Rijksorganisatie voor Ontwikkeling, Digitalisering en Innovatie
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Bias-toetsing 'Kort Verblijf Visa' aanvragen

April 2023

5.1.2e (Rijks ICT Gilde)

Versies

versie	status	titel	datum	auteurs	wijzigingen t.o.v. vorige versie
0.1	Concept	Bias-toetsing 'Kort Verblijf Visa' aanvragen	03-03-2023	5.1.2e	-
1.0	Definitief	Bias-toetsing 'Kort Verblijf Visa' aanvragen	06-04-2023		Kleine aanpassingen in het document op basis van feedback BZ; additionele bijlage met reactie op feedback BZ.

Inhoudsopgave

Managementsamenvatting

Het Ministerie van Buitenlandse Zaken (BZ) heeft het Rijks ICT Gilde (RIG) verzocht een bias-toetsing uit te voeren van de Buitenlandse Zaken Analyse Omgeving (BAO), als onderdeel van het Informatie Ondersteunend Beslissen Kort Verblijf Visa (IOB/KVV proces).

In een rapport van 24 mei 2022 concludeerde Privacy Management Partners onder meer dat het IOB/KVV-proces als behoorlijk in de zin van de AVG kan worden beschouwd op voorwaarde dat er geen bias blijkt te zitten in de BAO.[1]

De vraag of er sprake is van bias in de BAO kan worden beantwoord aan de hand van een aantal deelvragen:

1. **Welk typen bias kunnen zich voordoen binnen het IOB/KVV proces (paragraaf 2.3)?**

Bias kan zich in principe in alle stappen in het IOB/KVV proces voordoen. Specifiek voor de BAO geldt dat mogelijk bias zijn intrede kan doen in (i) de door de BAO gebruikte databronnen o.a. in de vorm van *historical bias*, (ii) tijdens het modelleren en het trainen van de BAO o.a. in de vorm van *confirmation bias*, (iii) als gevolg van de ingebruikname van de BAO o.a. in de vorm van *feedback bias*, en (iv) in het dagelijks gebruik van de BAO door medewerkers o.a. in de vorm van *automation bias*.

De lijst met mogelijke typen bias is lang, maar op basis van de beschikbare data voor deze bias-toetsing kan slechts worden geanalyseerd wat het netto-effect van al deze potentiële bronnen van bias tezamen is. Geconstateerde bias kan om die reden niet direct worden terug herleid tot individuele oorzaken.

2. **Welke definitie van bias wordt gehanteerd voor dit onderzoek (paragraaf 3.2)?**

Voor het uitvoeren van dit onderzoek wordt de volgende bias definitie gehanteerd: de mate waarin, voor groepen met verschillende demografische kenmerken (nationaliteit, geslacht en leeftijd), uiteindelijk goedgekeurde aanvragen desondanks een hoog-risico (*intensive track*) aanduiding vanuit de BAO hebben gekregen. Dit wordt ook wel de *False Positive Rate* genoemd.

Op basis van deze bias definitie wordt vervolgens deze definitie voor fairness (ook wel 'behoorlijkheid' genoemd) gehanteerd: de waarschijnlijkheid dat een bonafide visumaanvraag desondanks een intensive track aanduiding krijgt mag niet

[1] Zie Privacy Management Partners (2022).

(wezenlijk) afhankelijk mag zijn van de nationaliteit, geslacht of de leeftijd van de aanvrager.

3. **Is er sprake van bias in het opzetten en gebruik van profielen (paragraaf 4.24.4)?**

Op basis van de onderzoeksresultaten kan worden geconcludeerd dat er sprake is van (i) aanzienlijke bias-discrepanties op basis van nationaliteit, (ii) beperkte bias-discrepanties op basis van geslacht, en (iii) geen substantiële bias-discrepanties op basis van leeftijd.

4. **Wat zijn de voornaamste conclusies ten aanzien van geconstateerde bias (paragraaf 4.4)?**

Wat betreft de mate waarin de aangetroffen bias-discrepanties mogelijk van invloed zijn op de uiteindelijke besluitvorming over visumaanvragen kunnen op basis van de beschikbare data geen conclusies worden getrokken (dit zou op basis van toekomstige dataverzameling nader onderzocht kunnen worden).

Op basis van literatuur over de effecten van bias in risicoprofilering kan wel worden gesteld dat er een aanzienlijk risico bestaat dat de aangetroffen bias-discrepanties (die met name op basis van nationaliteit aanzienlijk zijn) de besluitvorming over visumaanvragen in enige mate beïnvloedt.

5. **Welke stappen kunnen genomen worden om geconstateerde bias te verminderen? (hoofdstuk 5)**

Dit rapport geeft een aantal aanbevelingen voor zowel de korte als de middellange termijn.

De kortetermijnmaatregelen zijn bedoeld als tijdelijke oplossingen om de geconstateerde bias in de BAO aanzienlijk terug te brengen. Op volgorde van meest tot minst effectief zijn dit (i) het beëindigen van het gebruik van de profielscore in de BAO, (ii) het opheffen van *intensive track* als uitkomst bij afwezigheid van 'hits' op aanvrager/referent/werkgever, (iii) het beëindigen van het gebruik van risico-groepen (profielen uitsluitend op basis van weigeringspercentage), of (iv) het beëindigen van het gebruik van het gegeven *nationaliteit* als variabele in de BAO.

De voorgestelde maatregelen voor de middellange termijn beogen een duurzame reductie van bias in zowel de BAO als het overkoepelende IOB/KVV proces. De set van complementaire maatregelen bestaat uit (i) een evaluatie en herziening van de BAO aan de hand van het Impact Assessment Mensenrechten en Algoritmen (IAMA), (ii) verbetering van de inrichting en beheer van de BAO, (iii) toepassing van technische biasmitigatie methoden, (iv) onderzoek naar de mogelijkheden van aanvullende databronnen ten behoeve van de profielscore, (v) toepassing van niet-

technische biasmitigatie methoden, en (vi) inrichting van monitoring van bias in de besluitvorming.

1 Inleiding

Het Ministerie van Buitenlandse Zaken (BZ) heeft het Rijks ICT Gilde (RIG) verzocht om een biastoetsing van de Buitenlandse Zaken Analyse Omgeving (BAO) tool. De BAO wordt gebruikt ter ondersteuning van de behandeling van visumaanvragen als onderdeel van het proces 'Informatie-Ondersteunend Beslissen: Kort Verblijf Visa' (IOB/KVV).

1.1 Aanleiding

Naar aanleiding van een intern DPIA traject ten aanzien van het IOB/KVV heeft BZ adviesbureau 'Privacy Management Partners' (PMP) verzocht om een analyse van de toelaatbaarheid van het gebruik van het gegeven *nationaliteit* in de BAO. In haar rapport van 24 mei 2022 concludeerde PMP onder meer dat het IOB/KVV-proces behoorlijk (en daarmee toelaatbaar) is in de zin van de AVG "mits er geen bias blijkt te zitten in het [BAO] algoritme".

Naar aanleiding van deze deelconclusie heeft BZ het RIG om een analyse gevraagd van mogelijke bias in de BAO, aan de hand van de volgende vragen:

1. Welk type bias kunnen zich voordoen binnen het IOB/KVV proces (en daarbinnen de BAO)?
2. Welke definitie van bias wordt gehanteerd voor dit onderzoek?
3. Is er sprake van bias in het opzetten en gebruik van profielen?
4. Wat zijn de voornaamste conclusies ten aanzien van geconstateerde bias?
5. Welke stappen moeten/kunnen genomen worden om geconstateerde bias te verminderen?

Feedback van BZ op een eerste conceptversie van dit rapport heeft geresulteerd in enkele verduidelijkingen in de tekst van dit rapport. Voor een reactie op de individuele opmerkingen in deze feedback verwijzen wij naar bijlage 7.

1.2 Context

1.2.1 Informatie Ondersteunend Beslissen – Kort Verblijf Visa (KVV)

BZ is belast met het verstrekken van Schengen Kort Verblijf Visa (KVV). BZ vervult hiermee de taak van toelating voor de migratieketen. Om deze taak zo goed mogelijk uit te voeren wordt er gebruik gemaakt van een op data-analyse gebaseerde werkwijze, het Informatie Ondersteund Beslissen (IOB). Het doel van deze werkwijze is beslismedewerkers van informatie voorzien die ze in staat stelt

een goede en objectievere beslissing te nemen op visumaanvraagdossiers.[2] De kwaliteit van KVV beslissingen is belangrijk om Nederlandse economische belangen maximaal te faciliteren en misbruik van de visumprocedure te beperken. De informatie wordt beschikbaar gesteld aan beslismedewerkers middels een technische applicatie (de BAO).

1.2.2 Buitenlandse Zaken Analyse Omgeving (BAO)

Het specifieke doel van de gegevensverwerking middels de BAO is om informatie ondersteund te beslissen op aanvragen voor een KVV. De BAO geeft hier invulling aan door een inschatting te presenteren over het risicoprofiel van een aanvraag en de gewenste intensiteit van de behandeling hiervan. De BAO wordt op drie manieren ingezet:

1. Controle of er informatie binnen het Nieuw Visum Informatie Systeem (NVIS) beschikbaar is over de aanvrager, referent, werkgever en/of reisdocument (zogenoemde 'hits');
2. Controle of er informatie binnen bronnen van de migratie ketenpartners beschikbaar is over de aanvrager en/of referent (zie Tabel 1);
3. Het vergelijken van de visumaanvraag met de op dat moment in de BAO aanwezige profielen en lokale informatie.

Ten behoeve van dit laatste punt maakt de BAO gebruik van zogeheten *profielen* (ook wel patronen in de vorm van kansen, risico's en trends genoemd). Dit is informatie die opgesteld is op basis van geaggregeerde gegevens die BZ zelf tot haar beschikking heeft en de gegevens uit de bronnen van ketenpartners (zie punt 1 en 2). Voor het opstellen van en toetsen aan de profielen wordt onder meer gebruik gemaakt van de *nationaliteit*, het *geslacht* en de *leeftijd* van de aanvrager. De profielen zelf worden periodiek herzien en gecreëerd op basis van historische hit-percentages (zie punt 1 en 2) en weigeringspercentages. Er is sprake van een zogenaamde hit als er al positieve of negatieve informatie bekend is gerelateerd aan een individuele aanvraag.

Tabel 1. Beschrijving van verschillende ketenpartnerbronnen en de informatie die deze bevatten over de aanvrager

Ketenpartner/bron	Informatie over aanvrager
Koninklijke Marechaussee (KMar)	Weigeringen aan de grens
	Verblijfstermijn verstreken
	Doorlating onder voorwaarden

[2] Zo wordt in het Data Protection Impact Assessment voor de BAO (Ministerie van Buitenlandse Zaken, 2021) onder meer gesteld dat "[de BAO] biedt de beslismedewerkers een mogelijkheid om sneller risico's in kaart te kunnen brengen, en beter onderbouwd een beslissing te kunnen maken. Dit zorgt ook voor objectievere besluitvorming, omdat de beschikbare informatie op objectieve en uniforme wijze is gewogen en wordt teruggekoppeld" en "de beslismedewerker neemt geen beslissingen op grond van de BAO, maar weegt informatie die uit IOB instrumenten zoals de BAO [...] komt mee in het te nemen besluit".

Immigratie- en Naturalisatiedienst (IND)	Asielaanvraag
	Machtiging tot voorlopig verblijf geweigerd
Dienst Terugkeer en Vertrek (DT&V)	Begeleid vertrek uit Nederland
Inspectie Sociale Zaken en Werkgelegenheid (ISZW)	Strafbare feiten zoals arbeidsmarktfraude, arbeidsuitbuiting en uitkeringsfraude
Nationale Politie, Afdeling Vreemdelingen Identificatie en Mensenhandel & -smokkel (AVIM)	Overige strafbare feiten
Attenderingen	Negatieve informatie over aanvrager of referent (Aanvrager Intensive Track AIT, en Referent Intensive Track RIT)
	Positieve informatie over aanvrager of referent (Aanvrager Fast Track AFT, en Referent Fast Track RFT)

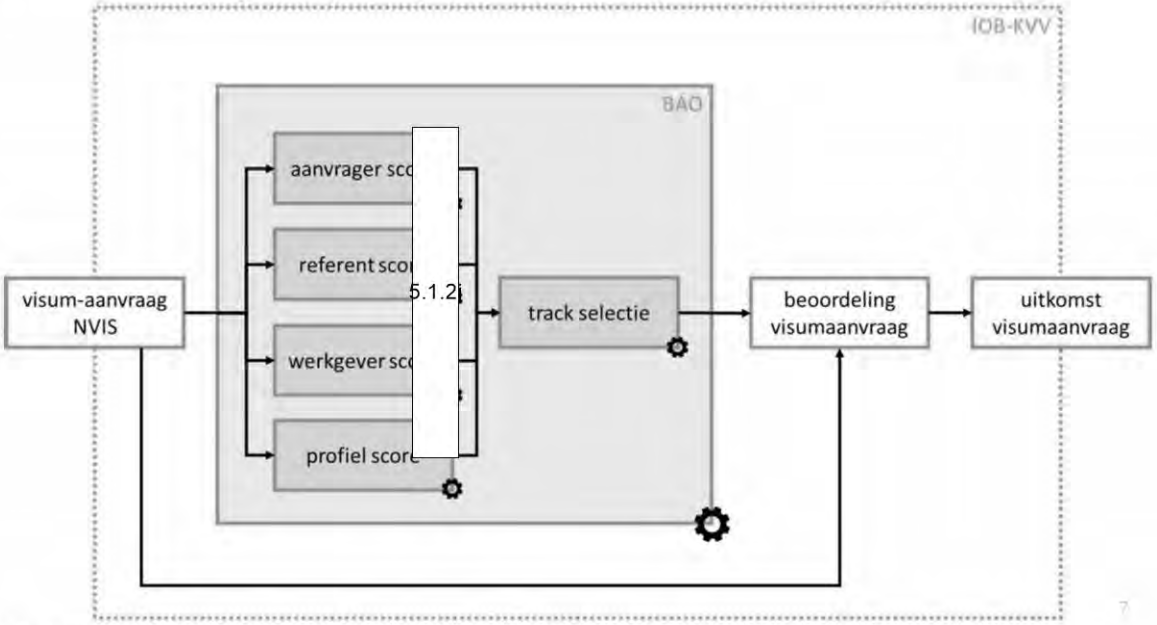
Op basis van eventuele hits in de onder punt 1 en 2 genoemde informatiebronnen en/of de onder punt 3 genoemde profielen presenteert de BAO een risicoprofilering in de vorm van een trackselectie, te weten *fast*, *regular*, en *intensive track* (zie Figuur 1 voor een versimpelde weergave van de BAO binnen het IOB/KVV proces).

Fast track
Aanvraag valt in groep met weinig tot geen risico op misbruik visumprocedure. Dit betekent dat er alleen positieve informatie bekend is in relatie tot de aanvraag. De aanvraag kan niet blindelings worden goedgekeurd maar kan wel sneller worden afgehandeld. Bij een kansprofiel is de motivatie die de beslismedewerker te zien krijgt als volgt: *'Aanvraag valt in groep met weinig tot geen risico op misbruik visumprocedure'*.

Regular track
Aanvraag waarop op basis van de beschikbare data geen risico-inschatting mogelijk is. Dit betekent dat er geen informatie bekend is in relatie tot de aanvraag. De aanvraag kan via normale processtappen worden afgehandeld.

Intensive track
Aanvraag valt in groep met verhoogd risico op misbruik visumprocedure. Dit betekent dat er negatieve informatie bekend is in relatie tot de aanvraag en het dossier dus extra aandacht verdient. Bij een risicoprofiel is de motivatie die de beslismedewerker te zien krijgt als volgt: *'LET OP: Aanvraag valt in groep met verhoogd risico op misbruik visumprocedure'*.

De BAO track selectie is in zowel opzet, implementatie, beschrijving als communicatie naar de beslismedewerkers primair en expliciet een risico-profileringsysteem.[3]



Figuur 1. Versimpelde weergave van het IOB/KVV

1.2.3 Het onderliggende BAO beslisboomalgoritme

Het algoritme binnen de BAO dat de profielen genereert is een zogenaamd beslisboomalgoritme dat werkt op basis van beslisregels die door de bouwers van het algoritme zijn opgesteld en verwerkt in het algoritme.

Het beslisboomalgoritme verwerkt alle aanvragen van de afgelopen vijf jaar en deelt deze op in groepen op basis van de volgende zeven kenmerken in deze vaste volgorde:

[3] Zo wordt er in de BAO documentatie (Richtlijnen Informatie Ondersteunend Beslissen, FAQ Informatie Ondersteunend Beslissen, DPIA BAO vastgesteld 25 juni 2021) herhaaldelijk gesproken over een positief (fast track) of negatief (intensive track) advies vanuit de BAO. Daarnaast wordt meermaals gesproken over de fast en intensive track als "kansprofiel en risicoprofiel", "meer of minder betrouwbare aanvragen", "positieve en negatieve aanvragers", "positieve en negatieve referenten", "laag of verhoogd risico", "weinig of geen risico" / "verhoogd risico", en "alleen positieve informatie bekend in relatie tot de aanvraag" / "negatieve informatie bekend is in relatie tot de aanvraag".

1. Verblijfsdoel
2. Post (plaats van aanvraag)
3. Nationaliteit
4. Geslacht
5. Leeftijdsklasse
6. Burgerlijke staat
7. Beroep

Op basis van bovenstaande kenmerken maakt het algoritme subgroepen met minimaal 200 aanvragen. Deze subgroepen zijn gemaakt op basis van minimaal drie kenmerken (verblijfsdoel, post en leeftijd boven de 18 jaar) die aan specifieke regels voor hit- en weigeringspercentage voldoen. Op basis van deze subgroepen zullen de uiteindelijke profielen worden opgesteld, waarbij moet worden voldaan aan de specifieke regels voor een gegeven profiel (zie Tabel 2 voor een volledig overzicht van de regels onderliggend aan de profielen).

Tabel 2. Regels op basis waarvan het beslisboomalgoritme de verschillende typen profielen genereert

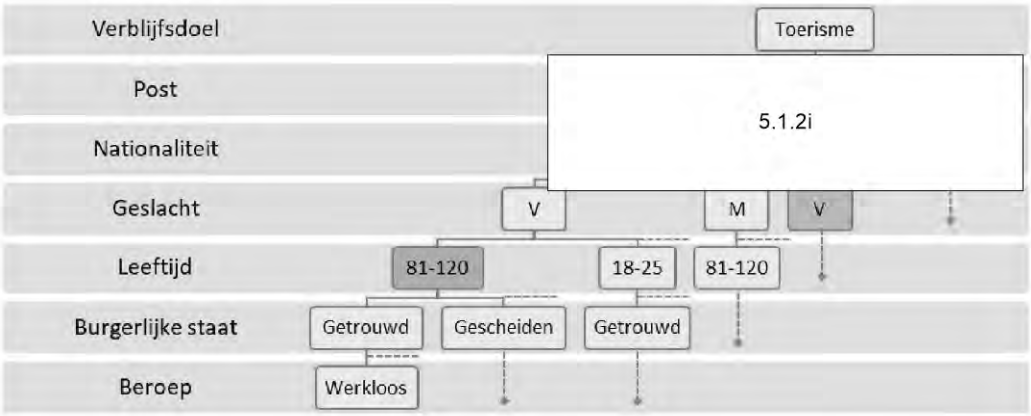
Type profiel		Hitpercentage	Weigeringspercentage
Risico	Type 1	Minimaal 5%	-
	Type 2	Tussen de 1% en 5%	Minimaal 16%
Kans	Type 1	Maximaal 0.25%	Maximaal 5%
	Type 2	Tussen 0.25% en 0.5%	Maximaal 2.5%
Trend	Type 1	Minimaal 5%	-
	Type 2	Tussen de 1% en 5%	Minimaal 16%
Groep	Fast	-	Lager dan gemiddeld
	Intensive	-	Hoger dan gemiddeld

Een voorbeeld van een dergelijke subgroep gevonden door het algoritme kan bijvoorbeeld zijn: 212 aanvragen met verblijfsdoel *toerisme,* post 5.1.2i nationaliteit 5.1.2i geslacht *vrouwelijke* en leeftijdsklasse *81-120* met een gemiddeld hitpercentage van 0.3% en een weigeringspercentage van 2%. Deze subgroep voldoet aan de regels voor een kansprofiel type 2 en dus worden bovenstaande kenmerken vervolgens geregistreerd als een kansprofiel type 2 (Figuur 2). Een andere subgroep kan zijn: 460 aanvragen met verblijfsdoel *toerisme,* post 5.1.2i nationaliteit 5.1.2i met een hitpercentage van 6%. Deze subgroep voldoet aan de regels voor een risicoprofiel type 1. Bij het identificeren van deze subgroepen worden nooit onnodig extra kenmerken toegevoegd, ook al wordt nog steeds aan de regels voldaan. Dit wordt ook wel het 'snoeien' van de beslisboom genoemd.

Er zijn vier verschillende typen profielen: risico-, kans- en trendprofielen en een vierde categorie van de zogenaamde groepsprofielen. Risico- en kansprofielen zijn gebaseerd op een combinatie van hit- en weigeringspercentage. Deze informatie is meestal afkomstig van migratie ketenpartners (zie Tabel 1). Trendprofielen zijn

hetzelfde als risicoprofielen, maar gebaseerd op basis van data van de afgelopen zes maanden om de korte termijn risicotrends op te vangen. Groepsprofielen worden gegenereerd nadat er geen subgroepen geïdentificeerd kunnen worden die voldoen aan de regels voor risico-, kans- en trendprofielen. De overgebleven subgroepen (de gele blokjes in Figuur 2) worden verdeeld in *fast* en *intensive track* op basis van het weigeringspercentage van die subgroep: is deze lager dan het gemiddelde weigeringspercentage van de afgelopen 5 jaar, dan wordt dit een *fast* groepsprofiel; is deze hoger, dan wordt dit een *intensive* groepsprofiel.

Op ieder gegeven moment zijn er circa 1.000 profielen actief. De resulterende profielen worden vervolgens geladen in de BAO. Per nieuwe aanvraag die in NVIS binnenkomt wordt vervolgens gecheckt of deze aanvraag een match heeft op een profiel. Profielen draaien voor 1 maand mee in de BAO, waarna een geheel nieuwe set aan profielen wordt gegenereerd voor de komende maand.



Figuur 2. Gedeelte van een mogelijke beslisboom voor verblijfsdoel *toerisme* gegenereerd door het beslisboomalgoritme onderliggend aan de BAO

2 Theoretisch kader bias-toetsing

2.1 Maatschappelijke context

Potentiële bias in algoritmische besluitvorming en -ondersteuning is geen opzichzelfstaand vraagstuk, maar maakt onderdeel uit van het bredere vraagstuk van verantwoorde en betrouwbare inzet van algoritmen, en de ethische context waarbinnen dergelijke systemen gebruik worden.

Mede door enkele spraakmakende zaken in de laatste jaren is de maatschappelijke en politieke aandacht voor het thema bias in algoritmes aanzienlijk toegenomen, zowel in Nederland als internationaal. In Nederland staat de zogenoemde 'Toeslagenaffaire' nog vers op het netvlies. In de VS zorgde in 2016 het onderzoek van *ProPublica* naar het COMPAS recidive algoritme voor veel ophef,[4] en in het Verenigd Koninkrijk trok de uitvoeringsorganisatie *Home Office* in 2020 de stekker uit een algoritme ter ondersteuning van de beoordeling van visa aanvragen (in opzet grotendeels vergelijkbaar met het BAO algoritme) nadat de *Joint Council for the Welfare of Immigrants* (JCWI) en burgerrechtenorganisatie *Foxglove* een rechtszaak aanspanden.[5]

In het Coalitieakkoord van 2021 is eveneens aandacht voor verantwoorde en betrouwbare AI, inclusief de aankondiging van een wettelijke regeling voor controles van algoritmes op onder meer discriminatie en willekeur. Daarnaast besteedt het Kabinet aandacht aan bias-problematiek in diverse voortgangsrapportages over de uitvoering van "Motie 21" (in het bijzonder ten aanzien van het gebruik van afkomstgerelateerde indicatoren in risico-modellen).[6] Verder heeft aandacht voor mogelijke bias een centrale rol in het in ontwikkeling zijnde implementatiekader voor de inzet van algoritmes en instrumenten als het Impact Assessment Mensenrechten en Algoritmes (IAMA) en de handreiking Non-discriminatie by design.[7] Tot slot worden in de op handen zijnde AI verordening eisen gesteld aan de beheersing van bias, met name voor "hoog-risico" toepassingen als de BAO.[8]

2.2 Definities

Het gebruik van terminologie rondom bias is vaak niet eenduidig en inconsistent, en termen als bias, discriminatie en fairness worden regelmatig door elkaar gebruikt. In dit rapport worden de volgende definities gebruikt voor deze begrippen:

[4] Zie onder meer Angwin et al. (2016).
[5] Zie onder meer BBC (2020).
[6] Motie van het lid Klaver c.s., 19 januari 2021.
[7] Zie Ministerie van Binnenlandse Zaken (2021a, 2021b).
[8] Europese Commissie (2021).

Bias
Systematische afwijking tussen modeluitkomsten en de werkelijke waarden (voor een bepaalde groep). Wanneer bijvoorbeeld sprake is van een stelselmatige onderschatting (danwel overschatting) van het risicoprofiel voor een bepaalde groep, dan bevat het model een vooringenomenheid ten gunste (danwel ten nadele) van deze groep.

Discriminatie
Verschillen in behandeling van bepaalde personen in vergelijkbare gevallen (direct of indirect op basis van beschermde persoonskenmerken).

Fairness ('behoorlijkheid' in termen van de AVG)
Gewenste type en mate van gelijke behandeling van (of modeluitkomsten voor) individuen of groepen, op basis van relevant geachte kenmerken.

