). 640 studenten die door een beoordelaar zijn geselecteerd voor een huisbezoek hadden het label 'laag risico' en 140 hadden het label 'onbekend risico' toegekend gekregen door het risicoprofileringsalgoritme (resp. 20% en 5%).[29] Niet alleen als 'hoog risico' geclassificeerde gevallen werden dus geselecteerd voor vervolgonderzoek.

[29] Supra noot 5 p.40. Door afronding op tientallen wijkt de som der delen af van het totaal.

Praktijkvoorbeeld met betrekking tot Stap 4.4: Een beoordelaar van een autodeelplatform besluit in 40-50% van de gevallen tegen de aanbeveling van een risicoprofileringsalgoritme in te gaan en een gebruiker van het platform geen waarschuwing te sturen voor risicovol rijgedrag. De beoordelaar is getraind, heeft duidelijke werkinstructies en voldoende tijd om een afweging te maken. Ook kunnen gebruikers in beroep gaan tegen de beslissing en kunnen zij de verwerkte data bij het platform opvragen. Er is sprake van betekenisvolle menselijke tussenkomst. In deze stap van het besluitvormingsproces vindt geen verboden geautomatiseerde besluitvorming plaats.[30]

[30] Supra noot 6.



Figuur 1 - Bepalen hoe vaak beoordelaars afwijken van een door het algoritme aanbevolen vervolgactie kan ondersteunend zijn bij het bepalen of er sprake is van betekenis menselijke tussenkomst

Stap 5 – Veldexperiment *automation bias*

LET OP: als in Stap 4 is vastgesteld dat beoordelaars vaak tegen de voorspelling van het algoritme ingaan, dan hoeft deze stap niet uitgevoerd te worden. Mocht dit niet het geval zijn, dan kan het volgende experiment worden uitgevoerd om na te gaan of een beoordelaar door het zien van een door een risicoprofileringsalgoritme gegenereerd label beïnvloed wordt in het nemen van een beslissing.[31]

5.1 Formuleer de volgende hypotheses:
- > H_0: Zichtbaarheid van een door een risicoprofileringsalgoritme gegenereerd label voor een casus beïnvloedt de beslissing die wordt genomen door de beoordelaar;
- > H_A: Zichtbaarheid van een door een risicoprofileringsalgoritme gegenereerd label voor een casus beïnvloedt de beslissing die wordt genomen door de beoordelaar niet.

5.2 Selecteer een aantal (bijvoorbeeld 10) realistische casussen. Deze mogen ook fictief zijn. Zorg dat de uitkomsten bekend zijn, bijvoorbeeld 'hoog risico'- of 'minder hoog risico'-categorie.

5.3 5.3 Deel beoordelaars op in twee groepen (bijvoorbeeld 10 beoordelaars per groep). Groep A krijgt het door het risicoprofileringsalgoritme gegenereerd label niet te zien. Groep B krijgt dit label wel te zien. Groep A is de controlegroep.

5.4 Bepaal de steekproefgrootte: bepaal hoe vaak van de in Stap 5.2 gegenereerde casussen aan de in Stap 5.3 opgedeelde groepen met beoordelaars worden voorgelegd.[32]

5.5 Voer het volgende experiment uit:
- > Groep A (controlegroep): krijgt de casussen te zien met alle relevante informatie voor beoordeling en de werkinstructie die ook in het reguliere proces gebruikt wordt. Deze groep krijgt geen door het risicoprofileringsalgoritme gegenereerd label te zien. Zij beoordelen de casus op basis van deze informatie;
- > Groep B: krijgt de casussen te zien met daarbij de informatie "deze casus is door het risicoprofileringsalgoritme beoordeeld als hoog risico" of "deze casus is door het risicoprofileringsalgoritme beoordeeld als minder hoog risico". De risicocategorie is willekeurig toegewezen: de helft van groep B krijgt een casus met label 'hoog risico' en de andere helft van groep B krijgt een casus met label 'minder hoog risico'. Bij het toekennen van risicocategorieën wordt dus niet het risicoprofileringsalgoritme gebruikt.[33] Daarbij krijgt de beoordelaar ook in deze groep alle relevante informatie voor beoordeling en de werkinstructie die ook in het reguliere proces gebruikt wordt. Zij beoordelen de casus op basis van deze informatie, inclusief de risicocategorie.

5.6 Label, op basis van de beslissingen van de beoordelaars, correcte beslissingen met 1 en incorrecte beslissingen met 0 voor zowel groep A als B. Of een casus correct beoordeeld is, hangt af van de echte uitkomsten die in Stap 5.2 zijn vastgesteld. Bereken het percentage correcte beslissingen in groep A en in groep B.

5.7 Pas een tweezijdige Z-toets toe om te bepalen of er sprake is van een statistisch significant verschil tussen correcte beoordeling in groep A en groep B. In lijn met Stap 5.1, zijn de volgende hypotheses hierbij van toepassing:
- > H_0: Proportie correct beoordeelde casussen is gelijk in groep A als in groep B, of te wel $p_A \neq p_B$;
- > HA: Proportie correct beoordeelde casussen is groter is groep A dan in groep B oftewel $p_A > p_B$

5.8 Accepteer of verwerp H_0. Houd een significantieniveau aan van $p<0.05$.

---

[31] Het veldexperiment beschreven in Stap 5 is geïnspireerd door Kamerstukken 2023/24 2024D17779.

[32] Zie Random sample size for single and multiple hypothesis tests, Algorithm Audit (2024).

[33] Door uitkomsten willekeurig toe te wijzen, is dit experiment niet afhankelijk van de kwaliteit of werking van een bestaand algoritme.

Over Algorithm Audit

Algorithm Audit is een Europees kennisplatform voor AI bias testing en normatieve AI-standaarden.
De doelen van de stichting zijn drieledig:

Kennisplatform	Samenbrengen van kennis en experts voor collectief leerproces over verantwoorde inzet van algoritmes, bijvoorbeeld ons AI Policy Observatory en position papers
Normatieve adviescommissies	Adviseren over ethische kwesties in concrete algoritmische toepassingen door het samenbrengen van deliberatieve, diverse adviesommissies, met algoprudentie als resultaat
Technische hulpmiddelen	Implementeren en testen van technische methoden voor bias-detectie en -mitigatie, zoals onze bias detection tool
Projectwerk	Ondersteuning bij specifieke vragen vanuit de publieke en private sector over de verantwoorde inzet van algoritmes.

Structurele partners van Algorithm Audit

SIDN Fonds

Het SIDN Fonds staat voor een sterk internet voor iedereen. Het Fonds investeert in projecten met lef en maatschappelijke meerwaarde, met als doel het borgen van publieke waarden online en in de digitale democratie.

European AI&Society Fund

Het European AI&Society Fund ondersteunt organisaties uit heel Europa die AI beleid vormgeven waarin mens en maatschappij centraal staan. Het fonds is een samenwerkingsverband van 14 Europese en Amerikaase filantropische organisaties.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Het ministerie van BZK maakt zich sterk voor een democratische rechtsstaat, met een slagvaardig bestuur. Ze borgt de kernwaarden van de democratie. BZK staat voor een goed en digitaalvaardig openbaar bestuur en een overheid waar burgers op kunnen vertrouwen.

Opbouwen van *publieke kennis*

over verantwoorde AI *zonder winstoogmerk*

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Algorithm Audit

Stichting Algorithm Audit is geregistreerd bij de
Kamer van Koophandel onder nummer 83979212