Adviescommissie
Analytics

Advies Datafundamenten & Analytics

Oktober 2024

Inhoudsopgave

1. Introductie

1.1 Goede bedoelingen versus ongewenste effecten

Bekend is dat het voor een deel van de burgers in onze samenleving moeilijk is om mee te komen. Bijvoorbeeld omdat ze laaggeletterd of digitaal niet zo vaardig zijn, of omdat ze een verstandelijke beperking hebben. Maar misschien ook wanneer ze in moeilijke omstandigheden (komen te) zitten, bijvoorbeeld als gevolg van schulden of door een ingrijpende levensgebeurtenis, zoals een scheiding of het overlijden van een naaste. De inzet van analytics zou behulpzaam kunnen zijn om dergelijke groepen in het vizier te krijgen en hen proactief en gericht te helpen. Deze inzet geeft echter ook vragen rond het 'labelen' van mensen en mogelijke ongewenste bemoeizucht vanuit de overheid. Goede bedoelingen zijn niet voldoende om ongewenste effecten te voorkomen. Er zijn binnen de organisatieonderdelen van de overheid diverse initiatieven die hieraan raken en waar een blik van buiten heel waardevol is. Het idee is dat deze initiatieven baat kunnen hebben bij advisering op dit vlak.

1.2 Behandeling adviesvragen

Een adviesvraag wordt opgehangen aan de kapstok 'Willen, Mogen, Kunnen'. Een adviesvraag aan de Adviescommissie Analytics (verder: de Adviescommissie[1]) is bedoeld om helderheid en richting te krijgen over de wenselijkheid ('Willen') van een bepaalde analytics toepassing, daar waar dilemma's plaatsvinden en lastig een (ethische) afweging gemaakt kan worden. Vragen met betrekking tot 'Mogen' en 'Kunnen' dienen organisatieonderdelen eerst zelf te beantwoorden. Het gaat hier namelijk om vragen die alleen beantwoord kunnen worden wanneer men beschikt over gedetailleerde inhoudelijke kennis (over wet- en regelgeving, kaders, ICT, data-science, uitvoering, etc.).

Een adviesvraag zal gaan over de maatschappelijke en morele wenselijkheid (van een bepaald handelen). Het advies gaat over het afwegen van belangen vanuit een externe blik. Een advies zal dan ook eerder richtinggevend dan kaderbepalend zijn.

1.3 Managementsamenvatting advies

De Adviescommissie heeft gereflecteerd op de hoofdvraag: 'De Belastingdienst vraagt de Adviescommissie om te reflecteren op de methode die door Dienst Datafundamenten & Analytics (DF&A) gehanteerd wordt bij de bias- en fairnesstoets en om een verdiepend advies te geven bij de vraagstukken die ontstaan tijdens de uitvoering van deze methode' en op de bijhorende deelvragen.

De Adviescommissie stelt vast dat DF&A systematisch te werk gaat en al veel stappen heeft gezet die in lijn zijn met het data science en engineering karakter van de afdeling. De adviesvraag legt bloot dat een puur technische aanpak grenzen kent. Hoewel het goed is om te optimaliseren vanuit technisch perspectief, moeten we ook durven constateren dat 100% biasvrije en accurate systemen niet bestaan. Er wordt gewerkt in een maatschappelijke context die voortdurend in verandering is en met technologie die feilbaar is, waardoor risico's en restbias niet uit te sluiten zijn. Maar de mogelijke aanwezigheid van risico's bij het inzetten van een systeem, maakt niet dat dit systeem per definitie onbruikbaar is. Het vergt echter wel een proactieve en kritische houding, waarbij voldoende wordt geïnvesteerd in een responsieve governancestructuur waarin het systeem wordt ingebed.

Een goede Data&AI governance zorgt ervoor dat checks and balances worden gewaarborgd. In dit schrijven komt naar voren dat er een duidelijke Data&AI governance nodig is. Te allen tijde moet dit een gedeelde verantwoordelijkheid zijn van alle partners in de keten. Er is in de bredere organisatie een continue inventarisatie en afweging nodig tussen de verschillende belangen en waarden die een rol

[1] Zie Appendix C 'Over de Adviescommissie Analytics' voor meer informatie.

spelen bij specifieke oplossingen (met inachtneming van maatstaven als proportionaliteit, billijkheid, redelijkheid). Daarbij moeten ook de volgende punten in beschouwing worden genomen:

- Transparantie over hoe modellen worden gebouwd, welke data gebruikt wordt, en hoe beslissingen worden genomen, is cruciaal.
- Het blijft noodzakelijk om modellen continu te monitoren en bij te stellen om te zorgen dat deze rechtvaardig blijven presteren en geen schade veroorzaken.
- Er moeten altijd mechanismen zijn om de schade die mogelijk toch ontstaat zo snel mogelijk te identificeren om zo getroffen individuen of groepen proactief te ondersteunen en de negatieve impact te minimaliseren.