Hoewel de bovengenoemde definities dichtbij elkaar liggen zijn er tevens belangrijke verschillen. Zo is het mogelijk dat een bepaald (slecht gekalibreerd) model een bepaalde bias heeft die voor alle onderscheiden groepen ongeveer even slecht uitpakt. In dat geval is er niet noodzakelijk sprake van discriminatie of oneerlijke behandeling. Daarnaast kan er sprake zijn van enige vorm van directe of indirecte discriminatie, zonder dat dit in strijd hoeft te zijn met het gewenste fairness principe. Tot slot kan een model uitmonden in een oneerlijke behandeling van personen of groepen zonder dat hier per sé sprake is van enige bias of discriminatie.

2.3 Bronnen en verschijningsvormen van bias

Bias is een veelkoppig monster. In door algoritmes ondersteunde beslisprocessen kunnen diverse bronnen van bias in alle processtappen een rol spelen. Er zijn veel verschillende (maar op hoofdlijnen vergelijkbare) typologieën van bias. Ten aanzien van door algoritmes ondersteunende beslisprocessen als het IOB/KVV kunnen mogelijke bronnen van bias grofweg in vier categorieën worden ingedeeld:

1. Bias in data
2. Bias in het modelleren
3. Bias als gevolg van ingebruikname van een systeem
4. Bias in het dagelijks gebruik door medewerkers

Bij het trainen van algoritmes (dan wel het bepalen van beslisregels) kan de gebruikte data een bron van veel bias zijn. Met name wanneer het data van externe bronnen, verouderde data, incomplete data, of data verzameld voor een ander doel betreft is het risico op bias groot. Tijdens de ontwikkeling van een algoritme worden veel impliciete en expliciete keuzes gemaakt die kunnen resulteren in bias. Denk hierbij aan de keuze voor de te gebruiken methodiek, voorbewerking van de data, gekozen waarden voor modelparameters, selectie van evaluatiecriteria, et cetera. Wanneer een systeem eenmaal in productie wordt genomen, kan het systeem zelf

onbedoeld een bron worden van bias. Tot slot kan bias optreden in het gebruik van het systeem door medewerkers, met name als gevolg van een onjuiste interpretatie en/of gebruik van modeluitkomsten.

In onderstaande tabel worden de verschillende bronnen van bias gepresenteerd. In Bijlage 2 is een nadere toelichting op de genoemde typen bias te vinden.

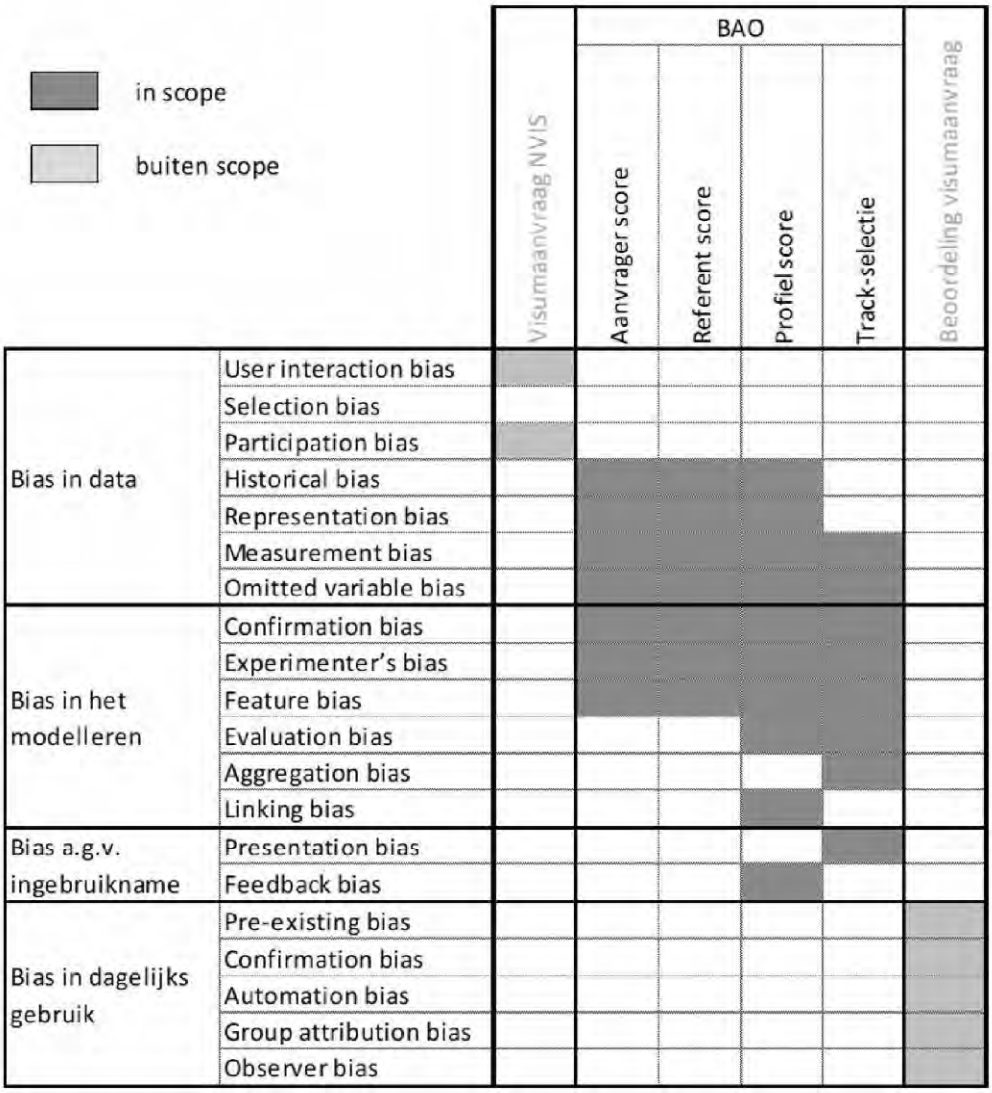
Tabel 3. Mogelijke bronnen van bias

Bias in data	**Bias in het modelleren**	**Bias a.g.v. ingebruikname**	**Bias in dagelijks gebruik**
User interaction bias	Confirmation bias	Presentation bias	Pre-existing bias
Selection bias	Experimenter's bias	Feedback bias	Confirmation bias
Participation bias	Feature bias		Automation bias
Historical bias	Evaluation bias		Group attribution bias
Representation bias	Aggregation bias		Observer bias
Measurement bias	Linking bias		
Omitted variable bias			

Het kan inzichtelijk zijn om een inventarisatie te maken van mogelijke bronnen van bias in het te toetsen systeem. Figuur 3 geeft een overzicht van hoe verschillende typen bias zich zouden kunnen manifesteren in bepaalde onderdelen van het IOB/KVV.

Wanneer bias wordt geconstateerd in een systeem is het echter vaak niet makkelijk om te achterhalen welke type bias (waar in het systeem) precies hiervoor verantwoordelijk is. Bij het toetsen van bias in systeemuitkomsten is het gemeten resultaat vrijwel altijd een netto resultante van alle mogelijke bronnen van bias in het gehele systeem. Mogelijk neutraliseren bepaalde vormen van bias elkaar gedeeltelijk, maar vaak zullen verschillende bronnen elkaar juist versterken. Het constateren van bias is vaak al een complexe onderneming. Het verkrijgen van inzicht in de precieze oorzaken van de geconstateerde bias vergt helaas vrijwel altijd diepgravende vervolganalyses.

Daarnaast is het belangrijk om te beseffen dat het nooit mogelijk zal zijn om bias geheel uit te bannen. Ieder model is een versimpelde benadering van de werkelijkheid, en zal om die reden altijd een bepaalde mate van systematische fouten vertonen (waarbij de eerdergenoemde potentiële bronnen van bias mede de omvang en richting van deze systematische fout bepalen). Om die reden is het beter te spreken over het (zoveel als mogelijk) reduceren van bias in plaats van het volledig wegnemen van bias. Tot slot, het is een misvatting te denken dat simpelere modellen (bijvoorbeeld gebaseerd op eenvoudige beslisregels) een goede manier zijn om bias te verminderen, aangezien modellen die een te simplistische benadering van de werkelijkheid zijn in de regel resulteren in hogere bias (een fenomeen dat bekend staat als '*underfitting*').



Figuur 3. Mogelijke bronnen van bias in het IOB/KVV

2.4 Definitie en meetbaarheid van fairness

Fairness, bezien als het gewenste type en mate van gelijke behandeling van (of modeluitkomsten voor) individuen of groepen op basis van relevant geachte kenmerken, dient per situatie te worden gedefinieerd en geoperationaliseerd. Hiervoor dient te worden gespecificeerd welk type van gelijke behandeling het meest relevant is voor het vraagstuk, welke mate van ongelijkheid in behandeling als acceptabel wordt beschouwd, en op basis van welke relevante (groeps)kenmerken geen onderscheid zou mogen worden gemaakt.

2.4.1 Type fairness

In de literatuur over algoritmes is inmiddels sprake van een rijk scala aan mogelijke fairness-definities, waarbij de focus aan de ene kant van het spectrum ligt op *individual fairness* (of *process fairness*) en aan de andere kant op *group fairness* (of

outcome fairness).[9] Andere, en meer verfijnde categorisaties zijn eveneens mogelijk, zoals een driedeling tussen fairness-definities die voldoen aan *independence*, *sufficiency*, of *separation* criteria.[10]

De meest relevante definitie van fairness is uiteindelijk afhankelijk van de aard van het vraagstuk, en volgt uit een precisering van wat als een (on)eerlijk uitkomst wordt beschouwd.[11]

Wanneer eenmaal duidelijk is wat als een (on)eerlijke uitkomst dient te worden beschouwd, kan met behulp van een aantal vragen de van toepassing zijnde fairness-definitie (en bijbehorende metriek) worden bepaald, bijvoorbeeld:

- Dient er sprake te zijn van een gelijke representatie van groepen in de uitkomsten, of een gelijke behandeling in het proces?
- Is gelijke representatie gewenst in termen van absolute getallen of naar rato van groepsgrootte?
- Is de uitkomst voor betrokkenen van positieve ('*assistive*') of negatieve ('*punitive*') aard?
- Welke subgroep van betrokkenen is het meest relevant: de groep die een bepaalde uitkomst heeft gekregen, of de groep die de uitkomst had moeten krijgen?

Een veelgebruikt hulpmiddel om tot de juiste fairness-definitie te komen is een zogenoemd fairness kompas (zie Bijlage 3 voor een voorbeeld van Aequitas).

2.4.2 Acceptabele mate van ongelijkheid

Het is, onafhankelijk van de gekozen fairness-definitie, praktisch niet mogelijk om in alle situaties een perfecte gelijkheid te creëren. Enerzijds is dit het gevolg van het feit dat de historische data waarop een model is getraind nooit perfect representatief is voor toekomstige gevallen. Anderzijds zal het bevorderen van fairness (op basis van de gekozen definitie) in veel gevallen ten koste gaan van andere aspecten van het model (bijvoorbeeld de nauwkeurigheid). Afhankelijk van de weging van de verschillende belangen die met het gebruik van bepaald model gepaard gaan, kan een bepaalde mate van ongelijkheid als acceptabel worden beschouwd. In andere woorden, de gekozen acceptabele mate van ongelijkheid resulteert in een bepaalde bandbreedte waarbinnen de uitkomsten als acceptabel mogen worden beschouwd.

---

[9] Zie bijvoorbeeld Narayan (2018).
[10] Zie bijvoorbeeld Ruf & Detyniecki (2021).
[11] Een belangrijk gegeven hierbij is dat het realiseren van gelijkheid op basis van een bepaald type fairness (in geval van groepen met verschillende prevalentie van eigenschappen) noodzakelijk tot gevolg heeft dat er sprake zal zijn van (enige mate van) ongelijkheid op basis van alle andere typen (het zogenoemde '*Impossibility Theorem*'). Zie bijvoorbeeld Kleinberg et al. (2016).

Een veel gebruikte vuistregel is de zogenoemde "80 procent regel".[12] Deze regel stelt dat een model oneerlijk is wanneer de negatieve impact voor een bepaalde groep meer dan 20 procent (positief of negatief) afwijkt ten opzichte van de impact op de bredere bevolking (of referentiegroep).

2.4.3 Relevante (groeps)kenmerken

Naast het gewenste type en mate van gelijke behandeling, is afbakening van de relevante (groeps)kenmerken waarvoor het model dient te voldoen aan de gekozen fairness-definitie van belang.[13] Afhankelijk van de context kan dit bijvoorbeeld gaan over beschermde persoonskenmerken als etniciteit, geslacht of leeftijd, en/of andere voor het vraagstuk relevante kenmerken zoals inkomen of opleidingsniveau.

[12] Deze regel wordt door de US Equal Employment Opportunity Commission (1979) als volgt verwoord: "*The agencies have adopted a rule of thumb under which they will generally consider a selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5ths) or eighty percent (80%) of the selection rate for the group with the highest selection rate as a substantially different rate of selection.*"
[13] Zie bijvoorbeeld Muhammad (2022) voor een duidelijke toelichting.

3 Werkwijze onderzoek

De bias toetsing is tot stand gekomen op basis van een analyse van de door BZ beschikbaar gestelde data. Voor de interpretatie van de data en een beter begrip van het functioneren van de BAO is met verschillende medewerkers van BZ gesproken.[14]

3.1 Context voor bias-toetsing: IOB/KVV

Hoewel deze bias-toetsing is gericht op het BAO algoritme is het voor het bepalen van de juiste fairness-definitie belangrijk dit te bezien in de context van het volledige IOB/KVV proces.

3.1.1 Fairness-definitie IOB/KVV

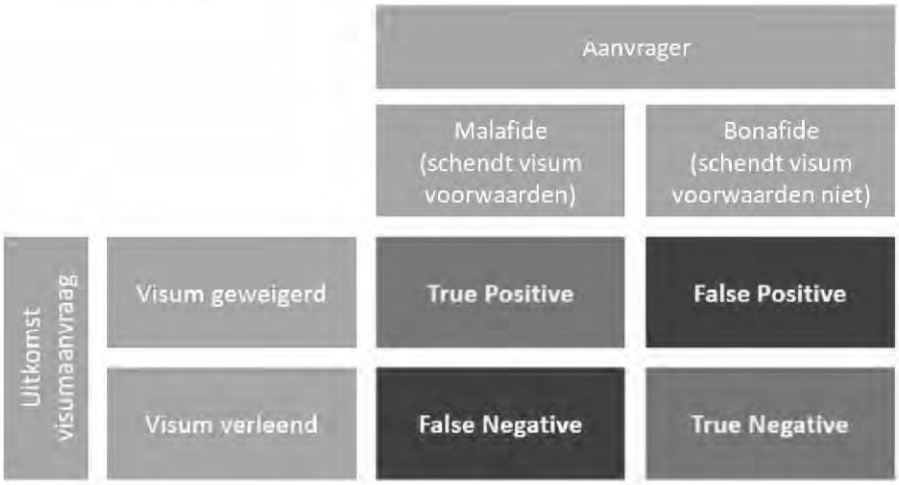
Om op een zinvolle wijze over bias en fairness te spreken dient er sprake te zijn van een belang dat wordt geschaad, oftewel: een onterecht nadeel (of onterecht onthouden voordeel). In de context van het IOB/KVV is het niet per se problematisch wanneer de afwijzingspercentages verschillen tussen groepen. Immers, wanneer in bepaalde groepen relatief vaker malafide aanvragers voorkomen is een hoger weigeringspercentage niet onterecht. Wat dient in de context van IOB/KVV dan wel als een onterecht nadeel te worden beschouwd?

"Waar mensen werken worden fouten gemaakt". Dit is eveneens het geval bij de besluitvorming over visumaanvragen. Het risico is aanwezig dat een aanvraag van een bonafide aanvrager per abuis wordt geweigerd of dat een malafide aanvraag er onbedoeld doorheen glipt. Wanneer we kijken naar het IOB/KVV zijn er conceptueel vier uitkomsten mogelijk:

[14] Tot op het moment van het schrijven van dit rapport heeft BZ geen toestemming verleend voor interviews met beslismedewerkers, zijnde de belangrijkste gebruikers van de BAO. Hierdoor is het niet mogelijk geweest om nadere duiding met betrekking tot het gebruik van track-uitkomsten in de dagelijkse praktijk te krijgen.

1. Een visumaanvraag van een bonafide aanvrager[15] wordt (terecht) goedgekeurd;
2. Een visumaanvraag van een bonafide aanvrager wordt (onterecht) afgewezen;
3. Een visumaanvraag van een malafide aanvrager wordt (onterecht) goedgekeurd;
4. Een visumaanvraag van een malafide aanvrager wordt (terecht) afgewezen.

Deze uitkomsten kunnen schematisch als volgt worden weergegeven (in een zogenoemde '*confusion matrix*'), waarbij de bovengenoemde uitkomsten worden getypeerd als, respectievelijk, (1) *True Negative*, (2) *False Positive*, (3) *False Negative*, en (4) *True Positive* (Figuur 4).



Figuur 4. Schematische weergave uitkomsten IOB/KVV (conceptueel)

Als uitgangspunt voor het IOB/KVV wordt gesteld dat er sprake is van een onterecht nadeel wanneer de visumaanvraag van een bonafide aanvrager wordt afgewezen (*False Positive*). De bijbehorende fairness-definitie is dat de waarschijnlijkheid dat visumaanvragen van bonafide aanvragers worden afgekeurd niet afhankelijk mag zijn van beschermde persoonskenmerken (zoals nationaliteit, leeftijd of geslacht).

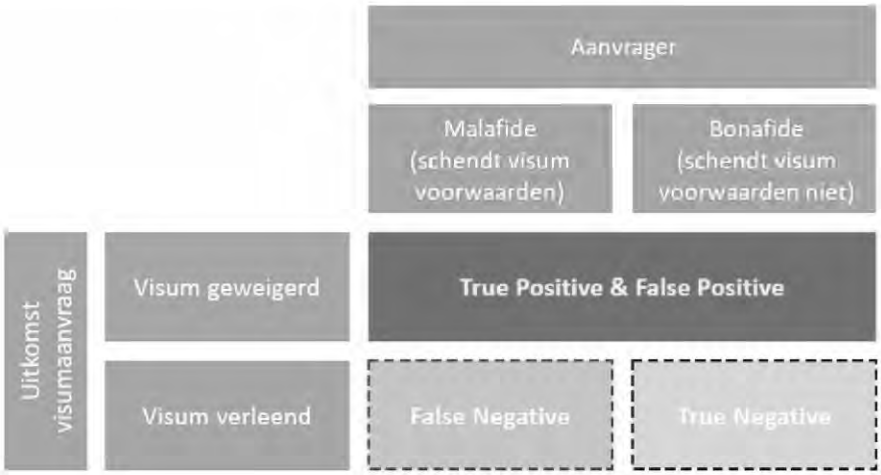
3.1.2 Onuitvoerbaarheid van bias-toetsing op niveau IOB/KVV

Door de aard van het visumverstrekking proces is het per definitie niet mogelijk om te bepalen in welke mate visumaanvragen van bonafide aanvragers worden afgewezen. Immers, van personen waarbij de visumaanvraag is geweigerd kan nimmer worden vastgesteld of zij zich bij verblijf aan de visumvoorwaarden gehouden zouden hebben.

[15] Met 'bonafide aanvragers' bedoelen we personen die aan alle visumcriteria voldoen, zich bij verblijf aan alle visumvoorwaarden (zouden) houden, en welbeschouwd voor een KVV in aanmerking zouden moeten komen.

Andersom is het niet met precisie vast te stellen welke aanvragers bij verblijf de visumvoorwaarden al dan niet schenden (aangezien niet alle overtredingen in beeld zijn bij de uitvoeringsorganisaties).

De eerder gepresenteerde schematische weergave ziet er vanuit een informatie-oogpunt dan dus eigenlijk uit als in Figuur 5. Er is geen onderscheid mogelijk tussen *True Positives* en *False Positives* (geel gebied). Daarnaast is er geen volledige informatie over *False Negatives* en *True Negatives* (onderbroken kader).



Figuur 5. Schematische weergave uitkomsten IOB/KVV (praktisch)

3.2 Bias-toetsing BAO

Hoewel het op basis van de beschikbare data zoals gezegd onmogelijk is om potentiële bias in de besluitvorming over visumaanvragen vast te stellen, kan wel worden geëvalueerd in welke mate er sprake is van bias in de risicoprofilering in de BAO (ten opzichte van de uiteindelijke besluitvorming over aanvragen). Indien er bovenmatige bias in deze risicoprofilering optreedt, bestaat het risico dat deze bias in enige mate doorwerkt in de besluitvorming en op die manier discriminatie in de hand werkt.[16]

De beoordeling van visumaanvragen geschiedt aan de hand van zowel harde (objectieve) criteria zoals de aanwezigheid van de vereiste documentatie, als zachte (subjectieve) criteria zoals de geloofwaardigheid en betrouwbaarheid van de verklaringen van de aanvrager. Wanneer de BAO een vertekend beeld presenteert over het risicoprofiel van bepaalde groepen aanvragers is het mogelijk dat dit van

[16] In het rapport 'Aandacht voor Algoritmes' (2021) noemt de Algemene Rekenkamer de BAO als voorbeeld van toepassingen waarbij risicoprofilering door een algoritme de uiteindelijke beslissing van de medewerker kan beïnvloeden. Dit risico wordt eveneens benoemd in andere beleids- en toezichthouderpublicaties (zie onder meer het Ministerie van Binnenlandse Zaken (2022a, 2022b, 2021a, 2021b), College voor de Rechten van de Mens (2022), Raad van State (2021), en Autoriteit Persoonsgegevens (2020)).

enige invloed kan zijn op de besluitvorming over de visumaanvragen van deze groep.

BZ geeft aan dat de BAO track selectie enkel iets zegt over de intensiteit rondom het behandelen van het dossier door de beslismedewerker. Uit de gegevens van BZ rondom de tijdsbesteding aan visumaanvragen blijkt echter dat aan aanvragen met een *intensive track* aanduiding gemiddeld ongeveer evenveel of zelfs minder tijd wordt besteedt dan aan aanvragen met een *regular track* aanduiding. Dit suggereert dat de track aanduiding in de praktijk niet zozeer bepalend is voor de *intensiteit* van de behandeling van de aanvraag, maar eerder voor de *rigiditeit* (en mogelijk vooringenomenheid) waarmee de aanvraag wordt beoordeeld.[17]

In welke mate (bias in) de risicoprofilering door de BAO daadwerkelijk van invloed is op de uiteindelijke besluitvorming door de beslismedewerker, bijvoorbeeld als gevolg van *automation bias* of *confirmation bias*, kan op basis van de beschikbare data helaas niet worden vastgesteld.[18] Recente onderzoeken laten echter zien dat de impact van (onjuiste) risicoprofilering op besluitvorming zelfs bij ervaren professionals, zoals rechters, aanzienlijk is.[19]

3.2.1 Fairness-definitie BAO

Zoals gesteld in paragraaf 3.1.1 dient, om op een zinvolle wijze over bias/fairness te spreken, sprake te zijn van een belang dat potentieel wordt geschaad.

Ten behoeve van de bias-toetsing veronderstellen we dat het uiteindelijke besluit over een visumaanvraag een redelijke proxy is voor de aard van de aanvraag (bonafide / malafide).[20] Wanneer we kijken naar de BAO zijn er, ten aanzien van de *fast track* (laag risico kwalificatie) en *intensive track* (hoog-risico kwalificatie) conceptueel dan vier uitkomsten mogelijk:

---

[17] Hoewel het expliciet beleid is dat afwijzingen van visumaanvragen te allen tijde worden voorzien van een weigeringsgrond heeft BZ aangegeven dat in het verleden (sporadisch) is voorgekomen dat slechts de track-uitkomst 'intensive track' als motivatie voor de weigering werd gegeven. Ondanks dat deze gevallen slechts anekdotisch van aard zijn – en niet noodzakelijk een systematisch probleem impliceren - onderstreept dit wel het *risico* van vooringenomenheid als gevolg van track-aanduidingen. Daarnaast stelt BZ in de 'FAQ Informatie Ondersteunend Beslissen' dat "het nadeel [van deze profielen] is dat er vooroordelen kunnen ontstaan waardoor een bepaalde groep als risico of kans wordt gezien terwijl dat niet terecht hoeft te zijn". Tot slot onderkent BZ, in de onderzoeksopdracht, "dat er altijd een risico bestaat dat het gebruikte algoritme de uiteindelijke beslissing van de medewerker beïnvloedt".

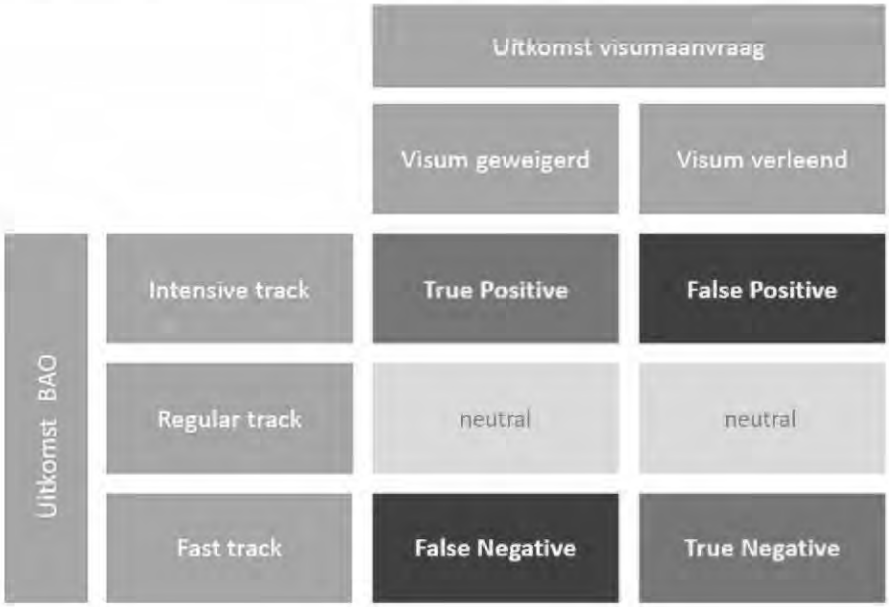
[18] Met een (in termen van de nieuwe Europese AI Verordening) hoog-risico toepassing als de BAO wordt de 'bewijslast' omgekeerd: BZ zal moeten kunnen onderbouwen dat de geconstateerde bias in de BAO niet doorwerkt in de besluitvorming door beslismedewerkers. Aangezien BZ dit niet monitort is er op dit moment geen data beschikbaar om een dergelijke claim (bias werkt niet door) te onderbouwen. Zie paragraaf 5.2.6 voor een aanbeveling ter inrichting van toekomstige monitoring hiervan.

[19] Zie onder meer Skeem et al. (2020), Stevenson, & Doleac (2022), en De-Arteaga et al. (2020).

[20] Zie paragraaf 3.2.2.

1. Een visumaanvraag met *fast track* aanduiding wordt goedgekeurd;
2. Een visumaanvraag met *fast track* aanduiding wordt desondanks afgewezen;
3. Een visumaanvraag met *intensive track* aanduiding wordt desondanks goedgekeurd;
4. Een visumaanvraag met *intensive track* aanduiding wordt afgewezen.

Deze uitkomsten kunnen schematisch als volgt worden weergegeven, waarbij de bovengenoemde uitkomsten worden getypeerd als, respectievelijk, (1) *True Negative*, (2) *False Negative*, (3) *False Positive*, en (4) *True Positive*. De gevallen waarin de BAO resulteert in een *regular track* (geen positieve of negatieve risico-informatie aanwezig) beschouwen wij hier als neutraal (zie Figuur 6).[21]



Figuur 6. Schematische weergave uitkomsten BAO

Vanwege het onder paragraaf 3.2 genoemde risico dat een onjuiste risico-classificering de besluitvorming over een visumaanvraag negatief beïnvloedt, wordt als uitgangspunt voor de BAO track selectie gesteld dat er sprake is van een potentieel nadeel wanneer de visumaanvraag van een bonafide aanvrager de *intensive track* krijgt toegewezen (*False Positive*).[22] De waarschijnlijkheid waarmee dit gebeurt voor bepaalde groepen – de *False Positive Rate* (ofwel het aantal *False*

[21] Deze classificering komt overeen met de wijze waarop BZ in 'Richtlijnen Informatie Ondersteunend Beslissen' de verhouding tussen track-uikomsten en het besluit over de visumaanvraag duidt: "Mocht blijken dat bijvoorbeeld het advies vanuit de BAO afwijkt van de beslissing om het visum wel of niet af te geven, dan zal deze feedback meegenomen in de doorontwikkeling van informatie ondersteund beslissen".

[22] Hoewel een *intensive track* aanduiding niet automatisch betekent dat een visumaanvraag wordt afgewezen is het aannemelijk dat dit wel resulteert in een hogere waarschijnlijkheid dat een visumaanvraag (onterecht) wordt afgewezen (zie ook paragraaf 3.2).

Positives gedeeld door het aantal aanvragen waarbij een visum uiteindelijk wordt verleend) – is daarmee de relevante bias-maatstaf waarin we geïnteresseerd zijn.

Fairness definitie

De waarschijnlijkheid dat visumaanvragen van bonafide aanvragers een intensive track krijgt toegewezen mag niet (wezenlijk) afhankelijk zijn van beschermde persoonskenmerken (zoals nationaliteit, leeftijd of geslacht)

De bijbehorende fairness-definitie is dat de waarschijnlijkheid dat visumaanvragen van bonafide aanvragers een *intensive track* krijgt toegewezen niet (wezenlijk) afhankelijk mag zijn van beschermde persoonskenmerken (zoals nationaliteit, leeftijd of geslacht).[23]

Zoals eerder gesteld in paragraaf 2.4.2 zal er in de praktijk altijd sprake zijn van enige bias-discrepanties tussen groepen. In overleg met BZ is ten aanzien van de BAO gekozen om de veelgebruikte "80% regel" als uitgangspunt te nemen om te bepalen wanneer aangetroffen discrepanties tussen groepen al dan niet acceptabel zijn.

3.2.2 Aannames

Ten behoeve van het kunnen toetsen van bias in de BAO doen we de volgende aannames:

1. Het besluit over de visumaanvraag is een redelijke proxy voor de aard van de aanvraag (bonafide / malafide)
2. De *regular track* aanduiding is een goede (neutrale) baseline.
3. Vergeleken met de baseline resulteert een *intensive track* aanduiding in een hogere kans op een onterechte afwijzing van de visumaanvraag.
4. Vergeleken met de baseline resulteert een *fast track* aanduiding in een hogere kans op een onterechte goedkeuring van de visumaanvraag.

1. Besluit visumaanvraag als proxy voor aard van de aanvraag
Deze veronderstelling lijkt op het eerste gezicht wellicht contra-intuïtief omdat de besluitvorming praktisch nooit geheel bias-vrij is. Het is echter een redelijke aanname dat de besluitvorming voor een grote meerderheid van de aanvragen juist zal zijn (onder meer omdat het aanvraag-dossier meer informatie bevat ter beoordeling door de beslismedewerker dan de zeven datapunten die door het algoritme wordt gebruikt). Substantiële statistische verschillen tussen de track-aanduiding van de BAO en het uiteindelijke besluit over de visumaanvraag geven daarom, zelfs met deze aanname, nog steeds een goed beeld over mogelijk aanwezige bias in de BAO.[24]

[23] Formeel wordt deze definitie ook wel met *False Positive Rate parity* aangeduid.
[24] Hierbij kan verder nog worden opgemerkt dat, vanwege de in de BAO aanwezige feedbackloop van historische weigeringspercentages, eventuele bias bij de beslismedewerkers waarschijnlijk de bias in de BAO versterkt (in plaats van afzwakt). Met andere woorden, de systematische onder- of overschatting van risico's ten aanzien van bepaalde groepen zal in de uiteindelijke besluitvorming eerder groter dan kleiner zijn dan in de BAO zelf (zie ook paragraaf 4.3.3).

2. *Regular track* als baseline
Wanneer een aanvraag de *regular track* aanduiding krijgt is de bijbehorende toelichting dat er geen enkele positieve of negatieve risico-informatie ten aanzien van de aanvraag bekend is. Dit betekent eveneens dat een aanvraag geen lage of hoge risico-indicatie meekrijgt.

3. Verhoogde kans op onterechte afwijzing door *intensive track*
Wanneer een aanvraag de *intensive track* aanduiding krijgt wordt een verhoogde risico-indicatie meegegeven aan de beslismedewerker: De medewerker krijgt het volgende bericht te zien: "*LET OP: Aanvraag valt in groep met verhoogd risico op misbruik visumprocedure*". Op basis van de onder paragraaf 3.2 gegeven argumenten hanteren wij de aanname dat de kans op een onterechte afwijzing van een bonafide visumaanvraag met een intensive track aanduiding (*False Positive*) hoger is dan in het geval van een *regular track* aanduiding (baseline).

4. Verhoogde kans op onterechte goedkeuring door *fast track*
Wanneer een aanvraag de *fast track* aanduiding krijgt wordt een lage risico-indicatie meegegeven aan de beslismedewerker. De medewerker krijgt het volgende bericht te zien: "*Aanvraag valt in groep met weinig tot geen risico op misbruik visumprocedure*". Op basis van de onder paragraaf 3.2 gegeven argumenten hanteren wij de aanname dat de kans op een onterechte goedkeuring van een malafide visumaanvraag met een *fast track* aanduiding (*False Negative*) hoger is dan in het geval van een *regular track* aanduiding (baseline).

3.2.3 Gebruikte data

Conform het huidige beleid van het Ministerie van Buitenlandse Zaken bedraagt de bewaartermijn van data vijf jaar. De gebruikte data heeft daarom betrekking op een periode van nooit langer dan vijf jaar geleden.

Door de aard van de opzet van de BAO en de interactie met NVIS hebben we twee verschillende datasets ontvangen: een NVIS dataset en een BAO dataset. De datasets bevatten afzonderlijke informatie over dezelfde set visumaanvragen en zijn voor het doel van de biasanalyse aan elkaar gekoppeld.

NVIS data

Huidige en historische visumaanvragen worden opgeslagen en bewaard in het Nieuw Visum Informatiesysteem (NVIS). De BAO maakt het mogelijk om deze visumaanvragen te koppelen aan geconfigureerde databronnen van de ketenpartners. Nadat deze koppeling is gemaakt, verstuurt de BAO informatie terug naar NVIS over de uitkomst van de track voor een gegeven aanvraag.

Per aanvraag bevat de NVIS dataset gegevens over trackuitkomst, beslissingsuitkomst en de volgende zeven kenmerken: hoofddoel van de reis, post van aanvraag, nationaliteit, geslacht, leeftijdsklasse, burgerlijke staat en beroep.

Van de zeven kenmerken zijn er drie gebruikt voor de analyse: nationaliteit, geslacht en leeftijdsklasse (zie Tabel 4).

De NVIS dataset bevat gegevens van de afgelopen vijf jaar, vanaf het begin van 2018.

BAO data

Binnen de BAO worden de profielen gegenereerd door het beslisboomalgoritme. Tijdens de koppeling met NVIS wordt een visumaanvraag gecontroleerd op een match met een profiel. De BAO dataset bevat per aanvraag gegevens over een eventuele match met een profiel (bijvoorbeeld kans, risico of trend) en de resulterende track-uitkomst. Daarnaast bevat de BAO dataset gegevens of een individuele aanvrager, werkgever of referent, gerelateerd aan een visumaanvraag, bekend is bij een of meerdere van de ketenpartners (zie ook paragraaf 1.2.3).

In het begin van 2020 is BAO 2.0 geïntroduceerd. De koppeling van visumaanvragen met databronnen van ketenpartners vindt gescheiden van het primaire proces plaats binnen de BAO. Dit betekent dat een selectie van bronnen en data niet toegankelijk is voor medewerkers. Vanwege deze opzet is er geen BAO data beschikbaar van vóór BAO 2.0 (zie Tabel 4).

De BAO dataset bevat alleen gegevens vanaf het begin van 2020.

Tabel 4. Gegevens in de twee gebruikte datasets

NVIS	BAO
Jan 2018—Nov 2022	Jan 2020—Nov 2022
Visumaanvraagnummer	Visumaanvraagnummer
Trackuitkomst	Trackuitkomst
Beslissingsuitkomst	Match op profiel
Kenmerken: hoofddoel reis, post, nationaliteit, geslacht, leeftijdsklasse, burgerlijke staat, beroep	Match op bronnen van ketenpartners

3.2.4 Gebruikte software

Voor de data-analyses is gebruikt gemaakt van de *Python* programmeertaal (versie 3.11.1) en de *Aequitas* package (versie 0.42.0). *Aequitas* is een open-source softwarepakket voor het evalueren van modellen op bias en fairness. Deze software is gebruikt om de resultaten van de BAO te evalueren op verschillende bias-criteria[25].

[25] Zie bijvoorbeeld Saleiro et al. (2018).

3.2.5 Selectie groepen op basis van persoonskenmerken

Voor de biasanalyses op basis van nationaliteit is een subselectie gemaakt van 60 nationaliteiten in samenspraak met BZ. Deze subselectie bevat zowel kwetsbare nationaliteiten als niet-kwetsbare landen geïdentificeerd door BZ, inclusief een selectie van overige landen om de totale subselectie zo divers mogelijk te maken. De definitie van kwetsbare nationaliteiten heeft hier betrekking op nationaliteiten met hoge gemiddelde weigeringspercentages en hoge gemiddelde aantallen aanvragen die resulteren in ongewenst gedrag.

4 Bevindingen

De ontvangen datasets zijn gebruikt om kwantitatieve biasanalyses uit te voeren op basis van gegevens van uitkomsten van visumaanvragen sinds december 2019 (vanaf lancering BAO 2.0). Deze analyses zijn uitgevoerd met ondersteuning van de softwarepackage *Aequitas* (zie 3.2.4 Gebruikte software). Naast de resultaten van deze kwantitatieve biasanalyses komen ook andere bevindingen ten aanzien van de BAO aan bod die mogelijk bias en onwenselijke effecten kunnen veroorzaken.

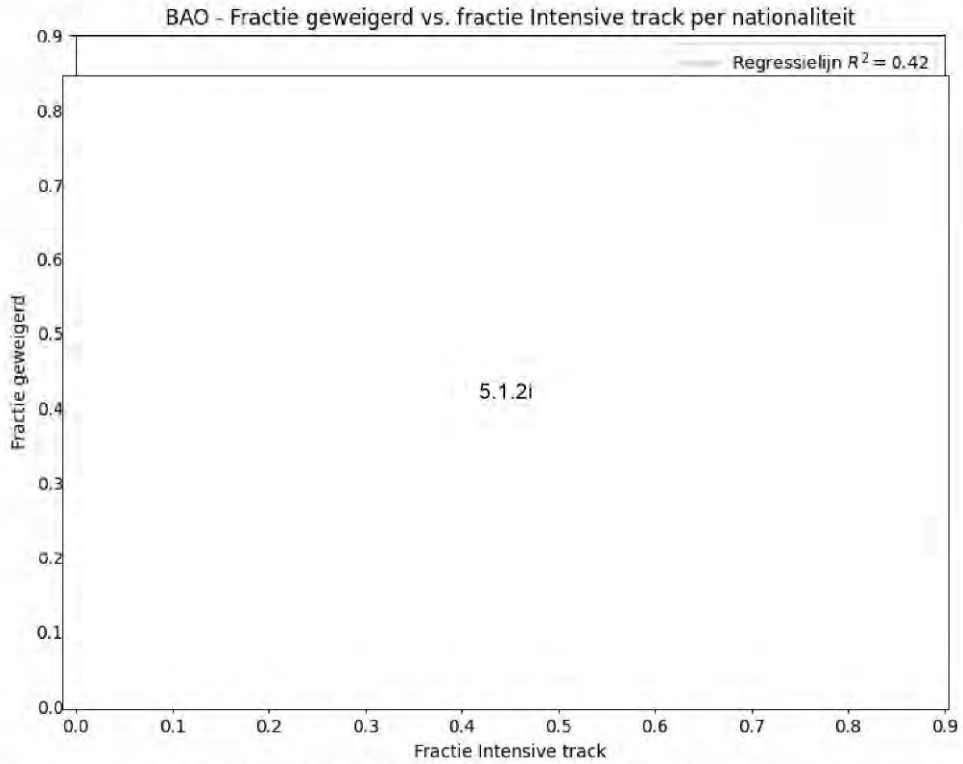
4.1 Kalibratie van de BAO

Een goed gekalibreerd model zorgt ervoor dat bias in de uitkomsten van het model wordt geminimaliseerd tussen verschillende demografische groepen. Bij een goed gekalibreerde BAO zal, met andere woorden, sprake zijn van een duidelijke proportionele relatie tussen de risicoclassificatie-percentages en de daadwerkelijke toelatings- en weigeringspercentages over nationaliteiten (al hoeft deze niet 1-op-1 te zijn).[26] Voordat de bevindingen van de biasanalyses worden besproken, wordt eerst een paragraaf gewijd aan de kalibratie van de BAO.

Figuur 7 toont per nationaliteit de fractie aanvragen met een *intensive track* uitkomst en de fractie aanvragen die zijn geweigerd. Hoewel deze figuur nog geen directe informatie bevat over bias, geeft het wel een overzicht van de spreiding in weigeringspercentages en *intensive track*-percentages. Meer informatie over de reden waarom deze selectie van nationaliteiten is gebruikt, is te vinden in paragraaf 3.2.5.

Aannemende dat beslismedewerkers over het algemeen nauwkeurig en redelijkerwijze objectief onderzoek verrichten tijdens het bekijken van een dossier van een aanvraag, kan men aannemen dat de fractie geweigerde aanvragen per nationaliteit informatie geeft over de hoogte van het risico van aanvragen voor een gegeven nationaliteit.[27] Met deze aanname kan weigeringspercentage als een grove en imperfecte indicatie worden gezien van hoe hoog de fractie *intensive track* zou moeten zijn voor een gegeven nationaliteit.

[26] Hierbij dient te worden opgemerkt dat wanneer er sprake is van een proportionele relatie tussen risicoclassificatie-percentages en toelatings/weigeringspercentages dit niet impliceert dat het model goed gekalibreerd is. Immers, een risicomodel dat simpelweg de intensive track toewijst met aan aanvragen op basis van slechts het weigeringspercentage zal resulteren in een proportionele relatie, maar is uiteraard geen goed gekalibreerd risicomodel.
[27] Het structureel gebruik van weigeringspercentage als risico-indicator kan op termijn wel bijdragen aan bias vanwege de feedback-loop die dit veroorzaakt.

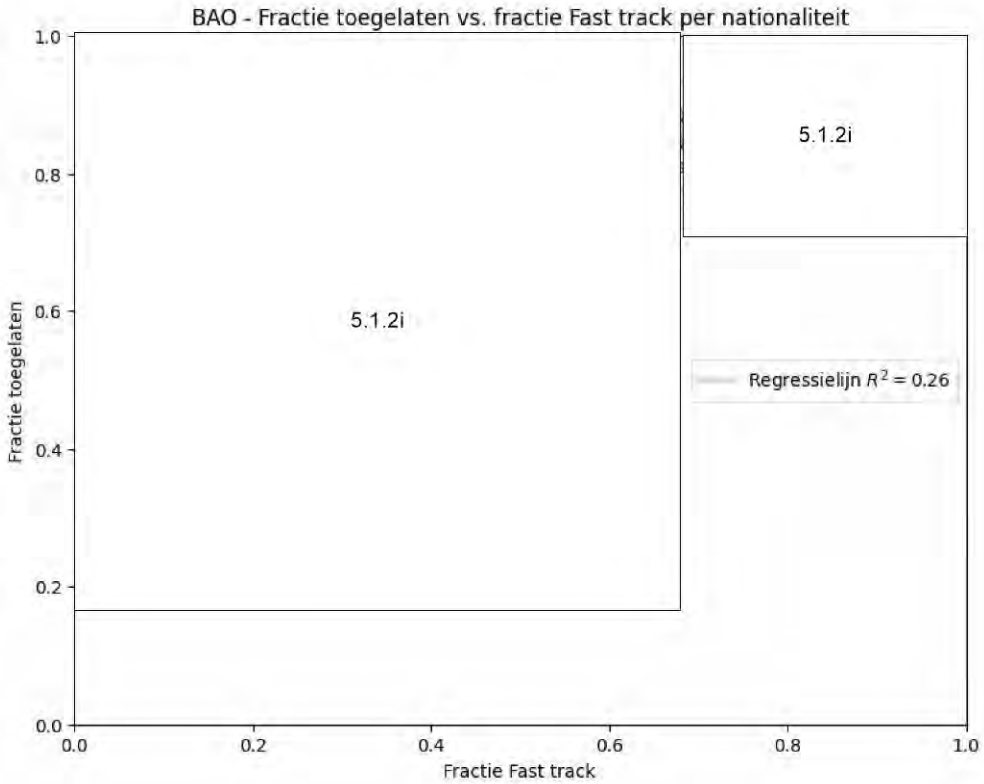


Figuur 7. Fractie aanvragen geweigerd vs. fractie aanvragen met intensive track uitkomst per nationaliteit (BAO data januari 2020—november 2022)

Uit Figuur 7 kunnen we opmaken dat er geen klaarblijkelijke relatie bestaat tussen weigeringspercentage en *intensive track*-percentage geproduceerd door de BAO per nationaliteit. De correlatie tussen fractie geweigerd en fractie *intensive track* is laag: slechts 42% van de variatie tussen de punten kan worden verklaard door een correlatie tussen weigering en intensive track uitkomst (R^2 = 0.42). Dit heeft een grote spreiding in *intensive track*-percentages tussen nationaliteiten als resultaat.

Hetzelfde kan worden gezegd over de kalibratie van toelatingspercentage en *fast track*-percentage, waar slechts 26% van de variatie tussen de punten kan worden verklaard door een correlatie tussen toelating en *fast track* uitkomst (Figuur 8; R^2 = 0.26).

Een mogelijke conclusie die hieruit zou kunnen worden getrokken, is dat de BAO niet goed gekalibreerd is: mogelijk bevat de BAO te weinig relevante informatie om accuraat een conclusie te kunnen trekken met betrekking tot het daadwerkelijke risicoprofiel van een aanvraag. Zo bevat de BAO bijvoorbeeld mogelijk niet genoeg informatie om 5.1.2i aanvragen vaker in te schatten als intensive (zie Figuur 7). Deze informatie kan voor een medewerker met toegang tot het dossier onmiddellijk duidelijk zijn, maar in de BAO ontbreekt deze informatie. Dit heeft als gevolg dat de BAO een misleidende risicoclassificering mee kan geven aan een beslismedewerker.



Figuur 8. Fractie aanvragen toegelaten vs. fractie aanvragen met fast track uitkomst per nationaliteit (BAO data januari 2020—november 2022)

In Bijlage 4 is een uitgebreide analyse van de prestatie van de BAO op basis van nationaliteit van de aanvrager terug te vinden, aan de hand van acht verschillende performance en error metrieken. De uitkomsten van deze analyse suggereren dat voor aanvragers met bepaalde nationaliteiten (afhankelijk van de gekozen metriek ongeveer 30-50% van de onderzochte nationaliteiten) het gegeven nationaliteit effectief de enige ter zake doende risicofactor is ten behoeve van de track aanduiding.[28]

False Positive Rate (FPR)

De FPR verwijst naar het percentage goedgekeurde aanvragen dat door de BAO als intensive wordt gecategoriseerd. Een aanvrager wordt door deze interventie mogelijk onterecht benadeeld.

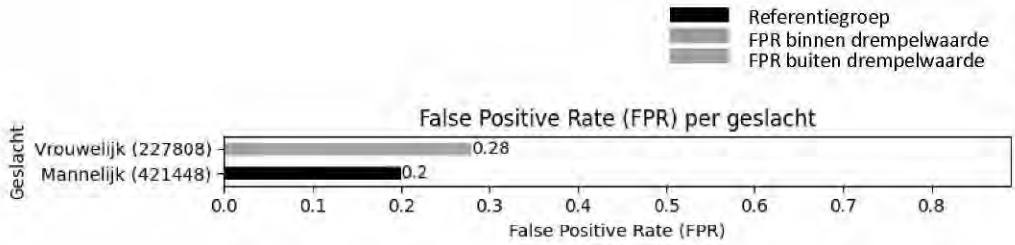
Resultaten bias-toetsing

Hieronder volgen de resultaten van de kwantitatieve biasanalyses voor drie demografische kenmerken: geslacht, leeftijd en nationaliteit. Deze resultaten zijn beschreven vanuit het perspectief van de intensive track (*False Positive Rate*; FPR). Het gevolg van een bias tussen *FPR's* is een mogelijk nadeel voor aanvragers behorende tot een bepaalde demografische groep. Zie Bijlage 5 voor de resultaten vanuit het perspectief van de *fast track* (*False Negative Rate*).

[28] In termen van het College voor de Rechten van de Mens (2021) is voor deze groepen het gegeven nationaliteit "het enige of doorslaggevende selectiecriterium binnen het risicoprofiel" (p. 42).

4.2.1 Geconstateerde bias op basis van geslacht

De resultaten van de analyse suggereren dat er mogelijk sprake is van benadeling op basis van geslacht veroorzaakt door de BAO. Op basis van Figuur 9 kan geconstateerd worden dat de *FPR* voor vrouwelijke aanvragers aanzienlijk hoger is dan die voor mannelijke aanvragers. Hierbij is de vooraf bepaalde bias drempelwaarde gehanteerd (een verschil met factor 0.8-1.25 wordt geaccepteerd, daarbuiten wordt het verschil onacceptabel geacht).



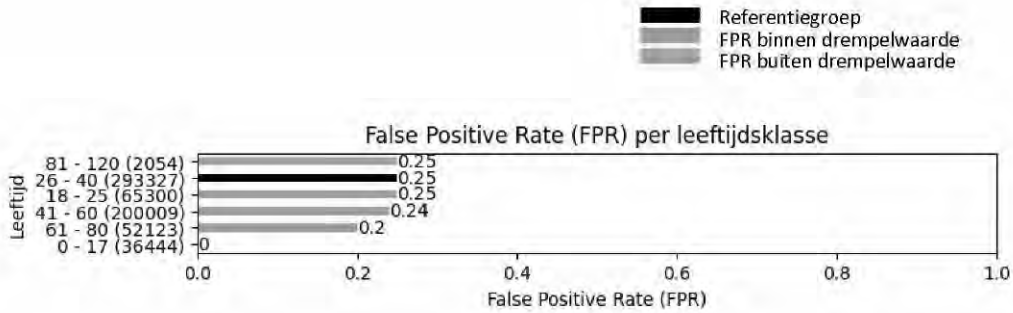
Figuur 9. False Positive Rate (FPR) per geslacht van de aanvrager. Een groene en rode staaf betekent dat de FPR van de betreffende groep respectievelijk binnen en buiten de acceptabele range van de biasdrempelwaarde valt ten opzichte van de referentiegroep (zwart).

Deze discrepanties suggereren dat er mogelijk sprake is van ongelijke behandeling op basis van geslacht in de BAO, namelijk dat bonafide vrouwelijke aanvragers vaker als intensive worden gecategoriseerd dan bonafide mannelijke aanvragers.

4.2.2 Geconstateerde bias op basis van leeftijd

De resultaten van de analyse suggereren dat er geen sprake is van een benadeling op basis van leeftijd veroorzaakt door de BAO. In Figuur 10 kan worden geconstateerd dat de *FPR's* tussen verschillende leeftijdsklassen niet aanzienlijk van elkaar verschillen. Hierbij is weer de vooraf bepaalde bias drempelwaarde gehanteerd (een verschil met factor 0.8-1.25 wordt geaccepteerd) waarbij de *FPR's* voor alle leeftijdsklassen binnen het bereik van de drempelwaarde vallen ten opzichte van de referentiegroep.[29]

[29] Het ontbreken van een *FPR* voor de leeftijdsklasse 0-17 wordt verklaard door het feit dat minderjarigen niet door de BAO worden geprofileerd en daarom wordt deze leeftijdsgroep in deze bias-analyse buiten beschouwing gelaten.