Door risico's expliciet te maken kunnen gepaste 'checks and balances' in de vorm van monitoring en menselijk toezicht worden geïmplementeerd, welke bepaalde risico's aanvaardbaar kunnen maken. Maatregelen om eventuele onwenselijke effecten te beperken en risico's te mitigeren vinden plaats op verschillende niveaus. Zo dient men kritisch te zijn op de methode van bias- en fairnesstoetsing en het proces waarin dit plaatsvindt (hoofdstuk 3.3). Continu monitoren is vereist, met een zogeheten *human in the loop*, omdat de gouden standaard niet beschikbaar is of onderhevig is aan verandering. Mogelijke restbias kan sommige gebruikers benadelen. Dit vraagt om een gedegen wisselwerking en feedback-loop met de doelgroep (hoofdstuk 3.4). Goed risicomanagement vereist een passende governancestructuur waarin er een structuur is om waarden te bepalen en tegen elkaar af te wegen. Dit stelt DF&A en ook de bredere dienst in staat om gericht te monitoren en te reflecteren (hoofdstuk 3.5). Praktische belemmeringen in het achterhalen van (de oorzaken van) bias en het meten van fairness, zijn tot op zekere hoogte te ondervangen (hoofdstuk 3.6). Wel dient steeds rekening te worden gehouden met het (veranderende) wettelijke kader. Daarnaast leidt het aanpassen van de definitie van fairness niet persé tot bruikbaardere uitkomsten. Tenslotte bevat de Appendix A concrete voorbeelden om het advies handen en voeten te geven.

2. Adviesvraag van Datafundamenten & Analytics

Deze adviesvraag komt vanuit de Belastingdienst, Corporate Dienst DF&A en heeft betrekking op de methodologie bias- en fairnessbepaling. Eigenaar van de adviesvraag is Benjamin Jansen, directeur Corporate Dienst DF&A.

2.1 Adviesvraag

Het doel van de adviesvraag is overwegingen en/of richting te krijgen bij de vragen die bij elke biastoetsing en elke fairnessbepaling terugkomen en waar ook andere (overheids)organisaties tegenaan lopen. De hoofdvraag is: *De Belastingdienst vraagt de Adviescommissie om te reflecteren op de methode die door DF&A gehanteerd wordt bij de bias- en fairnesstoets en om een verdiepend advies te geven bij de vraagstukken die ontstaan tijdens de uitvoering van deze methode.*

De hoofdvraag wordt verduidelijkt door drie deelvraagstukken die volgen uit de overwegingen bij de verschillende stappen van de gebruikte methode. Dit zijn: (i) feiten; (ii) praktische belemmeringen en (iii) de methode op zich. De Belastingdienst vraagt de Adviescommissie om hierbij een verdiepend advies te geven.

Feiten
Er kan nooit volledigheid worden bereikt wat betreft getoetste kenmerken en in geen geval kan definitief worden vastgesteld dat geen indirecte discriminatie plaats heeft gevonden. In dat geval zou per casus

beoordeeld moeten worden, waarvoor de aantallen veel te groot zijn om dat te doen. Het is niet mogelijk om de bias in de uitkomsten van een algoritme volledig te verklaren en te mitigeren. Er zal altijd sprake zijn van 'restbias'.

Vraag 1: Moeten deze feiten geaccepteerd worden als beperking van de methodologie?

Praktische belemmeringen
Beperkingen in de hoeveelheid data

Vraag 2a: Wat is het juiste om te doen als er te weinig data beschikbaar is om een volledige oorzaakanalyse te doen met de juiste diepgang?
Vraag 2b: Is het verzamelen van (extra) beschermde persoonsgegevens (bijvoorbeeld door koppeling met data van het CBS) een logische stap om te zetten?
Vraag 2c: In sommige gevallen kan fairness volgens de afgesproken fairness definitie niet gemeten worden vanwege ontbrekende data, variabelen of doelgroep(en). Is het wenselijk om dan een andere definitie te gebruiken?

Gouden standaard

Vraag 3a: De beoordeling van de behandelaar wordt vaak gebruikt als 'waarheid', terwijl we weten dat deze beoordeling ook bias kan bevatten. Is deze beoordeling te gebruiken als gouden standaard, wanneer het borgen van de kwaliteit van dit oordeel belegd is in andere processen (bijvoorbeeld door middel van werkinstructies)?
Vraag 3b: Moet de beoordeling van de behandelaar ook onderwerp zijn van de biastoetsing?
Vraag 3c: Is er een betere gouden standaard mogelijk?

Methode
Accepteren restbias

Vraag 4a: Discriminerend onderscheid is onrechtmatig. Is het mogelijk om een harde universele norm te stellen om te bepalen wat een acceptabele hoeveelheid restbias is, waarbij dus nooit sprake is van discriminerend onderscheid? Eenzelfde mate van bias kan voor verschillende modellen, verschillende kwetsbare groepen, een andere impact hebben. Dit maakt het bepalen van acceptabele hoeveelheid restbias context afhankelijk. De vraag is welke normatieve principes leidend zijn in deze bepaling.
Vraag 4b: Moet de te accepteren restbias vooraf vastgesteld zijn of kan dat ook nadat de bias gemeten is?