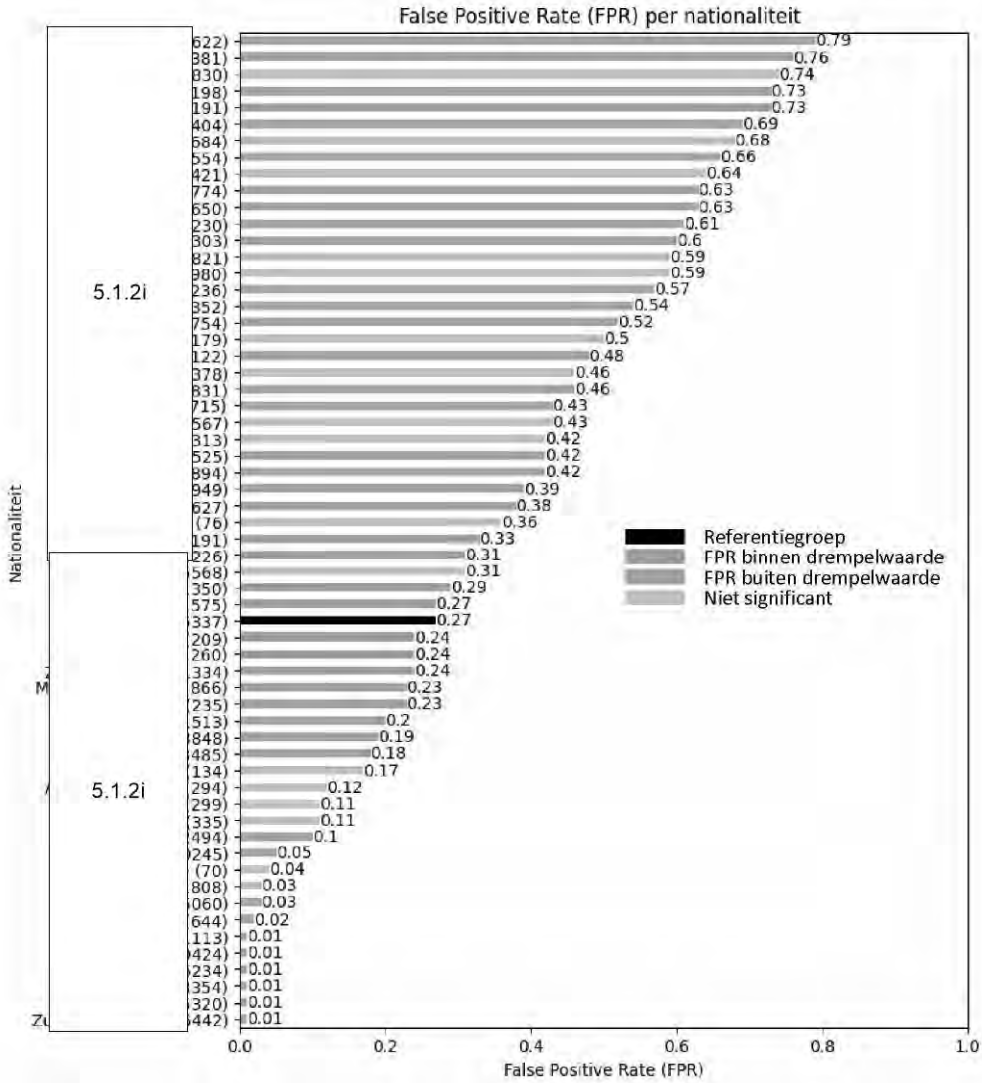
Figuur 10. False Positive Rate (FPR) per leeftijdsklasse van de aanvrager. Een groene en rode staaf betekent dat de FPR van de betreffende groep respectievelijk binnen en buiten de acceptabele range van de biasdrempelwaarde valt ten opzichte van de referentiegroep (zwart).

4.2.3 Geconstateerde bias op basis van nationaliteit

De resultaten van de analyse suggereren dat er sprake is van een mogelijke benadeling op basis van nationaliteit veroorzaakt door de BAO.

Er zijn zeer grote verschillen in *FPR* gevonden tussen nationaliteiten (zie Figuur 11 voor een staafdiagram en Figuur 12 voor een landkaart overzicht van de selectie van 60 nationaliteiten). Ten opzichte van de referentiegroep 5.1.2i vertonen verreweg de meeste nationaliteiten een aanzienlijk hogere of lagere *FPR* (zie de rode staven in Figuur 11). Ongeacht wat als referentiegroep zou zijn gekozen, zouden deze grote discrepanties aanwezig zijn [ ] is gekozen als referentiegroep omdat het overeenkomt met het gewogen gemiddelde van *FPR's*).

Deze discrepanties suggereren dat er mogelijk sprake is van ongelijke behandeling op basis van nationaliteit door de BAO, namelijk dat bonafide aanvragen met een nationaliteit voornamelijk afkomstig uit 5.1.2i 5.1.2i relatief vaker als intensive worden gecategoriseerd dan bonafide aanvragen met een andere nationaliteit.



Figuur 11. False Positive Rate (FPR) per nationaliteit van de aanvrager (selectie nationaliteiten). Een groene en rode staaf betekent dat de FPR van de betreffende groep respectievelijk binnen en buiten de acceptabele range van de biasdrempelwaarde valt ten opzichte van de referentiegroep (zwart). Een grijze staaf betekent geen statistisch significant verschil met de referentiegroep.

Figuur 12. False Positive Rate (FPR) per nationaliteit van de aanvrager (selectie landen)

4.2.4 Effect van de Covid-19 pandemie

Op verzoek is er onderzocht welke invloed de Covid-19 pandemie heeft op de resultaten van de biasanalyses. Hiervoor is dezelfde biasanalyse uitgevoerd op gegevens van zowel het jaar voorafgaand aan de pandemie (2019) als het meest recente jaar (2022). De resultaten van deze analyses zijn vervolgens met elkaar vergeleken. Helaas is een nauwkeurige vergelijking niet mogelijk, aangezien op het moment dat de pandemie uitbrak, er een nieuwe versie van de BAO is gelanceerd.

Desondanks tonen de resultaten aan dat er zowel voor als na de pandemie aanzienlijke verschillen aanwezig zijn in de *FPR* op basis van nationaliteit (zie Figuur 22 in Bijlage 6).

4.3 Overige bevindingen

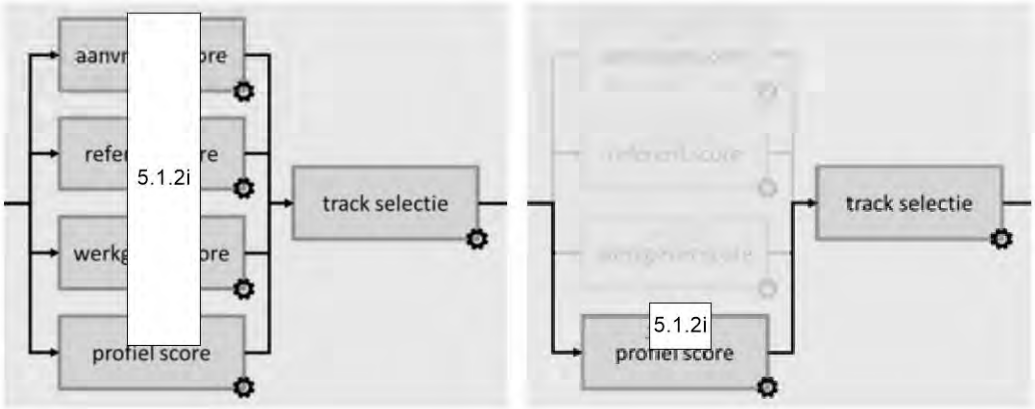
Naast de kwantitatieve biasanalyse, waarvan de resultaten hierboven staan beschreven, zijn er een aantal andere bevindingen gedaan gerelateerd aan het functioneren van de BAO die tot bias en andere mogelijke ongewenste effecten kunnen leiden.

4.3.1 Impact profielen op trackselectie

De impact van de profielen in de BAO wordt sterk onderschat. In zowel de documentatie van de werking van de BAO als in communicatie over de BAO naar beslismedewerkers en naar buiten wordt beweerd dat profielen maar voor 10%

meewegen in de bepaling van de uiteindelijke track van een aanvraag (zie Figuur 13). Informatie over de aanvrager, referent en werkgever zouden hierbij voor 90% meewegen.

Uit de data blijkt echter dat bij 98.4% van de KVV aanvragen informatie over de aanvrager, referent en werkgever ontbreekt.[30] In deze gevallen betekent dat, op het moment dat een aanvraag een match met een profiel heeft, deze niet voor 10% maar effectief voor 100% meeweegt in de bepaling van de track. Daarmee geeft de communicatie over het gebruik van profielen in de BAO (waarin wordt gesteld dat de profielscore slechts voor 10% meetelt) een onjuist beeld.



Figuur 13. Schematische weergave van de score onderdelen (aanvrager, referent, werkgever en profiel) met bijbehorende weging die resulteren in de uiteindelijke track. Situatie zoals gecommuniceerd (links) en situatie zoals aanwezig bij verreweg de meeste aanvragen (rechts).

4.3.2 Conclusies over potentiële medewerkersbias in het licht van de BAO dekkingsgraad

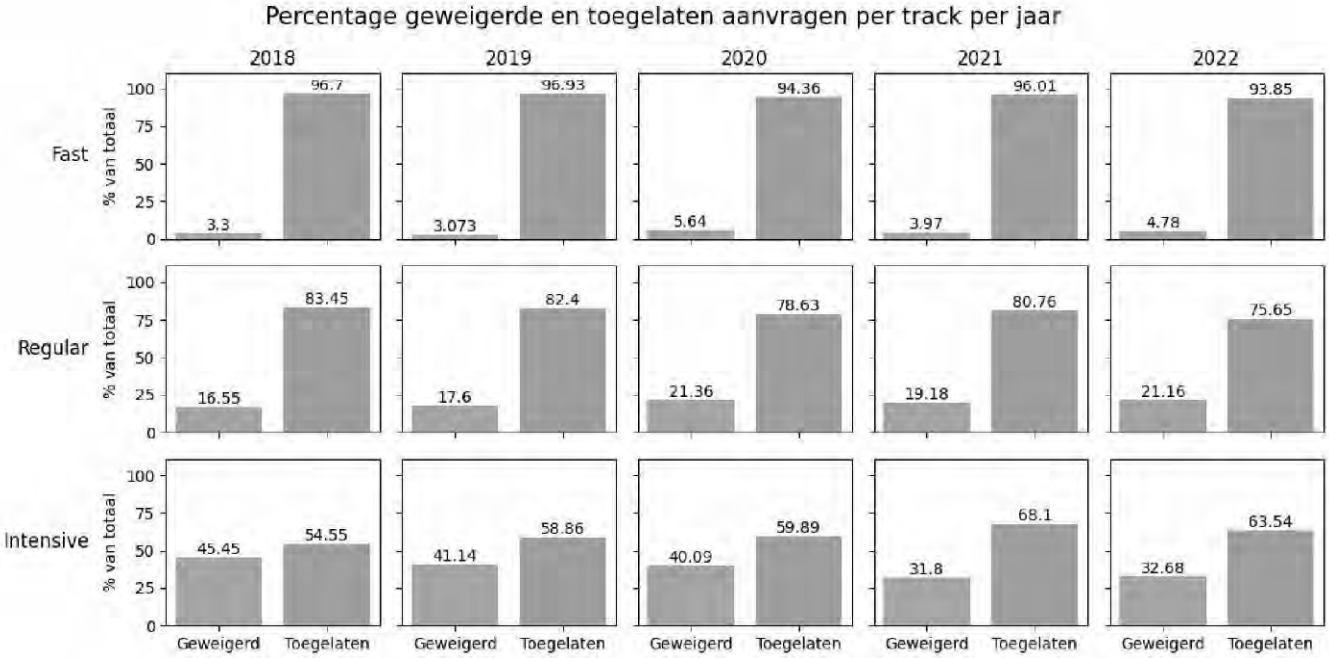
Een conclusie die is gecommuniceerd door BZ met betrekking tot de afwezigheid van medewerkersbias is hoogstwaarschijnlijk niet gerechtvaardigd op basis van de gegeven argumenten. Er wordt beweerd dat er geen signalen van medewerkersbias zijn waargenomen op basis van een daling in het weigeringspercentage voor de intensive track sinds 2017, van bijna 56% naar 33%.[31] Op basis van deze cijfers wordt geconcludeerd dat het gebruik van de BAO niet leidt tot een mechanisme van 'zoekt en gij zult vinden'. Dit zou namelijk resulteren in een steeds hoger weigeringspercentage voor de *intensive track*.

Hoewel de waargenomen daling in het weigeringspercentage wel degelijk vast te stellen is en dezelfde cijfers zijn gereproduceerd (zie

[30] Maar 10.668 van de 649.257 (1.64%) van de aanvragen heeft een match op een bron van een of meerdere ketenpartners. Voor de rest van de aanvragen is geen informatie beschikbaar.
[31] Zie Terms of Reference voor dit onderzoek

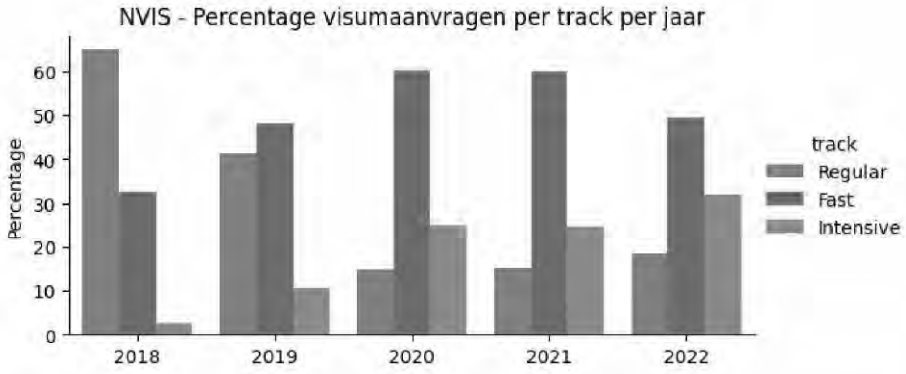
Figuur 14 Figuur 14 voor de cijfers van 2018 t/m 2022), kan bovenstaande bewering niet gerechtvaardigd worden op basis van deze cijfers.

Het is namelijk in lijn der verwachting dat het weigeringspercentage voor de intensive track de afgelopen jaren is afgenomen. Deze daling wordt namelijk veroorzaakt door een zeer sterke stijging in het aantal aanvragen in de *intensive track*, terwijl het aantal aanvragen in de *regular track* juist wordt geminimaliseerd (zie Figuur 15). Dit is in lijn met de wens van BZ om zoveel mogelijk aanvragen in te delen in de *fast* en *intensive track*. Dit heeft als implicatie dat er naast malafide aanvragen ook steeds meer bonafide aanvragen in de *intensive track* terechtkomen, wat logischerwijze resulteert in een dalend weigeringspercentage.[32]



Figuur 14. Percentages geweigerde en toegelaten aanvragen per track, per jaar vanaf 2018 t/m 2022. Met als doel het overzicht te verbeteren, is een derde categorie (geen beslissing) weggelaten.

[32] Daarnaast dient te worden opgemerkt dat op basis van (de ontwikkeling van) weigeringspercentages in de verschillende tracks op zichzelf geen enkele conclusies kunnen worden getrokken over (de ontwikkeling van) eventuele bias op medewerker niveau. Daarvoor is nader onderzoek nodig (zie tevens paragraaf 5.2.6).

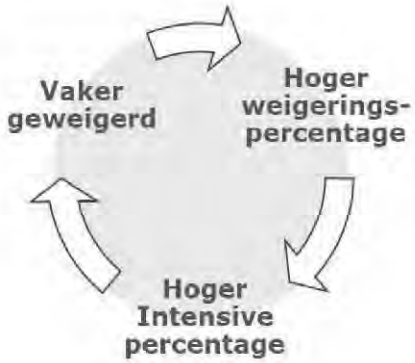


Figuur 15. Percentage visumaanvragen per track per jaar. Duidelijk te zien is dat het aantal *regular* aanvragen zwaar is afgenomen.

4.3.3 Het gebruik van weigeringspercentages ten behoeve van profilering

Het onderliggende beslisboomalgoritme dat de profielen genereert werkt op basis van door de bouwers vastgestelde beslisregels (zie paragraaf 1.2.3). Een aantal van deze regels die profielen genereren zijn puur en alleen gebaseerd op weigeringspercentage: een risico- of kansprofiel wordt opgesteld voor een groep wanneer het weigeringspercentage respectievelijk hoger of lager is dan het gemiddelde weigeringspercentage. Deze regels zijn toegevoegd om de dekkingsgraad voor de intensive en *fast track* van de BAO te verhogen.

Het gebruiken van weigeringspercentage in de bepaling van een track is echter zeer risicovol met het oog op bias: een bias feedback loop kan ontstaan waar een hoger weigeringspercentage zorgt voor een hoger aantal *intensive track* aanvragen, wat vervolgens weer een hoger weigeringspercentage kan veroorzaken voor een bepaalde demografische groep. Eenzelfde feedback loop kan ontstaan voor de *fast track*.



4.3.4 Het beslisboomalgoritme en de kwaliteit van profielen

Met betrekking tot het functioneren van het beslisboomalgoritme is een gebrek geconstateerd aan 1) documentatie over de totstandkoming van de regels op basis waarvan het algoritme functioneert (zie paragraaf 1.2.3), 2) de validatie van de performance van de resulterende profielen en 3) de continue monitoring en documentatie hiervan.

In de gedeelde documentatie van de BAO wordt genoemd dat de zeven variabelen op basis waarvan de profielen zijn samengesteld in deze specifieke volgorde gekozen zijn, omdat is onderzocht dat variabelen in deze volgorde het meest

onderscheidend zijn.[33] Echter is dit onderzoek nergens vastgelegd. Hetzelfde geldt voor de totstandkoming van de verschillende regels op basis waarvan de profielen worden gegenereerd (zie Tabel 2 in paragraaf 1.2.3). Wat is bijvoorbeeld de reden waarom een subgroep minimaal 200 aanvragen moet bevatten? Waarom is een risicoprofiel type 2 gebaseerd op een hitpercentage tussen de 1-5% en een weigeringspercentage van minimaal 16%? Informatie hierover is niet in de beschikbare documentatie terug te vinden.

Eveneens is niets in de documentatie te vinden over de validatie van deze profielen. Hoe wordt er gevalideerd dat deze profielen doen wat ze behoren te doen? Met andere woorden; worden de juiste profielen door de BAO uitgefilterd en in de juiste track geplaats? Hoe is dit getest? Hoe wordt dit blijvend en voortdurend gemonitord, en zo nodig geüpdatet, zeker in het geval van veranderende situaties, zoals de Covid-19 pandemie? Bij een gebrek aan monitoring loopt het systeem gevaar om trends van 5 jaar geleden uit te vergroten en in het hier en nu toe te passen. Dit zorgt voor mogelijke bias in de BAO en track uitkomsten.

Uit gesprekken met de ontwikkelaars van de BAO en het onderliggende algoritme blijkt dat de initiële intentie bij het gebruik van de BAO anders is dan wat terugkomt in de gedeelde documentatie. Vanuit het oogpunt van de ontwikkelaars fungeert de BAO niet met het doel van risicoclassificatie, maar om een indicatie te geven van de tijdsduur en intensiteit waarmee een aanvraag behandelt moet worden. Dit staat echter in veel gevallen haaks op wat er in de documentatie wordt beschreven. De berichten die de medewerkers te zien krijgen bij een aanvraag met een *fast track*: "Aanvraag valt in groep met weinig tot geen risico op misbruik visumprocedure", en met een *intensive track*: "LET OP: Aanvraag valt in groep met verhoogd risico op misbruik visumprocedure" impliceren dat dit een risicoclassificatie betreft met de mogelijke gevolgen van stigmatisatie en *automation bias* van dien.

4.3.5 Conformiteit met BZ fairness standaard

Aan de hand van het document *Beleid werken met data en algoritmen* is getoetst of de BAO conform de BZ fairness standaard functioneert. [34] Hieruit is gebleken dat de BAO niet voldoet aan alle negen basisprincipes voor het werken met data en algoritmen.

1. Bij BZ wordt bewust omgegaan met en uitgegaan van fundamentele rechten, ethiek en informatiebeveiliging bij werken met data en algoritmen.

Als deel van dit principe wordt fairness gedefinieerd als een algoritme dat geen oordeel over een persoon of een groep personen heeft. Op basis van de uitkomsten

[33] Richtlijnen Informatie Ondersteund Beslissen
[34] Beleidsnotitie *Werken met data en algoritmen* (november 2021) 5.1.2e 5.1.2e

van de biasanalyses kan niet worden gesteld dat de BAO aan dit fairness criterium voldoet.

2. Kwaliteit van data is bij BZ toereikend voor het doel waarvoor de data gebruikt worden.

Het gebruik van convenanten met ketenpartners en duidelijke afspraken over doelbinding en dataminimalisatie draagt bij aan de kwaliteit van data. Echter zal BZ moeten nadenken over de mogelijkheid van '*underfitting*' van het model door mogelijk te weinig relevante informatie waarop de BAO profielen genereerd.

3. BZ weet wie verantwoordelijk is en diegene neemt verantwoordelijkheid voor de registratie, inhoud en kwaliteit van de data, de algoritmen en de informatieproducten die BZ gebruikt.

4. BZ betrekt een brede diversiteit aan perspectieven bij het valideren van aannames en uitkomsten van data- en algoritmegebruik.

Het is gebleken dat de BAO langere tijd is gebruikt zonder validatie en monitoring van het functioneren ervan en zonder updates in het systeem die veranderingen in de tijd meewegen. BZ zal beter moeten kijken naar de invulling van de rollen en hun verantwoordelijkheid voor het beheer en gebruik van data en het BAO systeem, met name het algoritme en de generatie van profielen.

5. BZ werkt transparant en uitlegbaar met data en algoritmen.

6. Een informatieproduct, al dan niet met algoritme, is altijd van toegevoegde waarde voor een proces.

De documentatie over de BAO blijkt incompleet, ontoereikend en onoverzichtelijk. Het wordt geadviseerd om de documentatie van de BAO zo beknopt mogelijk te maken in zo min mogelijk verschillende documenten. Hierbij dient deze documentatie bij updates in de BAO onmiddellijk geactualiseerd te worden.

Communicatie over de BAO in de documentatie komt bovendien niet overeen met wat door de ontwikkelaars en managers met de BAO wordt beoogd. Deze discrepanties kunnen doorwerken op het niveau van de communicatie naar beslismedewerkers en uiteindelijk hoe deze de BAO uitkomsten gebruikt. Meer onderzoek moet worden gedaan over hoe beslismedewerkers de track uitkomst gebruiken in hun beslisproces om te kunnen oordelen of het algoritme van toegevoegde waarde is.

7. Algoritmen worden bij BZ alleen ingezet wanneer dit het best beschikbare middel voor het beoogde doel is (subsidiariteit).

Dit basisprincipe stelt dat bij de overweging een algoritme in te zetten *by design* door het multidisciplinaire team expliciet, door middel van vastlegging in het

informatieproduct document (IPD), uitgesloten wordt dat een ander beschikbaar middel niet beter is om het beoogde doel te behalen. Uit de documentatie is niet gebleken dat dit voldoende is getoetst, zoals het toetsen van de BAO ten opzichte van een eenvoudiger en simpeler basismodel. Om goed in te schatten of de BAO het best beschikbare middel is voor het beoogde doel, moet worden getoetst of de BAO beter functioneert dan mogelijke alternatieve versies van de BAO. Dit moet vervolgens duidelijk worden vastgelegd in de nieuwste versie van de officiële BAO documentatie. Waarom is bijvoorbeeld gekozen voor een drie track systeem, een geen twee track systeem (met alleen *regular* en *fast track*)? Waarom is gekozen voor een specifieke volgorde van zeven kenmerken en geen andere volgorde met andere kenmerken?

8. Algoritmegebruik staat bij BZ in verhouding tot het beoogde doel (proportionaliteit).

Zoals hierboven bij basisprincipe 6 al kort genoemd, blijkt uit de documentatie niet dat de utiliteit van het algoritme is onderzocht en gekwantificeerd. Wat levert het gebruik van de BAO concreet op en wanneer is de inzet van de BAO proportioneel?

9. Een overeenkomst is de basis voor afname en levering van databronnen van en door derden.

Geadviseerd wordt om een sjabloon overeenkomst te maken die bij afname of levering van data standaard wordt gebruikt en afgestemd kan worden op de context waarvoor deze data gebruikt dient te worden.

4.4 Conclusies

De BAO is geboren uit de intentie om het beslisproces voor KVV aanvragen objectiever en efficiënter te maken. De alternatieve optie om geen gebruik te maken van een BAO-achtig model is vatbaar voor vooringenomenheid zoals *confirmation bias* op het individuele niveau van beslismedewerkers. Desondanks kan de BAO mogelijk nieuwe vormen van bias teweegbrengen en bestaande bias juist versterken. Het mag duidelijk zijn dat het onmogelijk is om elke vorm van bias volledig te voorkomen. Het is daarom van belang om de ongunstige impact van de BAO zo grondig mogelijk uit te sluiten. Daarvoor dient dit onderzoek en de daaruit voorgekomen aanbevelingen.

4.4.1 Conclusies ten aanzien van de onderzoeksresultaten

Op basis van de resultaten kunnen een aantal conclusies worden getrokken met betrekking tot bias in het BAO.

Er is sprake van beperkte bias op basis van geslacht. Bonafide aanvragen van vrouwen hebben een 1.4x zo hoge waarschijnlijkheid om als intensive track te

worden aangeduid als bonafide aanvragen van mannen. Afhankelijk van de mate waarin deze bias al dan niet doorwerkt in de besluitvorming over visumaanvragen kan deze bevinding worden gezien als een aanwijzing voor een mogelijke beperkte bias in de besluitvorming over visumaanvragen ten nadele van vrouwen.

Er is geen sprake van aanzienlijke bias op basis van leeftijd. De discrepanties in de waarschijnlijkheid dat bonafide aanvragen als intensive track worden aangeduid vallen voor iedere leeftijdsgroep binnen de "80% regel".

Er is sprake van aanzienlijke bias op basis van nationaliteit. De discrepanties in de waarschijnlijkheid dat bonafide aanvragen als intensive track worden aangeduid zijn zeer groot. Vergeleken met het gemiddelde over alle aanvragen, hebben bonafide aanvragen van aanvragers met bepaalde nationaliteiten van een 27x lagere tot een 2.9x hogere waarschijnlijkheid om een intensive track aanduiding te krijgen. In het meest extreme geval hebben bonafide aanvragers met bijvoorbeeld een 5.1.2i nationaliteit ongeveer een 70x zo hoge waarschijnlijkheid om een intensive track aanduiding te krijgen als bonafide aanvragers met een 5.1.2i 5.1.2i nationaliteit.

Deze waarden overschrijven ruimschoots de als acceptabel gestelde bandbreedte (van een 1.25x zo lage/hoge waarschijnlijkheid) op basis van de "80% regel". Afhankelijk van de mate waarin deze bias al dan niet doorwerkt in de besluitvorming over visumaanvragen kan deze bevinding worden gezien als een aanwijzing voor een mogelijke aanzienlijke bias in de besluitvorming over bonafide visumaanvragen op basis van de nationaliteit van de aanvrager.

4.4.2 Disclaimers ten aanzien van de onderzoeksresultaten

Op basis van de gepresenteerde resultaten zijn uitsluitend conclusies te trekken over onevenwichtigheden in de track selectie als onderdeel van de BAO. Over de mate waarin dit mogelijk van invloed is op de uiteindelijke besluitvorming (en daarmee op de daadwerkelijke belangen van aanvragers) is zonder vervolgonderzoek geen onderbouwde uitspraak te doen.

Het is evenmin mogelijk om enige onderbouwde uitspraak te doen over het mogelijke effect van het stopzetten van de BAO track-selectie als onderdeel van het beoordelingsproces van visumaanvragen. Het is weliswaar mogelijk dat eventuele bias in de besluitvorming als gevolg hiervan afneemt, maar het is evengoed mogelijk dat dit juist resulteert in nieuwe onevenwichtigheden en uiteindelijk een hogere bias.

De resultaten dienen wat de opstellers van dit rapport betreft dan ook te worden gezien als een aanleiding voor het reduceren van bias tot acceptabel geachte niveaus, door middel van een zorgvuldige herziening van de BAO (en de

bijbehorende beheersorganisatie). Het is in onze ogen niet verantwoord om de resultaten te interpreteren als argument voor het volledig en permanent stopzetten van enige algoritmische ondersteuning van het besluitvormingsproces over visumaanvragen.

5 Aanbevelingen

5.1 Korte termijn

De in deze paragraaf voorgestelde korte-termijn oplossingen beogen niet een duurzame oplossing voor het reduceren van mogelijke bias in het IOB/KVV, maar dienen slechts als tijdelijke maatregelen om de geconstateerde bias in de BAO aanzienlijk terug te brengen. De voorgestelde maatregelen worden op volgorde van verwachte effectiviteit gepresenteerd. Daarbij geldt dat lager gerangschikte maatregelen slechts zouden moeten worden overwogen wanneer de hoger gerangschikte maatregelen door BZ niet als haalbaar of opportuun worden geacht.

5.1.1 Beëindigen gebruik van profielscore in het BAO

Indien de profielscore niet langer wordt gebruikt voor het bepalen van de trackselectie wordt bij het overgrote deel (>98%) van de aanvragen per definitie de *regular track* toegepast. Alleen aanvragen met hits op aanvrager, referent en/of werkgever komen dan nog in aanmerking voor de *fast* en *intensive track*.

Het belangrijkste voordeel van deze maatregel is dat er het resultaat hiervan zal zijn dat er niet langer sprake is van bias in de BAO.[35] Het belangrijkste nadeel van deze maatregel is dat dit zal resulteren in een aanzienlijke stijging van de benodigde capaciteit voor het beoordelen van visumaanvragen.[36]

5.1.2 Opheffen van intensive track als uitkomst bij afwezigheid hits op aanvrager/referent/werkgever

Indien de *intensive track* niet langer kan worden toegepast op aanvragen zonder hits op aanvrager, referent of werkgever vervalt deze track als optie voor het overgrote deel van de aanvragen. In tegenstelling tot de hier bovengenoemde optie blijft de *fast track* wel bestaan als optie voor deze groep.

Het belangrijkste voordeel van deze maatregel is dat de geconstateerde bias in de BAO vrijwel volledig wordt geëlimineerd zonder gevolgen voor de benodigde capaciteit.[37] Het belangrijkste nadeel van deze maatregel is dat er strikt genomen nog steeds bias in de BAO blijft ten aanzien van de *fast track* selectie. Omdat deze

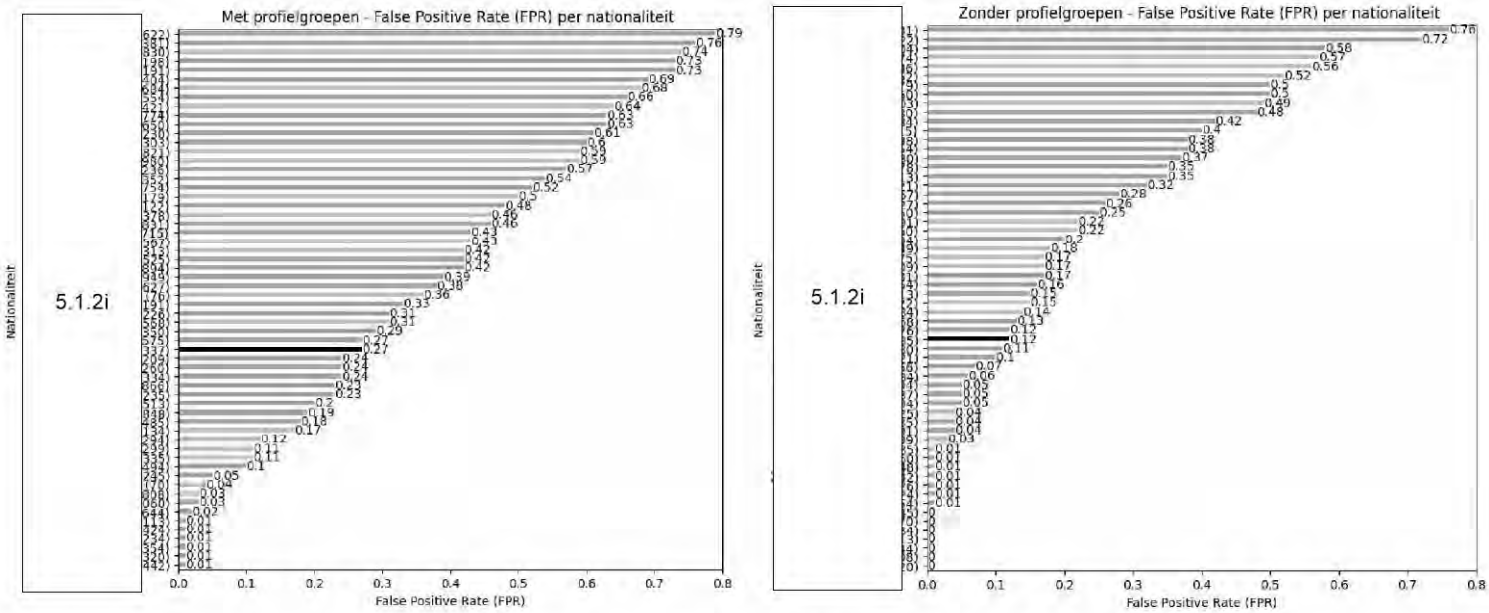
[35] Het is niet uitgesloten dat er enige mate van bias aanwezig is in de databronnen ten behoeve van de hits. Een analyse hiervan valt echter buiten de scope van dit onderzoek en heeft op een dusdanig laag aantal aanvragers betrekking dat de geaggregeerde effecten hiervan op de deel-populaties (bijvoorbeeld op basis van nationaliteit) nagenoeg verwaarloosbaar zullen zijn.
[36] Op basis van 700 duizend visumaanvragen per jaar, in combinatie met de gemiddelde tijdsbesteding per aanvraag in de verschillende tracks, resulteert dit naar verwachting in een benodigde extra capaciteit van ongeveer 20 fte.
[37] De gemiddeld benodigde tijd voor de beoordeling van een aanvraag met *intensive track* is volgens data van BZ vergelijkbaar met (of zelfs iets lager dan) een aanvraag met *regular track*.

bias met name malafide aanvragen betreft is dit type bias echter van een veel lagere orde van belang.[38]

5.1.3 Beëindigen gebruik van risicogroepen (uitsluitend op basis van weigeringspercentage)

Het beëindigen van de risicogroepen welke uitsluitend op basis van weigeringspercentage worden opgesteld zorgt voor een aanzienlijke reductie van bias in de BAO (zie Figuur 16), mede doordat deze groepen een bias-versterkend mechanisme vormen als gevolg van een feedback loop.

Het belangrijkste voordeel van deze maatregel is dat het huidige risico-classificatie systeem in grote lijnen kan worden behouden. Het belangrijkste nadeel is dat, ondanks de aanzienlijke reductie, de resulterende bias nog steeds grotendeels buiten de bandbreedte van de als acceptabel geachte "80% regel" valt.



Figuur 16. False Positie Rate (FPR) per nationaliteit met risico-groepen (links) en zonder risico-groepen (rechts)

5.1.4 Beëindigen gebruik van nationaliteit als variabele

Als minst effectief geachte oplossing kan het stopzetten van het gebruik van het gegeven nationaliteit (en in het verlengde hiervan het gegeven 'land van aanvraag') in de BAO worden overwogen.

Het grootste voordeel van deze optie is dat dit naar verwachting de geconstateerde bias-discrepanties op basis nationaliteit in de BAO aanzienlijk vermindert. Het belangrijkste nadeel is dat we op basis van de beschikbare data de effecten van

[38] Een onjuiste *fast track* aanduiding verhoogt met name de waarschijnlijkheid dat een malafide aanvraag onterecht wordt goedgekeurd (zie paragraaf 3.2)

deze maatregel niet kunnen doorrekenen. Mogelijk ontstaan als gevolg van deze maatregel nieuwe onevenwichtigheden in de risico-classificatie en wordt mogelijk de nauwkeurigheid van de BAO over de gehele linie slechter.

5.2 Middellange termijn

De in deze paragraaf voorgestelde middellange-termijn oplossingen beogen een duurzame reductie van bias in zowel de BAO als het overkoepelende IOB/KVV proces. In tegenstelling tot de voorgestelde kortetermijnmaatregelen zijn de hier voorgestelde oplossingen complementair aan elkaar.

5.2.1 Evaluatie en herziening van de BAO aan de hand van het Impact Assessment Mensenrechten en Algoritmen (IAMA)

Een zorgvuldige evaluatie en herziening van de BAO aan de hand van het IAMA zal naar verwachting de geconstateerde bias problematiek aanzienlijk reduceren en bijdragen aan een effectievere ondersteuning door de BAO van het IOB/KVV proces.

Het IAMA is een instrument voor discussie en besluitvorming voor overheidsorganen. Het instrument maakt een interdisciplinaire dialoog mogelijk door degenen die verantwoordelijk zijn voor de ontwikkeling en/of inzet van een algoritmisch systeem. Toepassing van dit instrument stelt BZ in staat om een uitgebreide *root-cause* analyse naar de mogelijke bronnen van bias in de data, het ontwerp van het algoritme, en als gevolg van de toepassing van het BAO uit te voeren (zie ook Bijlage 2).

Als onderdeel van het IAMA proces kan eveneens worden gekeken naar andere, mogelijk meer effectieve en minder bias-gevoelige alternatieven voor het huidige beslisboom model. Daarnaast kan de verslaglegging van het doorlopen van het IAMA proces gebruikt worden voor verantwoording aan de toezichthouder, of voor toekomstige audits en conformiteitsbeoordelingen (zie ook paragraaf 5.2.2).

Tot slot kan tijdens het doorlopen van het IAMA eveneens invulling worden gegeven aan de hieronder voorgestelde maatregelen.

5.2.2 Verbetering inrichting en beheer van de BAO

De huidige inrichting en het beheer van de BAO voldoen nog niet aan de eisen die in de op handen zijnde Europese AI Verordening aan hoog-risico toepassingen zoals de BAO worden gesteld.[39]

[39] Zie Europese Commissie (2021). De meest recente versie van de AI Verordening voorziet in een hoog-risico kwalificatie van "AI-systemen die bedoeld zijn om bevoegde overheidsinstanties te ondersteunen bij het onderzoek van asielaanvragen, aanvragen voor een visum en aanvragen voor een verblijfsvergunning, evenals gerelateerde klachten met betrekking tot de geschiktheid van de natuurlijke personen die een aanvraag voor een status indienen" (art. 7 sub d in Bijlage III van de AI Verordening).

De aankomende AI Verordening bevat onder meer strenge voorschriften ten aanzien van het te voeren systeem voor risicobeheer (art. 9), data en databeheer (art. 10), technische documentatie (art. 11 en 18), registratie (art. 12 en 19), transparantie en informatieverstrekking aan gebruikers (art. 13), menselijk toezicht (art. 14), nauwkeurigheid, robuustheid en cyberbeveiliging (art. 15), systeem voor kwaliteitsbeheer (art. 17), en conformiteitsbeoordeling (art. 19).

De herziening van het BAO biedt BZ een kans om de aankomende voorschriften van de AI Verordening mee te nemen en tijdig te implementeren als onderdeel van deze herziening.

5.2.3 Toepassing van technische bias-mitigatie methoden

Naast de hierboven genoemde organisatorische maatregelen zijn er diverse technische bias-mitigatie methoden voor handen die als onderdeel van een herziening van de BAO kunnen worden overwogen en geëvalueerd.

Dergelijke methoden worden doorgaans gecategoriseerd als 'pre-processing', 'in-processing', en 'post processing' methoden.[40] Pre-processing methoden zijn gericht op het aanpakken van bias in de gebruikte databronnen (bijvoorbeeld door data-bewerkingen zoals up-sampling). In-processing methoden zorgen dat bias wordt tegengegaan tijdens het trainen van het model (bijvoorbeeld door het integreren van de *False Positive Rate* in de *cost-function* waarop het algoritme wordt geoptimaliseerd). Post-processing methodes, tot slot, beogen een reductie van bias aan door automatische correcties op model uitkomsten.

De toepasbaarheid van deze verschillende methoden hangt onder meer af van het type algoritme waarvoor wordt gekozen, en zal dus nader dienen te worden uitgezocht als onderdeel van het voorgestelde herzieningsproces.

5.2.4 Onderzoeken mogelijkheden van aanvullende data-bronnen ten behoeve van de profiel-score

Voor het bepalen van de profielscore maakt de BAO gebruik van zeven datapunten ten aanzien van de aanvraag, te weten het verblijfsdoel, het land van aanvraag, de nationaliteit, het geslacht, de leeftijdsgroep, burgerlijke staat, en het beroep van de aanvrager.

Op basis van deze beperkte informatie kunnen, zelfs met behulp van de meest geavanceerde methodes, slechts grofmazige risicoprofielen worden gegenereerd. In de aanvraagdossiers is veel meer informatie aanwezig die mogelijk kan bijdragen aan een nauwkeurigere risicoclassificatie.

[40] Zie bijvoorbeeld Mehrabi et al. (2019).

Nader onderzoek naar het omzetten van deze informatie in bruikbare variabelen ter verrijking van de BAO kan in belangrijke mate bijdragen aan het reduceren van bias.

5.2.5 Toepassing van overige bias-mitigatie methoden

Naast het reduceren van bias in het BAO-algoritme kan eveneens worden gekeken naar methoden om mogelijke bias bij beslismedewerkers tegen te gaan, bijvoorbeeld door het (automatisch) uitvlakken/verbergen van niet relevante data. Er zijn verschillende AI-gebaseerde methoden om dit te automatiseren.

Een voorbeeld van mogelijke bronnen van dergelijke bias zijn de naam en pasfoto van de aanvrager. Zo hebben verscheidene studies aangetoond dat er sprake is van een aanzienlijke bias in strafmaat op basis van de naam en uiterlijke kenmerken van verdachten van criminele feiten.[41]

Hoewel in bepaalde situaties de naam en/of foto van een aanvrager relevant kan zijn, kan het effectief zijn om dergelijke informatie verborgen te houden voor medewerkers tot het moment dat deze informatie daadwerkelijk nodig is.

5.2.6 Inrichting van monitoring van bias in besluitvorming als gevolg van track selectie

Een van de belangrijkste vragen die op basis van de huidige beschikbare data niet kan worden beantwoord, is in welke mate bias van de BAO doorwerkt in de uiteindelijke besluitvorming.

Het verdient daarom de aanbeveling om monitoring hiervan mogelijk te maken en in te richten. Twee complementaire mogelijkheden die dit faciliteren zijn:

1. Het dupliceren van een percentage van de aanvragen in verschillende tracks, ten behoeve van kwantitatieve analyse.
2. Het periodiek laten evalueren van afgewezen aanvragen door een comité van ervaren beslismedewerkers, ten behoeve van kwalitatieve analyse.

Met het dupliceren van willekeurige aanvragen wordt bedoeld deze in een aantal kopieën in het systeem te zetten en ieder kopie te voorzien van een van de beschikbare risico-classificaties (tracks). Op die manier worden kopieën van de betreffende aanvragen dus zowel als *fast*, *regular* en *intensive track* aan verschillende medewerkers aangeboden. Zo kan inzicht worden gecreëerd over de mate waarin bias in de track selectie van invloed is op de besluitvorming.[42] De mate

[41] Zie voor onderzoeken in de juridische context onder meer Sah et al. (2015) voor een analyse naar bias op basis van namen, en Hollier (2017) voor een meta-analyse van onderzoeken naar bias op basis van uiterlijk.
[42] Wanneer bijvoorbeeld 1% van alle aanvragen op die manier wordt gedupliceerd resulteert dit bij een jaarlijks volume van 700 duizend aanvragen in 7.000 aanvragen per jaar die kunnen worden onderworpen aan een statistische analyse naar de invloed van track selectie.

waarin weigeringspercentages op deze aanvragen verschillen als gevolg van de track-aanduiding geeft dan een betrouwbaar beeld van de mate waarin bias-discrepanties in de BAO doorwerken op de besluitvorming.

Door het periodiek laten evalueren van afgewezen aanvragen kan meer inzicht worden gecreëerd in mogelijke bronnen van bias bij beslismedewerkers, zowel op basis van informatievoorziening uit de BAO en anderszins. Verkregen inzichten kunnen worden toegepast in verdere reductie van bias door aanpassingen in de BAO en/of het gevoerde beleid.

Bijlage 1: Geraadpleegde bronnen en documenten

Wetenschappelijke literatuur

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). *Machine Bias: There's software used across the country to predict future criminals. And its biased against blacks*, ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Hollier, R. (2017). Physical attractiveness bias in the legal system. *The Law Project*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Narayan, A. (2018). *21 Fairness Definitions and Their Politics*. Conference on Fairness, Accountability, and Transparency

Ruf, B., & Detyniecki, M. (2021). Towards the right kind of fairness in AI. *arXiv preprint arXiv:2102.08453*.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*. Zie de Terms of Reference geschreven door BZ voor dit onderzoek

Sah, S., Robertson, C. T., & Baughman, S. B. (2015). Blinding prosecutors to defendants' race: A policy proposal to reduce unconscious bias in the criminal justice system. *Behavioral Science & Policy*, 1(2), 69-76.

Shah, P. (2015). Against caste in British law: a critical perspective on the caste discrimination provision in the Equality Act 2010. Springer.

Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and human behavior*, *44*(1), 51.

Stevenson, M. T., & Doleac, J. L. (2022). Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440*.

U.S. Equal Employment Opportunity Commission (1979) Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines

Beleidsdocumenten en overige bronnen

Algemene Rekenkamer. (2021). *Aandacht voor algoritmes*. Algemene Rekenkamer.

Autoriteit Persoonsgegevens. (2020). *Toezicht op AI & Algoritmes*. Autoriteit Persoonsgegevens.

BBC. (2020). *Home Office drops 'racist' algorithm from visa decisions*. Opgehaald van https://www.bbc.com/news/technology-53650758

College voor de Rechten van de Mens. (2022). *Aanhoudend foutief gebruik algoritmes door overheden vraagt om bindende discriminatietoets*. Opgehaald van https://www.mensenrechten.nl/actueel/toegelicht/toegelicht/2022/aanhoudend-foutief-gebruik-algoritmes-door-overheden-vraagt-om-bindende-discriminatietoets

College voor de Rechten van de Mens. (2021). *Discriminatie door risicoprofielen: Een mensenrechtelijk toetsingskader*. Utrecht: College voor de Rechten van de Mens.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020, April). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Europese Commissie. (2021). *Voorstel voor een Verordening van het Europees Parlement en de Raad tot vaststelling van geharmoniseerde regels betreffende artificiële intelligentie (Wet op de Artificiële Intelligentie) en tot wijziging van bepaalde wetgevingshandelingen van de Unie*. Brussel: Europese Commissie.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2022a). *Kabinetsreactie op het Rathenau onderzoek 'Algoritmes Afwegen' - 7 oktober 2022.*

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2022b). *Kamerbrief met reactie op Amnesty rapport Xenofobe Machines - 4 juli 2022.*

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2021a). *Impact Assessment Mensenrechten en Algoritmes*.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2021b). *Non-discriminatie by design*.

Ministerie van Buitenlandse Zaken. (2021). *DPIA – Informatie ondersteund Beslissen - Kort Verblijf Visum*.

Muhammad, S. (2022). *The Fairness Handbook*. Gemeente Amsterdam, https://openresearch.amsterdam/image/2022/7/14/fairness_handbook.pdf

Privacy Management Partners. (2022). *Informatie Ondersteund Beslissen - Kort Verblijf Visum (IOB/KVV): Second Opinion*.

Raad van State. (2021). *Voorlichting met betrekking tot het wetsvoorstel Wet gegevensverwerking door samenwerkingsverbanden*. Opgehaald van https://www.raadvanstate.nl/@126518/w16-21-0223-ii-vo/

Bijlage 2: Typen bias

Bias in data	Bias in het modelleren	Bias a.g.v. ingebruikname	Bias in dagelijks gebruik
User interaction bias	Confirmation bias	Presentation bias	Pre-existing bias
Selection bias	Experimenter's bias	Feedback bias	Confirmation bias
Sample bias	Feature bias		Automation bias
Historical bias	Evaluation bias		Group attribution bias
Representation bias	Aggregation bias		Observer bias
Measurement bias	Linking bias		
Omitted variable bias			

Bias in data

User interaction bias
Dit kan voorkomen wanneer de manier waarop een systeem informatie of opties aan gebruikers presenteert, of de manier waarop het reacties van gebruikers vraagt, invloed heeft op de verzamelde gegevens.

Selection bias
Dit gebeurt wanneer de gegevens worden verzameld op een manier die niet representatief is voor de hele populatie, waardoor er sprake is van een vertekende steekproef. Dit kan gebeuren wanneer de gegevens worden verzameld van een zelfgeselecteerde groep of wanneer de gegevens worden verzameld met behulp van een vertekende steekproefmethode.

Participation bias
Dit kan gebeuren wanneer slechts een deel van de populatie bereid of in staat is om deel te nemen, wat leidt tot een vertekende steekproef.

Historical bias
Dit verwijst naar de invloed die eerdere gebeurtenissen of omstandigheden (bestaande bias) hebben op de verzamelde gegevens of op het gebruikte model voor machine learning.

Representation bias
Dit kan gebeuren wanneer de gegevens niet goed gebalanceerd zijn of wanneer bepaalde kenmerken of categorieën niet adequaat worden vertegenwoordigd in de gegevens.

Measurement bias
Dit kan gebeuren wanneer het meet- of dataverzamelingsproces niet nauwkeurig of consistent is, of wanneer er slechte procedures worden gebruikt, waardoor er sprake is van vertekende gegevens.

Omitted variable bias
Dit kan gebeuren wanneer belangrijke variabelen niet in het model zijn opgenomen, waardoor er sprake is van vertekende of incorrecte resultaten.

Bias in het modelleren

Confirmation bias
Dit gebeurt wanneer mensen ontwerpkeuzes maken voor het model die overeenstemmen met hun bestaande overtuigingen (tijdens modellering), terwijl ze informatie die deze overtuigingen tegenspreekt, negeren of bagatelliseren.

Experimenter's bias
Dit kan gebeuren wanneer de experimentator onbewust of bewust invloed uitoefent op de uitkomst van de studie, wat leidt tot vertekende of misleidende resultaten.

Feature bias
Dit gebeurt wanneer de kenmerken die worden gebruikt om het model te trainen, niet relevant of belangrijk zijn voor het voorspellen van de doelvariabele. Dit kan leiden tot slechte prestaties bij de taak, omdat het model beslissingen kan leren nemen op basis van irrelevante of misleidende kenmerken.

Evaluation bias
Dit kan gebeuren wanneer het evaluatieproces niet goed is ontworpen of wanneer de gegevens die worden gebruikt om het model te evalueren, niet representatief zijn voor de populatie.

Aggregation bias
Dit kan gebeuren wanneer de gegevens niet goed worden geaggregeerd of wanneer bepaalde kenmerken of trends in de gegevens niet adequaat worden vertegenwoordigd in de geaggregeerde data.

Linking bias
Dit kan voorkomen wanneer de gegevens niet goed gekoppeld zijn of wanneer bepaalde relaties of patronen in de gegevens niet adequaat worden vertegenwoordigd in de gekoppelde gegevens.

Bias a.g.v. ingebruikname

Presentation bias
Dit gebeurt wanneer de percepties of beslissingen van mensen worden beïnvloed door de manier waarop informatie wordt gepresenteerd of geframed. Dit kan leiden tot bevooroordeelde gegevens als mensen waarschijnlijker zijn om op informatie te reageren op een bepaalde manier, afhankelijk van hoe het wordt gepresenteerd.

Feedback bias
Dit gebeurt wanneer de uitkomst van het model (direct of indirect) als toekomstige invoer voor het model wordt gebruikt (wat resulteert in feedbacklussen).

Bias in dagelijks gebruik

Pre-existing bias
Dit gebeurt wanneer de mensen die gebruik maken van het systeem vooringenomenheid introduceren door hun eigen vooroordelen, voorkeuren of aannames.

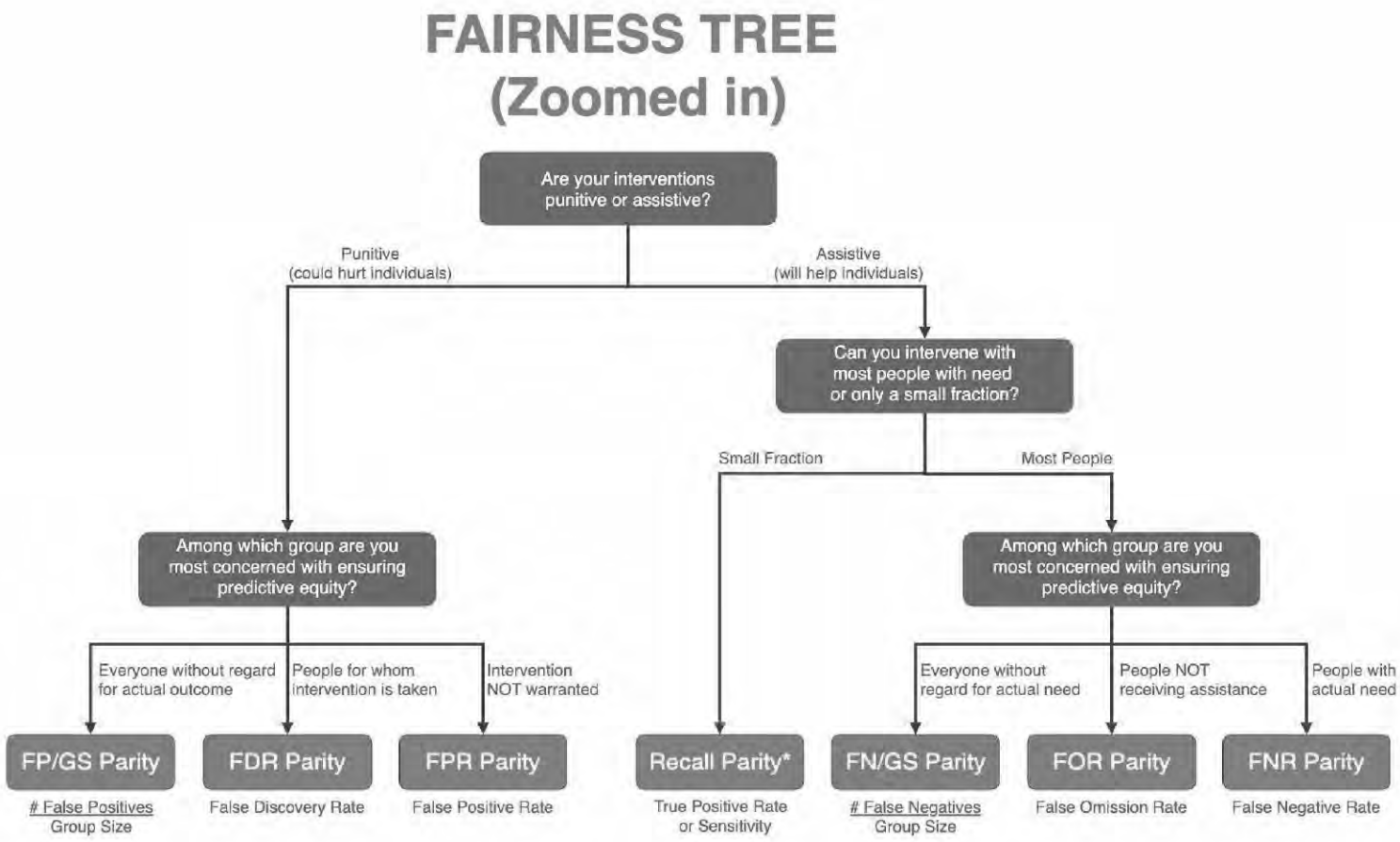
Confirmation bias
Dit gebeurt wanneer mensen op zoek gaan naar of meer aandacht besteden aan informatie die hun bestaande overtuigingen bevestigt, terwijl ze informatie die die overtuigingen tegenspreekt negeren of afzwakken.

Automation bias
Dit verwijst naar de neiging van mensen om te veel te vertrouwen op geautomatiseerde systemen, zoals machine learning-modellen, en om meer gewicht te geven aan de beslissingen of aanbevelingen die door deze systemen worden gedaan, zelfs wanneer ze mogelijk onjuist of gebrekkig zijn.

Group attribution bias
Dit verwijst naar de neiging van mensen om leden van hun eigen groep meer te begunstigen, prefereren of positievere attitudes te tonen ten opzichte van leden van andere groepe (in-group bias); of de neiging van mensen om leden van andere groepen minder te waarderen, of meer negatieve attitudes te tonen ten opzichte van leden van hun eigen groep (out-group bias).

Observer bias
Dit verwijst naar de neiging van mensen om bij observaties of interpretaties van gebeurtenissen beïnvloed te worden door hun persoonlijke overtuigingen, attitudes of verwachtingen.

Bijlage 3: Aequitas Fairness Kompas



Figuur 17. Relevante deel van het fairness kompas (the fairness tree) zoals ontwikkeld door de makers van biasanalyse softwarepakket Aequitas[43]

[43] Carnegie Mellon University, Data Science and Public Policy, Aequitas Fairness tree: http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Bijlage 4: Performance analyse BAO op basis van nationaliteit aanvrager

Performance metrics (hogere scores zijn beter)

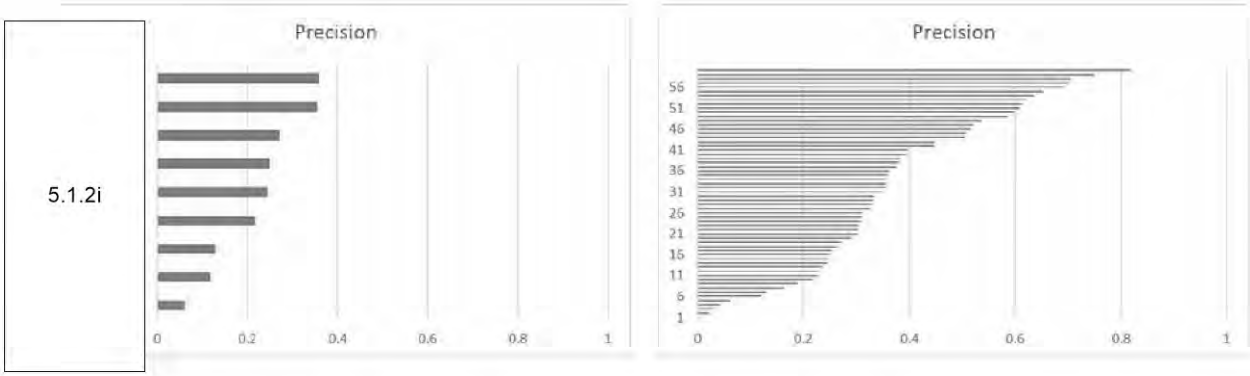
De hieronder genoemde performance metrics geven op vier verschillende manieren een beeld van hoe goed de track selectie van de BAO overeenkomt met de uiteindelijke besluitvorming over visumaanvragen.

Met betrekking tot de top-9 landen (qua volume) scoort de BAO alleen zowel goed als consistent op de *Negative Predictive Value*. Met betrekking tot alle landen scoort de BAO op geen enkele metric zowel goed als consistent.

Met betrekking tot de Intensive track

Precision
Praktische definitie: Ratio van aanvragen met intensive track dat inderdaad wordt afgewezen.

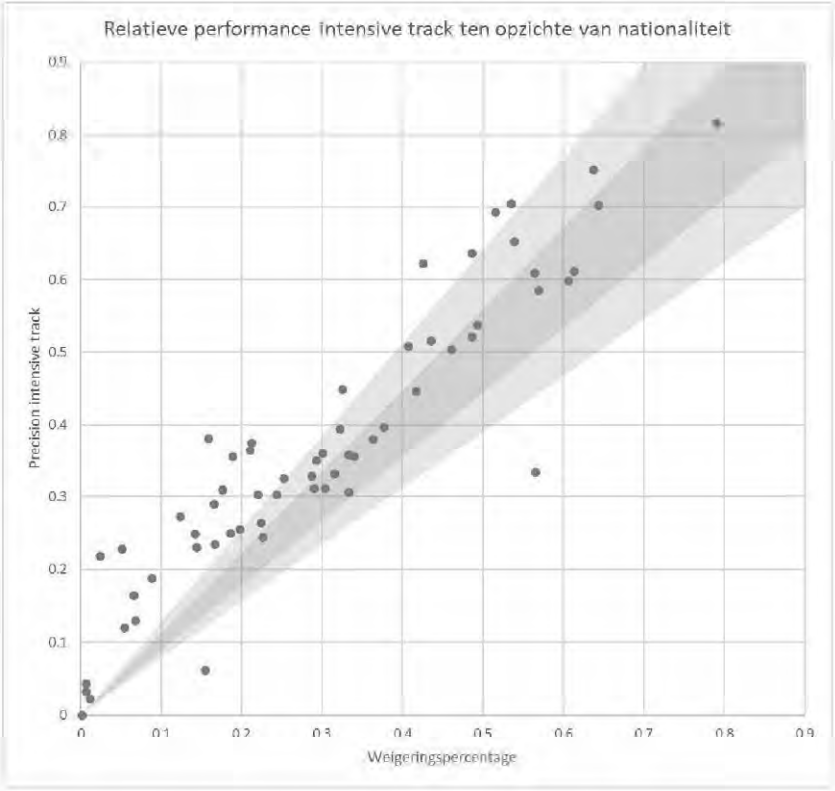
5.1.2i



- Top 9 landen: Gemiddeld 22% van aanvragen met *intensive track* wordt daadwerkelijk afgewezen. Variëteit **groot** (range 6-36%).
- Alle landen: Gemiddeld 37% van aanvragen met *intensive track* wordt daadwerkelijk afgewezen. Variëteit **zeer groot** (range 0-82%).

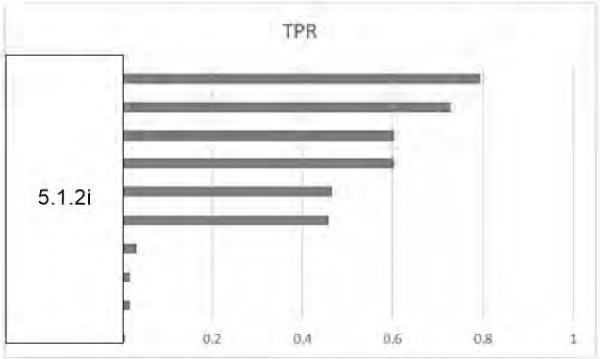
Ten aanzien van de *Precision* is een belangrijke kanttekening op zijn plaats. Aangezien de BAO naast nationaliteit verschillende andere kenmerken meeneemt in de risicoprofilering is de verwachting dat de *Precision* voor een bepaalde nationaliteit substantieel hoger is dan het weigeringspercentage voor die nationaliteit. Immers, indien dit niet het geval is dan is nationaliteit effectief het enige doorslaggevende kenmerk (anders gezegd, wanneer je binnen een bepaalde nationaliteit een aantal aanvragen willekeurig aan de *intensive track* zou toewijzen verwacht je een *Precision* gelijk aan het weigeringspercentage voor die nationaliteit).

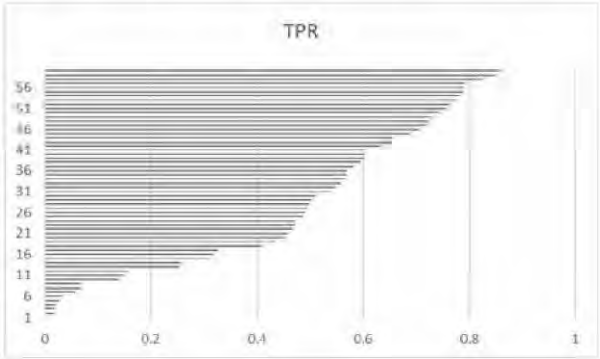
Wanneer we de *Precision* afzetten tegen het weigeringspercentage per nationaliteit ontstaat het volgende beeld:



Uit bovenstaande grafiek kan worden geconcludeerd dat voor een groot aantal nationaliteiten de BAO op basis van *Precision* niet (donkere zone) of nauwelijks (lichte zone) beter presteert dan kans. Dit suggereert dat voor deze landen het gegeven nationaliteit effectief de enige ter zake doende risicofactor is ten behoeve van de *intensive track*.

True Positive Rate (recall)
Praktische definitie: Ratio afgewezen aanvragen dat inderdaad *intensive track* had.

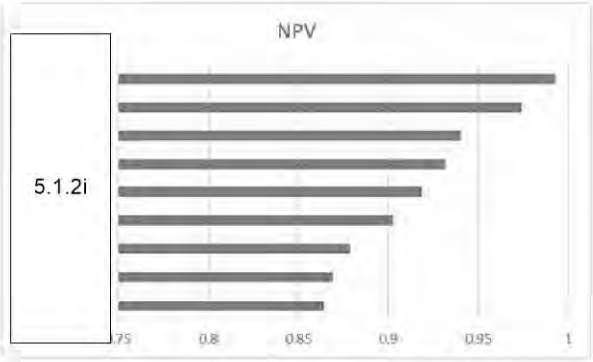


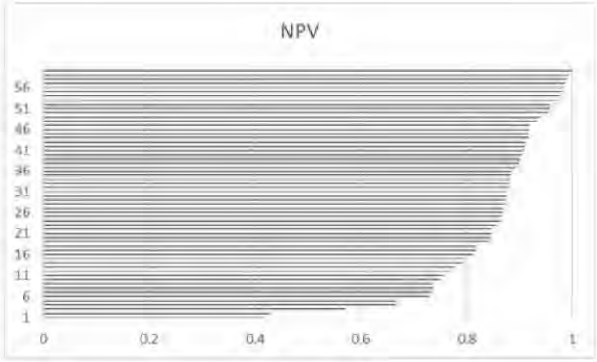


- Top 9 landen: BAO kwalificeert gemiddeld 41% van de afgewezen aanvragen inderdaad als *intensive track*. Variëteit **zeer groot** (range 2-79%)
- Alle landen: BAO kwalificeert gemiddeld 48% van de afgewezen aanvragen inderdaad als *intensive track*. Variëteit **zeer groot** (range 0-86%)

Met betrekking tot fast track

Negative Predictive Value
Praktische definitie: Ratio van aanvragen met *fast track* dat inderdaad wordt goedgekeurd.

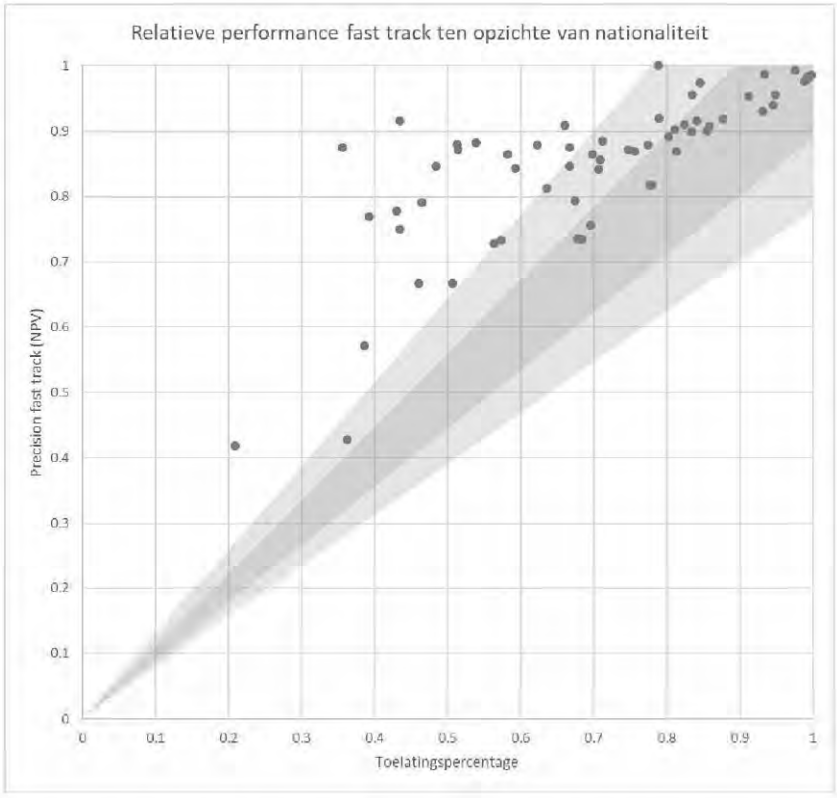




- Top 9 landen: Gemiddeld 92% van aanvragen met *fast track* wordt daadwerkelijk goedgekeurd. Variëteit **beperkt** (range 86-99%).
- Alle landen: Gemiddeld 85% van aanvragen met *fast track* wordt daadwerkelijk goedgekeurd. Variëteit **zeer groot** (range 42-100%).

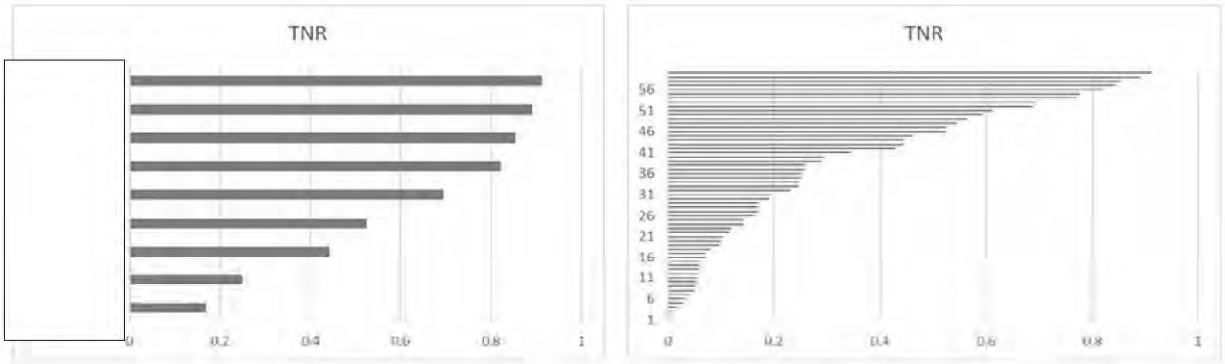
Ten aanzien van de *Negative Predictive Value* (NPV) is dezelfde kanttekening op zijn plaats als voor de *Precision* (zie 1.1.1.). Aangezien de BAO naast nationaliteit verschillende andere kenmerken meeneemt in de risicoprofilering is de verwachting dat de *NPV* voor een bepaalde nationaliteit substantieel hoger is dan het toelatingspercentage voor die nationaliteit. Immers, indien dit niet het geval is dan is nationaliteit effectief het enige doorslaggevende kenmerk (anders gezegd, wanneer je binnen een bepaalde nationaliteit een aantal aanvragen willekeurig aan de *fast track* zou toewijzen verwacht je een NPV gelijk aan het toelatingspercentage voor die nationaliteit).

De *NPV* afgezet tegen het toelatingspercentage per nationaliteit geeft het volgende beeld:



Uit bovenstaande grafiek kan worden geconcludeerd dat voor een redelijk groot aantal nationaliteiten de BAO op basis van NPV niet (donkere zone) of nauwelijks (lichte zone) beter presteert dan kans. Dit suggereert dat voor deze landen het gegeven nationaliteit effectief de enige ter zake doende risicofactor is ten behoeve van de *fast track*.

True Negative Rate (specificity)
Praktische definitie: Ratio goedgekeurde aanvragen dat inderdaad *fast track* had.



- Top 9 landen: BAO kwalificeert gemiddeld 62% van de goedgekeurde aanvragen inderdaad als *fast track*. Variëteit **zeer groot** (range 17-91%)
- Alle landen: BAO kwalificeert gemiddeld 30% van de goedgekeurde aanvragen inderdaad als *fast track*. Variëteit **zeer groot** (range 0-91%)

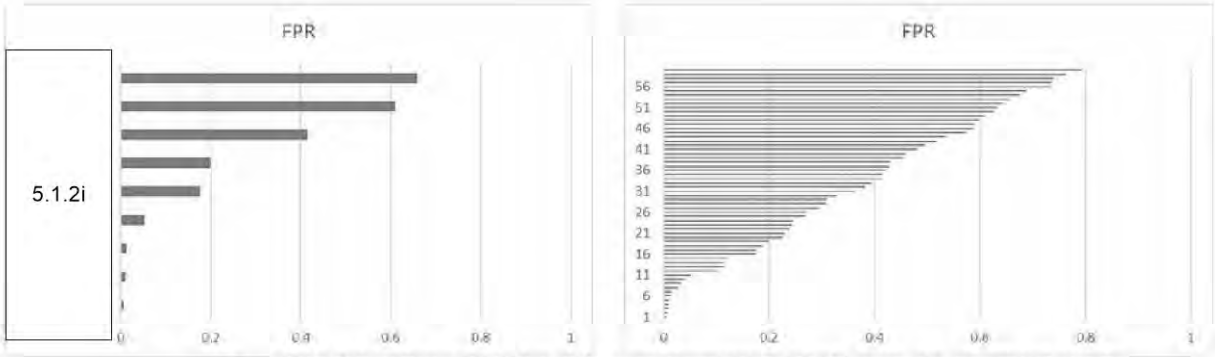
Error metrics (lagere scores zijn beter)

De hieronder genoemde error metrics geven op vier verschillende manieren een beeld van de mate waarin de track selectie van de BAO niet overeenkomt met de uiteindelijke besluitvorming over visumaanvragen.

Noch met betrekking tot de top-9 landen (qua volume), noch met betrekking tot alle landen, scoort de BAO op enige metric zowel goed én consistent.

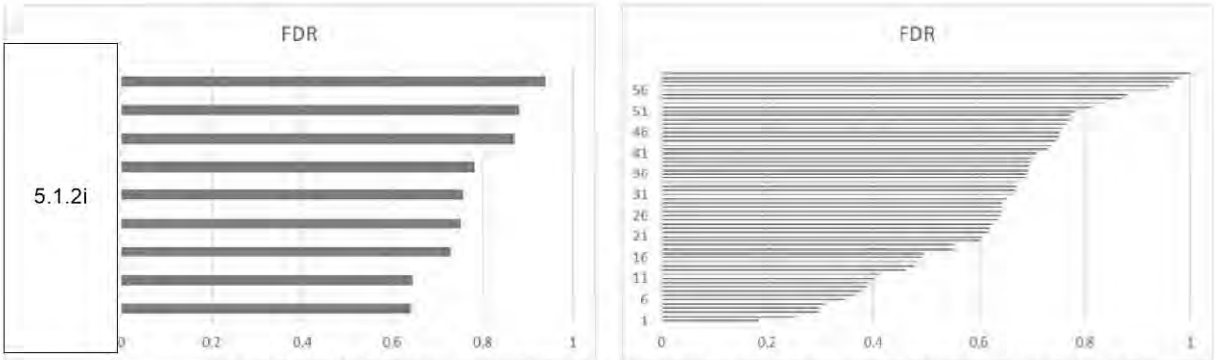
Met betrekking tot Intensive track

False Positive Rate
Praktische definitie: Ratio goedkeurde aanvragen dat desondanks *intensive track* had.



- Top 9 landen: BAO kwalificeert gemiddeld 24% van de goedgekeurde aanvragen desondanks als *intensive track*. Variëteit **zeer groot** (range 0-66%)
- Alle landen: BAO kwalificeert gemiddeld 35% van de goedgekeurde aanvragen desondanks als *intensive track*. Variëteit **zeer groot** (range 0-79%)

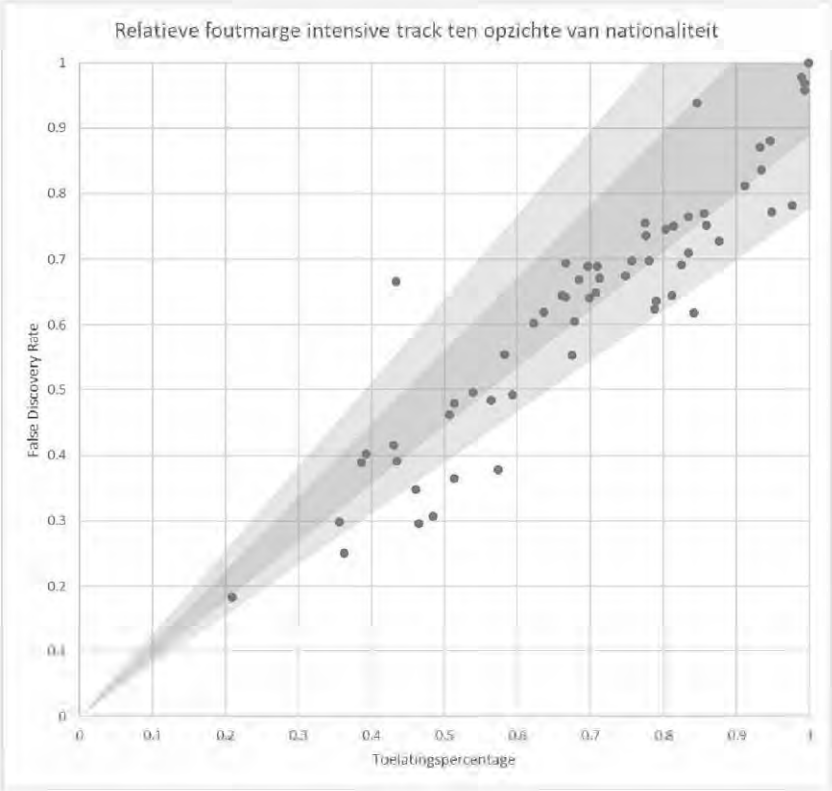
False Discovery Rate
Praktische definitie: Ratio van aanvragen met *intensive track* dat desondanks is goedgekeurd.



- Top 9 landen: Gemiddeld 77% van aanvragen met *intensive track* wordt desondanks goedgekeurd. Variëteit **beperkt** (range 64-94%).
- Alle landen: Gemiddeld 63% van aanvragen met *intensive track* wordt desondanks goedgekeurd. Variëteit **zeer groot** (range 18-100%).

Ten aanzien van de *False Discovery Rate* (feitelijk de foutmarge van de *intensive track*) is dezelfde kanttekening op zijn plaats als voor de *Precision* en *NPV*. Aangezien de BAO naast nationaliteit verschillende andere kenmerken meeneemt in de risicoprofilering is de verwachting dat de *FDR* voor een bepaalde nationaliteit substantieel lager is dan het toelatingspercentage voor die nationaliteit. Immers, indien dit niet het geval is dan is nationaliteit effectief het enige doorslaggevende kenmerk (anders gezegd, wanneer je binnen een bepaalde nationaliteit een aantal aanvragen willekeurig aan de intensive track zou toewijzen verwacht je een *FDR* gelijk aan het toelatingspercentage voor die nationaliteit).

De *FDR* afgezet tegen het toelatingspercentage per nationaliteit geeft het volgende beeld:



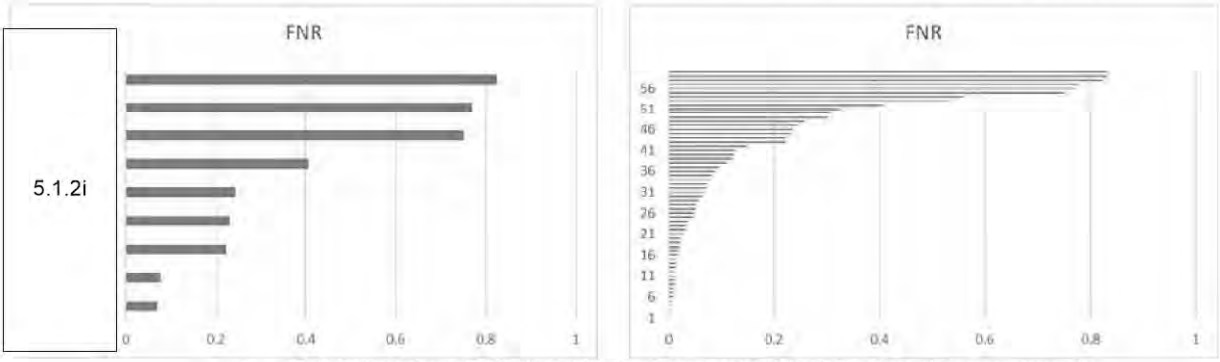
Uit bovenstaande grafiek kan worden geconcludeerd dat voor een redelijk groot aantal nationaliteiten de BAO op basis van *FDR* niet (donkere zone) of nauwelijks (lichte zone) beter presteert dan op basis van kans mag worden verwacht. Dit suggereert dat voor deze landen het gegeven nationaliteit effectief de enige ter zake doende risicofactor is ten behoeve van de *intensive track*.

Met betrekking tot fast track

False Negative Rate

Praktische definitie: Ratio afgewezen aanvragen dat desondanks *fast track* had.

5.1.2i

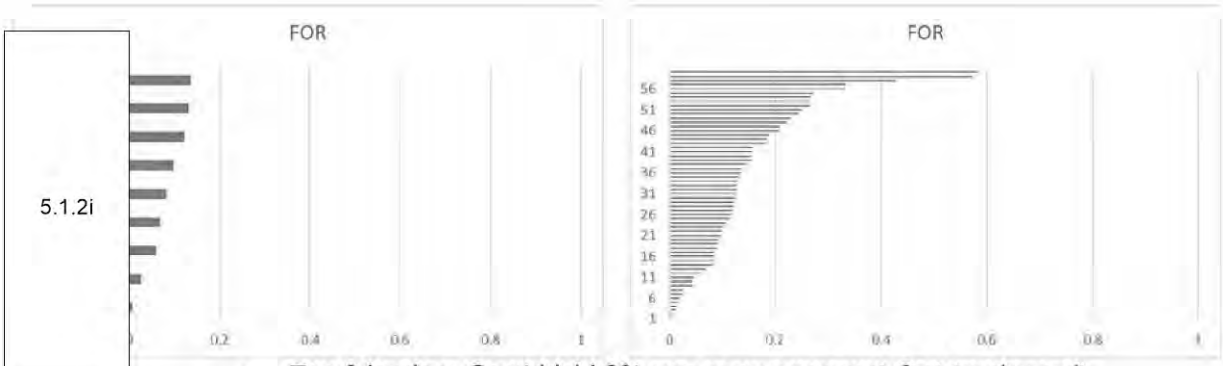


- Top 9 landen: BAO kwalificeert gemiddeld 40% van de geweigerde aanvragen desondanks als *fast track*. Variëteit **zeer groot** (range 7-82%)
- Alle landen: BAO kwalificeert gemiddeld 17% van de geweigerde aanvragen desondanks *als fast track*. Variëteit **zeer groot** (range 0-83%)

False Ommission Rate

Praktische definitie: Ratio van aanvragen met *fast track* dat desondanks is geweigerd.

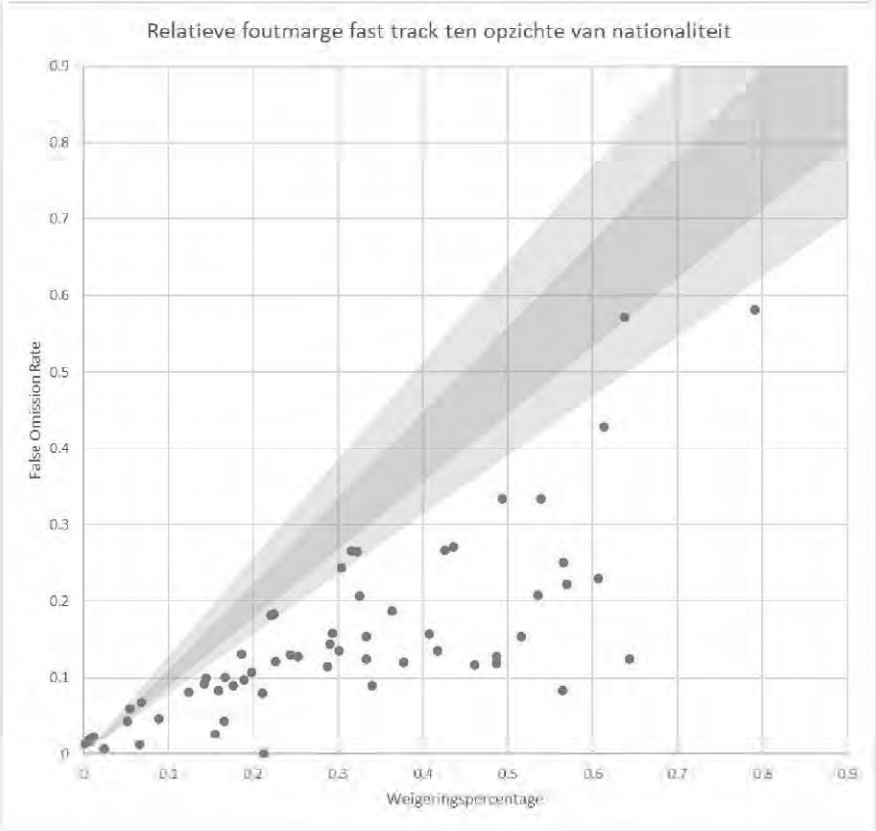
5.1.2i



- Top 9 landen: Gemiddeld 8% van aanvragen met *fast track* wordt desondanks geweigerd. Variëteit **groot** (range 0-14%).
- Alle landen: Gemiddeld 15% van aanvragen met *fast track* wordt desondanks geweigerd. Variëteit **zeer groot** (range 0-58%).

Ten aanzien van de *False Ommission Rate* (feitelijk de foutmarge van de *fast track*) is dezelfde kanttekening op zijn plaats als voor de *Precision*, *NPV*, en *FDR*. Aangezien de BAO naast nationaliteit verschillende andere kenmerken meeneemt in de risicoprofilering is de verwachting dat de *FOR* voor een bepaalde nationaliteit substantieel lager is dan het weigeringspercentage voor die nationaliteit. Immers, indien dit niet het geval is dan is nationaliteit effectief het enige doorslaggevende kenmerk (anders gezegd, wanneer je binnen een bepaalde nationaliteit een aantal aanvragen willekeurig aan de *fast track* zou toewijzen verwacht je een *FOR* gelijk aan het weigeringspercentage voor die nationaliteit).

De *FOR* afgezet tegen het weigeringspercentage per nationaliteit geeft het volgende beeld:



Uit bovenstaande grafiek kan worden geconcludeerd dat voor veruit de meeste nationaliteiten de BAO op basis van *FOR* substantieel beter presteert dan kans (buiten de donkere en lichte zone). Dit impliceert dat voor deze landen het gegeven nationaliteit niet de enige ter zake doende risicofactor is ten behoeve van de *fast track*.

Bijlage 5: Biasanalyse vanuit fast track perspectief (False Negative Rate)

In hoofdstuk 4.2 zijn de resultaten van de kwantitatieve biasanalyse vanuit het perspectief van de *intensive* track (False Positive Rate) beschreven. Het gevolg van een bias tussen FPR's is een mogelijk nadeel voor aanvragers behorende tot een bepaalde demografische groep.

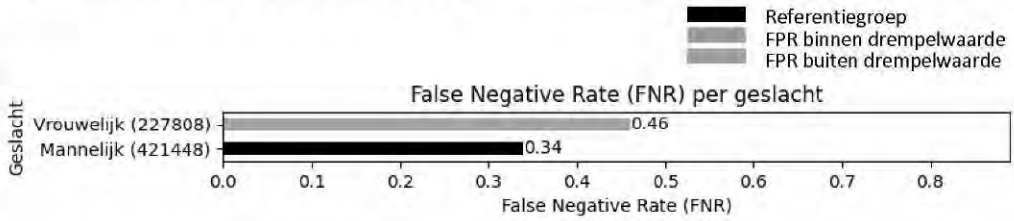
Hier worden de resultaten van de biasanalyse vanuit het perspectief van de *fast* track beschreven. Het gevolg van een bias tussen FNR's is een mogelijk onthouden voordeel voor aanvragers behorende tot een bepaalde demografische groep.

Geconstateerde bias op basis van geslacht

False Negative Rate (FNR)

De FNR verwijst naar het percentage goedgekeurde aanvragen dat door de BAO niet als fast wordt gecategoriseerd. Een aanvrager wordt hierdoor mogelijk onthouden van een voordeel.

Aan hand van dezelfde drempelwaarde die is gebruikt voor de FPR bias analyse kan op basis van Figuur 18 geconstateerd worden dat de False Negative Rate (FNR) voor vrouwelijke aanvragers aanzienlijk hoger is dan de FNR voor mannelijke aanvragers.



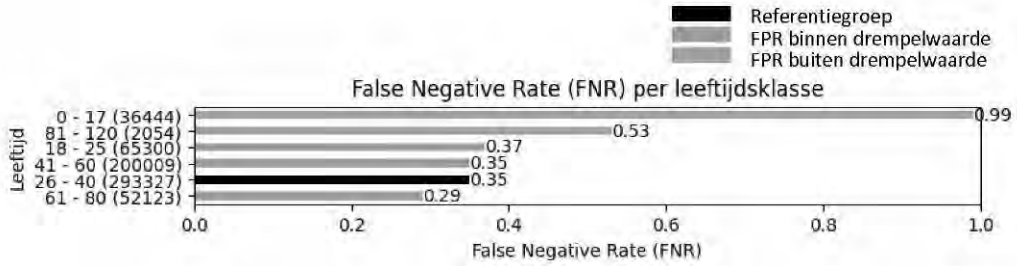
Figuur 18. False Negative Rate (FNR) per geslacht van de aanvrager. Een groene en rode staaf betekent dat de FPR van de betreffende groep respectievelijk binnen en buiten de acceptabele range van de biasdrempelwaarde valt ten opzichte van de referentiegroep (zwart).

Deze discrepantie suggereert dat er mogelijk sprake is van ongelijke behandeling op basis van geslacht in de BAO, namelijk dat bonafide vrouwelijke aanvrager minder vaak als *fast* worden gecategoriseerd dan bonafide mannelijke aanvragers.

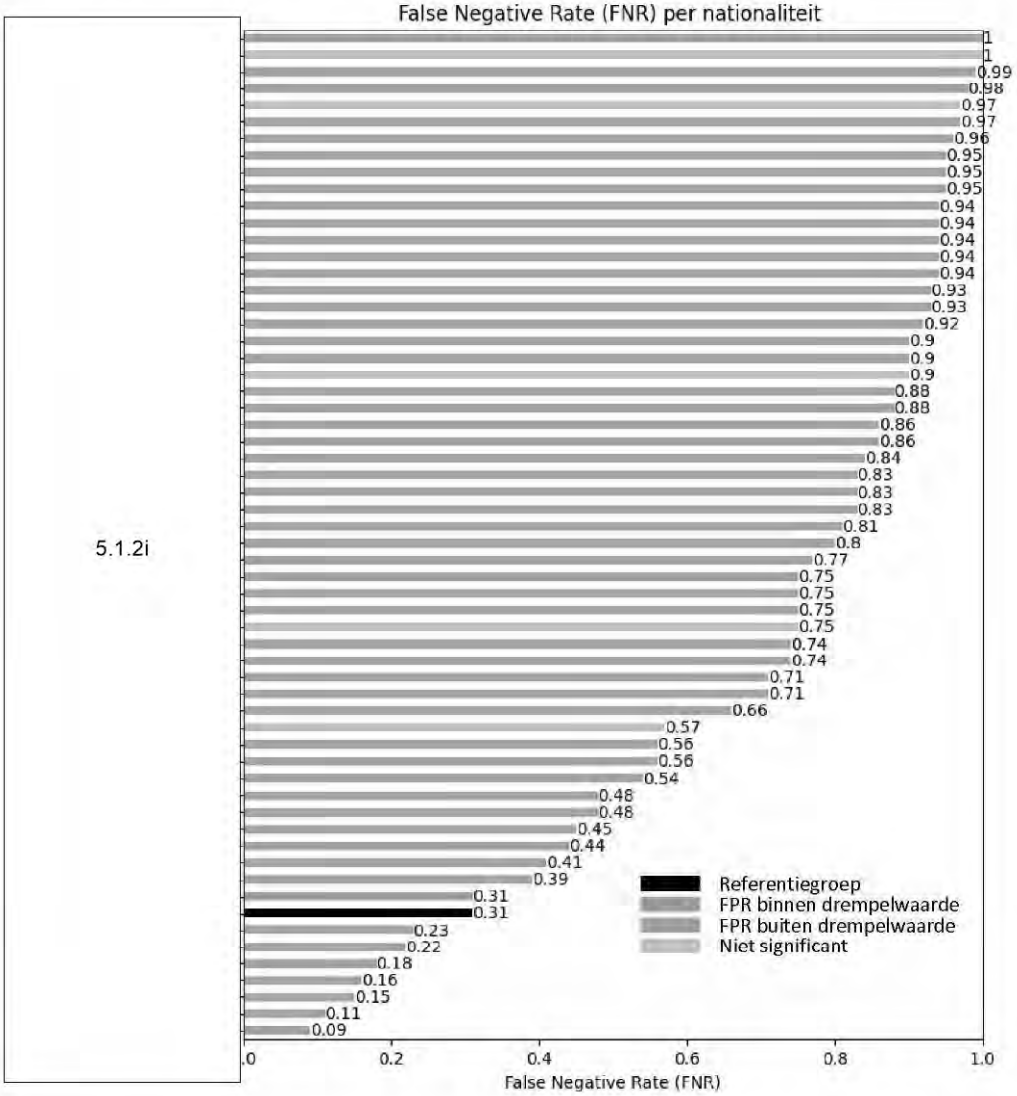
Geconstateerde bias op basis van leeftijd

Ook kan worden geconstateerd dat voor de leeftijdsklassen 0-17 en 81-120 de FNR aanzienlijk hoger is dan de leeftijd referentiegroep (Figuur 19). Voor de leeftijdsklasse 0-17 is dit gemakkelijk te verklaren: deze leeftijdsklasse wordt uitgesloten van profilering en wordt daardoor niet door de BAO gecategoriseerd als *intensive* of *fast*.

Dit resultaat suggereert dat er mogelijk sprake is van een onthouden voordeel voor de leeftijdsklasse 81-120: bonafide aanvragers met leeftijd tussen de 81-120 worden minder vaak als *fast* gecategoriseerd dan bonafide aanvragers met een jongere leeftijd.



Figuur 19. False Negative Rate (FNR) per leeftijdsklasse van de aanvrager. Een groene en rode staaf betekent dat de FPR van de betreffende groep respectievelijk binnen en buiten de acceptabele range van de biasdrempelwaarde valt ten opzichte van de referentiegroep (zwart).



Figuur 20. False Negative Rate (FNR) per nationaliteit van de aanvrager (selectie nationaliteiten)

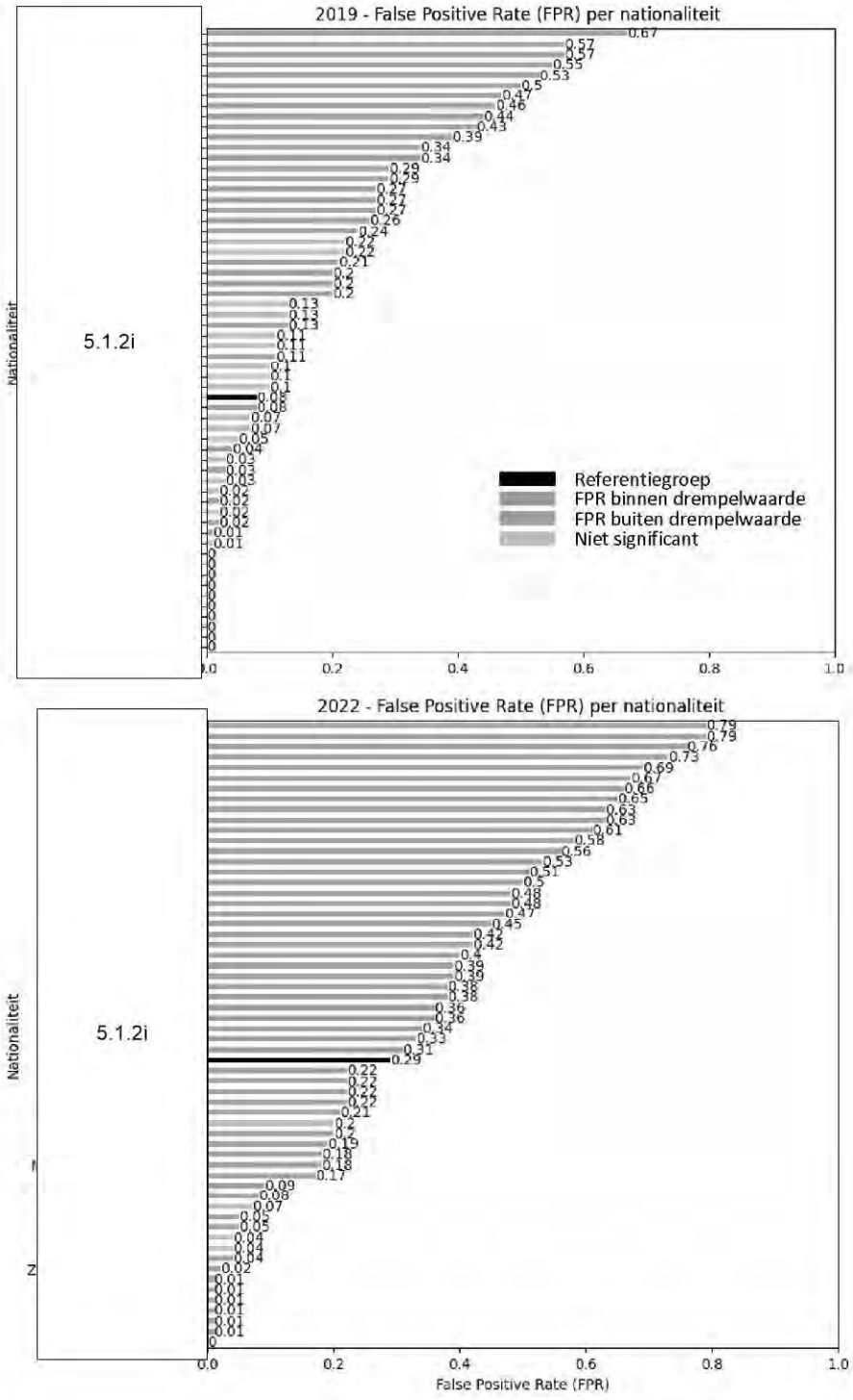
Geconstateerde bias op basis van nationaliteit

Er zijn zeer grote verschillen geconstateert in FNR tussen nationaliteiten (zie Figuur 20). Ten opzichte van de referentiegroep (Indiase) vertonen verreweg de meeste nationaliteiten een aanzienlijk hogere of lagere FNR. Ongeacht wat als referentiegroep zou zijn gekozen (Indiase is gekozen als referentiegroep omdat het overeenkomt met het gewogen gemiddelde FNR) zouden deze grote discrepanties aanwezig zijn.

Deze resultaten suggereren dat er sprake is van een mogelijke benadeling op basis van nationaliteit veroorzaakt door de BAO: bonafide aanvragen met een nationaliteit voornamelijk afkomstig uit Afrika en het Midden Oosten worden minder vaak als *fast* gecategoriseerd dan bonafide aanvragen met een andere nationaliteit (zie Figuur 21).

Figuur 21. False Negative Rate (FNR) per nationaliteit van de aanvrager (selectie landen)

Bijlage 6: Biasanalyse voor en na de Covid-19 pandemie



Figuur 22. False Positive Rate (FPR) per nationaliteit van de aanvrager in 2019 (boven) en 2022 (onder) (selectie nationaliteiten)

Bijlage 7: Reactie op feedback BZ

Het Rijks ICT Gilde (RIG) heeft een biastoets uitgevoerd voor het ministerie van Buitenlandse Zaken (BZ). Deze biastoets heeft betrekking op het Buitenlandse Zaken Analyse Omgeving (BAO) systeem dat wordt gebruikt ter ondersteuning van het Kort Verblijf Visa (KVV) aanvraagproces. Een eerste conceptversie van de bevindingen is door het RIG opgeleverd in maart 2023 en als reactie hierop heeft BZ drie documenten met feedback aangeleverd geschreven door 5.1.2i en de ontwerpers van de BAO. In dit document geeft het RIG een reactie op de geleverde feedback.

Overzicht van de geleverde feedback

Allereerst zijn we blij om te horen dat onze bevindingen en adviezen BZ tot nadenken heeft aangezet dat leidt tot het reflecteren over de BAO aanpak. Hierbij ondersteunen en onderschrijven we volledig het plan van BZ om de volgende punten aan te pakken:

- De fundamenten van de BAO/IOB zorgvuldiger beschrijven
- De inrichting en toepassing van de BAO continu monitoren en valideren
- Scherper zijn op de onderbouwing van de BAO/IOB en lacunes in de documentatie opvullen
- BAO robuuster maken door het betrekken van meer en relevantere databronnen, m.a.w. de kwaliteit en validatie van de profielen borgen en risico op *'underfitting'*[44] verminderen
- De woordvoering op de BAO herzien wat betreft de misleidende boodschap dat het profiel slechts 10% meeweegt in de bepaling van de uiteindelijke track
- Herziening van de regels op basis waarvan het algoritme de profielen genereert, inclusief onderbouwing van deze regels
- Toewerken naar een sterkere data-uitwisseling tussen BZ en de IND en deze feedback meenemen in het herzien van de regels en de databronnen van de BAO

Zoals beschreven in ons rapport, delen we het inzicht dat het gebruik van een tool als de BAO de beoordeling van een visumaanvraag kan ondersteunen en het proces objectiever en efficiënter kan maken. Het kan daarbij, indien goed ingericht, zelfs juist bias in het beslisproces verminderen.

[44] Bij *'underfitting'* is er sprake van een model met een te simplistische benadering van de werkelijkheid met onjuiste uitkomsten als gevolg

Echter tonen de bevindingen van ons onderzoek sterke indicaties dat de BAO in zijn huidige vorm een benadelend effect kan hebben op bepaalde groepen aanvragers. Daarom concluderen wij dat, in aanvulling op bovengenoemde punten, de mogelijk nadelige effecten van de BAO verder onderzocht moeten worden en vervolgens op meerdere vlakken verbeterd moet worden.

Hieronder gaan we punt voor punt dieper in op de door BZ geleverde feedback.

1. **Het is nooit de insteek van de BAO geweest om een intensive track gelijk te trekken met een malafide aanvraag/aanvrager**

We begrijpen dat risicoclassificatie nooit de originele insteek van de BAO is geweest. Echter werkt dit waarschijnlijk in de praktijk anders uit omdat de intensive track wel degelijk informatie geeft over de kans op een malafide/risicovolle aanvraag en de fast track over de kans op een bonafide/kansvolle aanvraag. Dit is ten slotte direct in lijn met de manier waarop de BAO wordt getraind: kenmerken van aanvragen die historisch tot ongewenst dan wel gewenst gedrag hebben geleid, zijn de directe input voor risico- dan wel kansprofielen.[45] De verhoogde kans op weigering voor intensive track wordt ook gereflecteerd in de weigeringspercentages per track (Figuur 14). Dit is uiteraard verwacht en ook gewenst, maar heeft ook als implicatie dat een deel van bonafide aanvragen met een intensive track kan worden gezien als mogelijk malafide aanvragen. Dit komt door een *priming* effect van de trackuitkomst en de resulterende bias hiervan is wederom veelvuldig wetenschappelijk aangetoond.[46]

Dit onderschrijft het belang om te garanderen dat aanvragen alleen met een goed onderbouwde reden in de intensive track belanden en dat dit een zo hoog mogelijk aantal daadwerkelijk malafide aanvragen betreft (en andersom ook voor de fast track). Deze redenering wordt ondersteund op basis van de wijze waarop BZ de verhouding tussen trackuitkomsten en het besluit over de visumaanvraag duidt: "Mocht blijken dat bijvoorbeeld het advies vanuit de BAO afwijkt van de beslissing om het visum wel of niet af te geven, dan zal deze feedback meegenomen in de doorontwikkeling van informatie ondersteund beslissen".[47]

---

[45] Zo wordt er in de BAO-documentatie (Richtlijnen Informatie Ondersteunend Beslissen, FAQ Informatie Ondersteunend Beslissen, DPIA BAO) herhaaldelijk gesproken over een positief (fast track) of negatief (intensive track) advies vanuit de BAO. Daarnaast wordt meermaals gesproken over de fast en intensive track als "kansprofiel en risicoprofiel", "meer of minder betrouwbare aanvragen", "positieve en negatieve aanvragers", "positieve en negatieve referenten", "laag of verhoogd risico", "weinig of geen risico"/"verhoogd risico", en "alleen positieve informatie bekend in relatie tot de aanvraag"/"negatieve informatie bekend is in relatie tot de aanvraag".

[46] Zie onder meer Skeem *et al.* (2020), Stevenson, & Doleac (2022) en De-Arteaga *et al.* (2020)

[47] Zie 'Richtlijnen Informatie Ondersteunend Beslissen'

2. Intensive track uitkomsten leiden niet tot een 1-op-1 negatief besluit: de BAO leidt daarom niet tot een misleidende risicoclassificering (en introduceert dus ook geen bias op het niveau van de medewerker)

We zijn het er mee eens dat intensive track uitkomsten inderdaad niet leiden tot een 1-op-1 negatief besluit (en trekken deze conclusie ook niet in het rapport). Dit is te zien aan de weigeringspercentages op aanvragen met een intensive track (dat de afgelopen jaren zelf is afgenomen, zie ook Figuur 14). Echter gebruikt BZ dit gegeven om de onjuiste conclusie te trekken dat de BAO daarom geen misleidende risicoclassificering kan meegeven aan beslismedewerkers (en niet resulteert in bias).

Een eerste reden waarom dit een onjuiste conclusie is, is dat er geen 1-op-1 correlatie nodig is om te spreken van een mogelijk misleidende risicoclassificering. Misleidende risicoclassificering kan op een kleinere schaal plaatsvinden welke moeilijker te detecteren valt, maar wel resulteert in bias op groepsniveau. Immers, zelf wanneer een misleidende risicoclassificering theoretisch gezien niet 1-op-1 maar 1-op-20 resulteert in een negatief besluit, is er nog steeds sprake van bias.

Een tweede reden is het feit dat het aantal aanvragen in de verschillende tracks de afgelopen jaren sterk is veranderd. Zo is het percentage aanvragen in de intensive track gegroeid van 3% in 2018 naar 33% in 2022 (Figuur 15). Het is aannemelijk dat die 33% intensive track aanvragen in 2022 meer bonafide aanvragen bevatten dan de 3% in 2018 wat logischerwijze resulteert in een lager weigeringspercentage voor 2022.[48] Het is niet mogelijk om een verschil in bonafide aanvragen te verifiëren, maar dit voorbeeld laat zien dat de conclusie van BZ te kort door de bocht is. Daarbij komt dat op basis van weigeringspercentages het onmogelijk is om conclusies te trekken over bias.

Belangrijk om hier te noemen is dat een misleidende risicoclassificering nooit 100% voorkomen kan worden. Enige mate van bias is een gegeven in ieder proces, ondersteund door een algoritme of niet, en wanneer deze bias beperkt is hoeft dit niet per definitie tot problemen te leiden. Met het oog op onterechte bias is het van belang dat er geen onacceptabele verschillen bestaan in de mate van mogelijk misleidende risicoclassificering tussen demografische groepen. Echter hebben we zeer grote verschillen gevonden tussen nationaliteiten in percentages KVV-

[48] Immers waren de regels van het beslisboomalgoritme in 2018 waarschijnlijk strenger wat resulteerde in een selectievere groep van 3% intensive track aanvragen

aanvragen met een intensive track die zijn toegewezen (False Positive Rate, FPR), m.a.w. de mate waarin bonafide aanvragen worden gelabeld als hoger risico. De spreiding in FPR's (van 0.01 tot 0.79) beslaan bijna het theoretische maximale spectrum. Dit heeft als implicatie dat de kans aanwezig is dat voor bonafide aanvragers met bepaalde nationaliteiten visumaanvragen (veel) vaker onterecht worden geweigerd. In welke mate hier sprake van is, moeten toekomstige experimenten uitwijzen, zoals ook genoemd in de aanbevelingen van ons rapport. Echter laten recente onderzoeken al zien dat de impact van (onjuiste) risicoprofilering op besluitvorming zelfs bij ervaren professionals, zoals rechters, aanzienlijk is.[49] De kans is zeer klein dat de BAO hierop een uitzondering vormt.

3. Een intensive track is niet nadelig voor een aanvrager

Het is onjuist te stellen dat de trackuitkomst uitsluitend iets zegt over de tijd die nodig is om een aanvraag te behandelen. Het is daarom ook niet mogelijk om risico-indicatie en tijdsindicatie van een aanvraag los van elkaar te zien, ongeacht dat dit de originele basis en het doel van de BAO was. Het algoritme voorspelt ten slotte niet hoeveel tijd er nodig is voor een aanvraag: behandel- en doorlooptijden zijn niet deel van de inputvariabelen van het systeem. De stelling dat de trackuitkomst informatie geeft over de intensiteit van een aanvraag zit puur in de communicatie naar beslismedewerkers. We vermoeden dat dit mogelijk onvoldoende gecommuniceerd wordt naar beslismedewerkers en betwijfelen of communicatie überhaupt voldoende is om mogelijke risico's terug op *priming* te dringen. Immers krijgen beslismedewerkers onder andere de volgende informatie:

- "De BAO geeft een indicatie van het risico voor de openbare orde en veiligheid en de verblijfsintentie van de aanvrager."[50]
- "LET OP: Aanvraag valt in groep met verhoogd risico op misbruik visumprocedure"[51]

Daarnaast heeft BZ eerder erkend dat "het nadeel [van deze profielen] is dat er vooroordelen kunnen ontstaan waardoor een bepaalde groep als risico of kans wordt gezien terwijl dat niet terecht hoeft te zijn"[52] en "dat er altijd een risico bestaat dat het gebruikte algoritme de uiteindelijke beslissing van de medewerker beïnvloedt".[53]

---

[49] Zie onder meer Skeem *et al.* (2020), Stevenson, & Doleac (2022) en De-Arteaga *et al.* (2020)
[50] Zie 'FAQ IOB voor beslismedewerkers v0.6'
[51] Dit is het bericht een beslismedewerkers te zien krijgen bij een aanvraag in de intensive track
[52] Zie 'FAQ Informatie Ondersteunend Beslissen v0.6'
[53] Zie Terms of Reference van dit huidige onderzoek

Medewerkers die aanvragen krijgen met een intensive label zullen de aanvraag intensiever bekijken en mogelijk meer tijd besteden om te kijken of het risicolabel gerechtvaardigd is. Echter zoals ook met IDI gecommuniceerd op 24 februari, hebben we in een gesprek met een projectleider CSO gehoord dat intensive aanvragen nauwelijks langer worden beoordeeld dan regular track aanvragen. Als een intensive track aanduiding niet leidt tot intensiever onderzoek (m.a.w. meer tijd voor interviews, het opvragen van documenten, nauwkeuriger lezen etc.) dan is mogelijk het enige gevolg van een intensive track uitkomst een negatief *priming* effect. In dat geval is er sprake van een nog hoger risico op bias: immers kan intensiever onderzoek dan niet eens het tegendeel bewijzen.

We willen benadrukken dat we niet stellen dat een intensive track per definitie leidt tot benadeling of tot "een organiserend principe om tot een weigering te komen", maar dat de functie van de BAO als risicovoorspeller kan leiden tot **mogelijk onterechte benadeling** van een deel van aanvragen, en dat de mate waarin dit gebeurt verschillend kan uitpakken voor verschillende nationaliteiten. Een onvoldoende onderbouwd proces om te komen tot profielen kan leiden tot grote en onterechte verschillen in trackuitkomsten tussen demografische groepen. De bevindingen van de statistische toets geven een sterke indicatie dat hier inderdaad sprake is. Dit vraagt om maatregelen. Zoals ook in onze aanbevelingen genoemd, moet meer onderzoek plaatsvinden naar hoe beslismedewerkers omgaan met intensive track aanvragen, onder andere wat er wordt gedaan in het geval van twijfelgevallen en wat het effect is van verschillende trackuitkomsten bij eenzelfde aanvraag.

4. De waargenomen hoge bias-discrepanties voor sommige landen is te rechtvaardigen op basis van historische gegevens

We erkennen dat er terecht grote verschillen kunnen bestaan in trackuitkomsten tussen nationaliteiten. Dit kan worden verklaard door historische gegevens zoals asielinstroom, illegale immigratie, nationale veiligheid, mensensmokkel en meer. Het is daarom van belang dat sommige aanvragen zorgvuldiger worden beoordeeld en daarom een intensive track krijgen. We erkennen daarom ook het gerechtvaardigde belang van het gebruik van het gegeven 'nationaliteit' in de BAO.

Het probleem zit echter niet in grote verschillen in trackuitkomsten of weigeringspercentages. Het probleem zit in de grote verschillen tussen nationaliteiten in percentages bonafide aanvragen die een intensive track hebben gekregen (FPR). Deze informatie vertelt ons namelijk iets over de blootstelling van bonafide aanvragen aan een intensive track en daarmee een onterecht verhoogd risico op weigering. Het is van belang dat deze percentages niet te veel verschillen

tussen nationaliteiten, zodat sommige nationaliteiten niet onterecht veel vaker worden blootgesteld aan mogelijk misleidende risicoclassificering dan andere landen. Op dat moment is er namelijk sprake van onbehoorlijke algoritmische bias. Onze bevindingen hebben aangetoond dat sommige landen buitenproportioneel vaker worden blootgesteld aan mogelijk misleidende risicoclassificering.

Deze verschillen worden dus niet direct veroorzaakt door hoge weigeringspercentages en hoge aantallen intensive track aanvragen. Grote verschillen in weigeringspercentages en/of intensive track percentages en kleine verschillen in FPRs spreekt elkaar niet tegen. De observatie van BZ dat deze bias-discrepanties niet leiden tot en-masse weigering is daarom geen reden om de conclusie af te wijzen. Immers kan er nog steeds sprake zijn van onterechte weigering voor bepaalde nationaliteiten. Dat het algoritme niet blind gevolgd wordt is duidelijk, maar dit is geen bewijs dat er geen sprake is van bias. Doordat we de beslissing van de medewerker als *'ground truth'* hebben genomen, worden in de onderzoeksmethode alle onterecht geweigerde aanvragen als terecht beschouwd. De implicatie is dat de geconstateerde bias-discrepanties waarschijnlijk zelfs lager zijn dan de feitelijke discrepanties (zie ook feedback punt 5).

5. **De keuze om de beslissing van de medewerker als proxy voor de *ground truth* te beschouwen is problematisch en introduceert een andere bron van bias waardoor gevonden bias niet direct aan BAO toegeschreven kan worden**

In de geleverde feedback wordt gesteld dat door het nemen van de beslissing van de medewerker als proxy voor de *ground truth* eventuele gevonden bias niet direct aan de BAO toegeschreven kan worden.

Ten eerste erkennen we dat het inderdaad geen ideale situatie is om de beslissing van de medewerker als proxy voor de *ground truth* te beschouwen, maar het is (gegeven het ontbreken van betere alternatieven) wel een aannemelijke en verdedigbare aanname voor een statistische biastoets. Ten slotte beschikt de medewerker over meer relevante bronnen van informatie dan de BAO en daarmee vormt de beslissing van de medewerker naar verwachting een significant betere reflectie van de aard van de aanvraag dan de track uitkomst van de BAO zou kunnen suggereren. Daar komt bij dat, door te stellen dat deze (per definitie onvolmaakte) proxy niet bruikbaar is, BZ de BAO feitelijk immuniseert voor enige bias-toetsing, aangezien de daadwerkelijke *ground truth* fundamenteel onkenbaar is (zie paragraaf 3.1.2 van het rapport). De zeer grote geconstateerde verschillen in FPR's tussen nationaliteiten kunnen bovendien niet exclusief worden verklaard door de gekozen proxy voor de *ground truth*: daarvoor zijn de verschillen te groot.

Zoals door 5.1.2 wordt gesuggereerd, zou je inderdaad zelfs als de BAO bias-loos was geweest met deze *ground truth* methode wat verschillen waarnemen. Dit wordt veroorzaakt doordat met deze *ground truth* per definitie alle onterechte intensive track weigeringen door beslismedewerkers als terecht worden beschouwd (de beslissing wordt immers altijd gezien als correct). Dat wil zeggen dat medewerker-bias daarmee buiten scope wordt gelaten en dat de FPR's mogelijk lager zijn dan ze in werkelijkheid zouden zijn. Bovendien zouden, in het hypothetische geval van een bias-loze BAO, de geconstateerde FPR-discrepanties naar verwachting ordes van grootte lager zijn.

Daar komt bij dat de zeer grote variëteit in verschillen tussen trackuitkomsten en uiteindelijke besluiten, vanuit een bias perspectief, hoe dan ook zeer problematisch is. Immer zijn er dan drie mogelijke verklaringen:

1. De uiteindelijke besluitvorming is relatief unbiased en er is sprake van zeer hoge bias in de trackuitkomsten (in onze ogen is dit het meest waarschijnlijke scenario)
2. De trackuitkomsten zijn relatief unbiased en er is sprake van zeer hoge positieve bias in de uiteindelijke besluitvorming (waarbij de beslismedewerker intensive aanvragen uit kwetsbare landen vaker toewijst wat leidt tot hoge FPR's)
3. Er is sprake van (zeer) hoge bias in zowel de trackuitkomsten als in de besluitvorming

Alle drie de verklaringen resulteren naar alle waarschijnlijkheid in een onacceptabele situatie.

6. RIG heeft niet voldaan aan de *Terms of Reference*: er zijn binnenbochten genomen die de kwaliteit en betrouwbaarheid ondergraven

Bij de start van onze samenwerking hebben we gezamenlijk de scope van de opdracht bepaald: een statistische toets met een focus op het algoritme en met name de profilering door het beslisboomalgortime. Hierbij is het meenemen van de beslissingen van medewerkers noodzakelijk om bias in trackuitkomsten aan te kunnen tonen. Het is hierbij niet mogelijk om de precieze oorzaken van de aangetoonde bias onder de streep aan te wijzen en het concreet meten van eventuele automation bias (en ook andere vormen van bias op medewerker niveau, hierna voor het gemak allemaal automation bias genoemd) is hierbij buiten beschouwing gelaten, in tegenstelling tot wat door BZ wordt gesteld. Ten slotte heb je om biases op niveau van de beslismedewerker vast te stellen naast data science

ook andere expertises nodig, zoals bijvoorbeeld gedragswetenschappers. Om andere vormen van bias te meten is ook uitgebreider onderzoek nodig, o.a. naar de methode van verzameling van data van alle ketenpartners, wat definitief buiten de scope van dit onderzoek valt.

Bovendien wordt in de Terms of Reference niet geïnstrueerd om automation bias expliciet buiten beschouwing te laten: er wordt slechts beweerd dat automation bias al wordt gemonitord op basis van weigeringspercentages per track. We hebben hierboven al beargumenteerd dat deze monitoring niet voldoende is om automation bias uit te sluiten. Het is aannemelijk dat automation bias de belangrijkste en voornaamste vorm van bias is binnen de BAO en daarom wordt automation bias ook vaak in het rapport genoemd: we kunnen ten slotte wel iets zeggen over het risico op bias, zonder de bias zelf te meten (zoals bijvoorbeeld de aanwezigheid van profielgroepen puur op basis van weigeringspercentage die een feedbackloop kan veroorzaken).

De feedback over het spreken met beslismedewerkers heeft ons verbaasd. In onze aanpak hebben we aangegeven in gesprek te gaan maar tot nu toe hebben we geen toestemming gekregen van BZ[redacted] Bovendien zijn gesprekken met een klein aantal beslismedewerkers hier niet voldoende om sterke conclusies te trekken over het effect van de trackuitkomsten op het beslisproces. Hiervoor hebben we meerdere aanbevelingen gedaan in ons rapport over mogelijke experimenten en onderzoeken waar beslismedewerkers betrokken moeten worden.

Tot slot sluiten wij ons niet aan bij de suggestie van het nemen van "binnenbochten". Wij hebben in ons rapport zorgvuldig gecommuniceerd over de gekozen onderzoeksmethode en de daarbij behorende beperkingen, en geven rekenschap over deze beperkingen in de formuleringen van onze conclusies.

7. De observering dat de beslissing van medewerkers niet in lijn is met de trackuitkomst is juist een goed teken: er is geen sprake van automation bias

Als trackuitkomsten niets zeggen over de aard/risico van de aanvraag dan volgt er geen logische redenering waarom de BAO door BZ wordt gebruikt. Ten slotte is het voor BZ niet wenselijk dat er structureel vaak intensiever wordt gekeken naar bonafide aanvragen. Dit kost onnodig veel kostbare tijd van beslismedewerkers. Met dit perspectief zouden bonafide aanvragen zoveel mogelijk in de fast track moeten worden ingedeeld. Het is dus wenselijk dat de trackuitkomst wel iets zegt over de aanvraag, zoals het algoritme ook wordt geïnstrueerd om risicoprofielen en

kansprofielen op te stellen op basis van ongewenst en gewenst gedrag (dit hoeft geen 1-op-1 relatie te zijn; dit is zelfs onwenselijk door risico's op *'overfitting'*[54]). Hierbij hoeft de medewerker absoluut niet blind te varen op de trackuitkomst: goede monitoring op FPR's, training en communicatiezorgt ervoor dat o.a. automation bias wordt ingeperkt.

Goede kwalitatieve profielen zorgen voor betere relaties tussen trackuitkomst en aard/risico van de aanvraag. In de huidige vorm van de BAO is het aantal hits (dus de harde data over ongewenst gedrag) waarop de profielen wordt gemaakt zeer laag en toch worden daarop alle profielen gebaseerd. De gevonden hoge FPR's kunnen wel degelijk het resultaat zijn van kwalitatief lage profielen als gevolg van *'underfitting'*. Om risico op bias terug te dringen moet de kwaliteit van profielen continu worden gevalideerd om een goede relatie tussen trackuitkomst en aard/risico van de aanvraag te garanderen en *'underfitting'* te voorkomen.

8. De BAO verschilt fundamenteel van het verboden visumalgoritme in het Verenigd Koninkrijk[55]

De ontwerpers van de BAO stellen dat de werkwijze uit het VK fundamenteel anders is dan de BAO-werkwijze van BZ. Wij willen niet de indruk wekken dat het BAO identiek is met het door het VK eerder gebruikte algoritme, maar vinden het wel relevant om te benoemen dat dit op hoofdlijnen vergelijkbaar is (inclusief de track-selectie als basis voor de intensiteit/rigiditeit van het onderzoek). Behalve deze observatie doen wij geen verdere vergelijkende analyse tussen de BAO en het VK-algoritme in het rapport. De reden waarom we het VK-algoritme in het rapport noemen en de gelijkenissen met de BAO benadrukken, is dat we BZ bewust wilden maken over het functioneren van de BAO en specifiek de communicatie en rechtvaardiging hiervan richting de politiek en de buitenwereld.

Dat gezegd hebbende, willen wij op basis van de feedback graag nog een observatie delen. Een van de door de BAO-ontwikkelaars genoemde argumenten is dat in tegenstelling tot het VK-algoritme het weigeringspercentage in de BAO niet mee wordt genomen. Binnen de BAO wordt echter het weigeringspercentage wel degelijk meegenomen. Sterker nog, er zijn zelfs risicoprofielen die puur alleen op basis van weigeringspercentage worden gegenereerd. Dit lijkt wel degelijk overeenkomsten te

---

[54] Bij *'overfitting'* is er sprake van een model dat te nauwkeurig past bij de inputgegevens en daardoor niet goed presteert op nieuwe, ongeziene gegevens

[55] BBC. (2020). *Home Office drops 'racist' algorithm from visa decisions*. Opgehaald van https://www.bbc.com/news/technology-53650758

tonen met de werkwijze uit het VK waarbij nationaliteiten op een zwarte lijst werden geplaats aan de hand van het aantal weigeringen.

Hoewel het VK-algoritme zelf niet openbaar is gemaakt, wordt in alle berichtgeving gesteld dat *"Applications made by people holding 'suspect' nationalities received a higher risk score"*.[56] Dit is feitelijk wat er ook in de BAO gebeurt (direct of indirect): (mede) het kenmerk nationaliteit leidt via weigeringspercentages tot een hogere risicoscore.

Een paar woorden over de aanbevelingen van 5.1.2e en de ontwerpers van de BAO

5.1.2e doet de aanbeveling om te onderzoeken of de BAO voorspelt "in hoeverre er een indicatie is voor aanvullend onderzoek" of dat de BAO voorspelt "in hoeverre er een indicatie is voor afwijzing op basis van profilering". Zoals hierboven in dit document meermaals beschreven, worden de profielen expliciet opgesteld op basis van kenmerken (waaronder nationaliteit) van aanvragen met ongewenst dan wel gewenst gedrag in combinatie met weigeringspercentages. Hieruit kan worden opgemaakt dat de profielen informatie geven over de hoogte van het risico van de aanvraag welke kan worden geïnterpreteerd voor een indicatie voor afwijzing. Uiteraard moet dit worden geverifieerd door aanvullend onderzoek. Naar onze mening kunnen de twee door 5.1.2e voorgestelde vragen moeilijk los van elkaar worden gezien en dit slaat dan ook de plank mis in de context van een onderzoek naar algoritmische bias.

Wij onderschrijven echter wel een aantal van de aanbevelingen van 5.1.2e en de ontwerpers van de BAO, welke grotendeels overeenkomen met de door ons beschreven aanbevelingen in het rapport, zoals:

- Onderzoek in hoeverre trackuitkomsten überhaupt worden gezien en meegewogen in het beslisproces van een beslismedewerker
- Onderzoek in hoeverre een intensive track uitkomst de kans op weigering verhoogd door middel van het dupliceren van een aanvragen in verschillende tracks ten behoeve van kwantitatieve analyse
- Onderzoek de mate waarin de trackuitkomsten invloed hebben op de duur van de behandeling en de mate waarin de duur van behandeling invloed heeft op de beslissing van de aanvraag
- Voer rootcause-analyses uit voor bepaalde profielen/nationaliteiten en onderzoek de invloed en rol van politiek bestuurlijke informatie. Hierbij kan

[56] iNews. (2020) *Home Office to scrap algorithm which secretly assigns 'risk score' to some nationalities by design.* Opgehaald van https://inews.co.uk/news/home-office-algorithm-priti-patel-risk-score-visa-570830

mogelijk ook gebruik kan worden gemaakt van de methode "*counterfactual fairness*" zoals gesuggereerd door de ontwerpers van de BAO, alhoewel dit in de context van de BAO zeer moeilijk te operationaliseren is[57]

- Betrek ten alle tijden een ervaren jurist bij het doen van juridische uitspraken
- Blijf de data die de BAO gebruik zorgvuldig monitoren en let op verdelingen en verschuivingen over de tijd.

De suggestie van de ontwerpers van de BAO om *demographic parity* als biasmetriek te kiezen hebben wij vroeg in het proces uitgesloten.[58] Zoals BZ heeft aangegeven zijn er objectieve gronden voor verschillen in weigeringspercentages tussen groepen. Op basis daarvan is *demographic parity* geen zinvolle biasmetriek voor de BAO. De biasmetriek *False Positive Rate parity* is omwille van de in het rapport gegeven redenen de meest geschikte metriek in de context van de BAO.[59]

Conclusie

We zien graag dat BZ de aanbevelingen gedaan door ons 5.1.2e en de ontwerpers van de BAO zal gaan opvolgen. Op basis van de door BZ geleverde feedback documenten, opgesteld door 5.1.2e en de ontwerpers van de BAO, zien wij echter geen redenen om van onze originele conclusies af te wijken. We hopen dat we door middel van dit document onze standpunten duidelijk genoeg hebben onderbouwd en horen graag van BZ waar eventueel nog onduidelijkheden over bestaan.

[57] *Counterfactual fairness* is een methode om causale verbanden in het model te ontdekken door variabelen willekeurig te veranderen en het effect daarvan te bestuderen
[58] *Demographic parity* is een bias/fairness metriek welke stelt dat iedereen gelijke kansen moet hebben om een positief resultaat te krijgen, ongeacht bijvoorbeeld nationaliteit of geslacht.
[59] Zie de *Aequitas fairness tree* hier: Carnegie Mellon University, Data Science and Public Policy, Aequitas Fairness tree: http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/ en zie ook de wetenschappelijke publicatie over Aequitas hier: Saleiro, P. *et al.* (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577.*

Dit is een uitgave van:
Rijksorganisatie voor Ontwikkeling, Digitalisering en Innovatie
Postbus 20011 | 2500 EA Den Haag

Meer weten?
Kijk op www.rijksorganisatieodi.nl