Mitigatie restbias

Vraag 5a: Mag de impact van restbias gemitigeerd worden in het proces waar de uitkomst van het algoritme gebruikt wordt, bijvoorbeeld door menselijke tussenkomst?
Vraag 5b: In hoeverre is het een risico om restbias te mitigeren terwijl de oorzaak niet bekend is? Is er dan geen risico dat er per ongeluk een andere groep gediscrimineerd wordt?

2.2 Toelichting op adviesvraag

In de afgelopen jaren zijn de zorgen over het maken van onderscheid door algoritmes gegroeid omdat deze stigmatiserende, discriminerende of anderszins schadelijke of nadelige effecten kunnen hebben. Vanwege de schaal waarop algoritmische ondersteuning plaatsvindt, wil de Belastingdienst waarborgen dat de algoritmes op een rechtvaardige manier ingezet worden.

Het in kaart brengen en voorkomen van bias is onderdeel van het Impact Assessment Mensenrechten & Algoritmes (IAMA). Het Toetsingskader van de Algemene Rekenkamer en het Algoritmekader dat momenteel ontwikkeld wordt door het ministerie van Binnenlandse Zaken vragen hiervoor eveneens aandacht. Daarnaast blijkt uit het door de Belastingdienst ontwikkelde waarborgenkader voor selectie-instrumenten dat de dienstonderdelen een verantwoordelijkheid hebben om algoritmes te toetsen op

bias. Verder heeft de Adviescommissie aan Dienst Toeslagen geadviseerd[2] om te controleren of er direct of indirect ongewenst onderscheid wordt gemaakt bij de inzet van algoritmes. Dit alles bevestigt dat de Belastingdienst toetsing op zijn algoritmes wil en moet uitvoeren om te borgen dat zijn algoritmes niet discrimineren. Directe discriminatie is eenvoudig te voorkomen door de betreffende kenmerken waarop onterecht direct onderscheid gemaakt zou kunnen worden niet mee te nemen in het algoritme. Daarom gaat deze adviesvraag uitsluitend over het voorkomen van indirecte discriminatie, door het toetsen op bias en de resultaten daarvan te gebruiken voor het vaststellen van de fairness.

Er bestaat uitgebreide wetenschappelijke literatuur over het analyseren en voorkomen van bias en het borgen van fairness in algoritmes, maar er is weinig ervaring met het toepassen van deze methodes in de praktijk van de uitvoering door de overheid. Echter zijn er recent onderzoeksrapporten verschenen van instanties zoals het Centraal Planbureau en het College voor de Rechten van de Mens. Deze onderzoeken verschillen vaak in aanpak, methodes (waaronder metrieken) en definities, waardoor het voor de Belastingdienst moeilijk is om grip te krijgen op wat de juiste en beste aanpak voor deze toetsing is.

DF&A heeft een bias- en fairness methodiek ontwikkeld voor interne kwaliteitscontrole die sinds 2024 gebruikt wordt om algoritmes, die door DF&A gerealiseerd zijn en binnen de Belastingdienst ingezet worden, te beoordelen op zaken als rechtvaardigheid, behoorlijkheid en wenselijkheid (zie verder: Appendix B).

DF&A ontwikkelt algoritmes in opdracht van primaire directies van de Belastingdienst, die zelf verantwoordelijk zijn voor het proces waarin deze algoritmes ingezet worden. Daarnaast worden er ook op andere plekken in de Belastingdienst algoritmes ontwikkeld, waar DF&A niet bij betrokken is. Het bias- en fairnesstoets proces, zoals ontwikkeld bij DF&A, dient begrepen te worden als een interne controle van DF&A op door DF&A ontwikkelde algoritmes.

Deze bias- en fairnesstoets omvat onder andere een statistische analyse waarvan de uitkomsten meegenomen worden in de beoordeling van de genoemde zaken. Momenteel is DF&A ook bezig met het implementeren van de IAMA, inclusief deze bias- en fairness methodiek, in het voortbrengingsproces en het beheerproces (monitoring) van algoritmes. Onderdeel hiervan is ook een juridische toetsing van het algoritme. Als organisatie is men nog zoekende naar de juiste methode.

DF&A heeft op dit moment twee algoritmes getoetst via deze methode. In navolging op de eerdere adviesvraag die vanuit Dienst Toeslagen aan de Adviescommissie is gesteld, wil de Belastingdienst in deze verdiepende adviesvraag graag advies van de Adviescommissie inwinnen over deze toetsingsmethodiek. Het advies van de Adviescommissie wil men graag meenemen bij de doorontwikkeling en verdere toepassing van de methodiek.

Zie verder Appendix B voor de bias- en fairnesstoetsingsproces en -methode van DF&A.

3. Advies aan Datafundamenten & Analytics

3.1 Reflectie op de huidige aanpak bias- en fairnesstoetsing DF&A

De Adviescommissie stelt vast dat DF&A al heel systematisch te werk gaat en zelf investeert in kennisvergroting en samenwerking binnen de Belastingdienst. DF&A heeft al veel stappen gezet om zo

[2] https://www.rijksoverheid.nl/documenten/publicaties/2023/09/01/bijlage-3-advies-dienst-toeslagen

zorgvuldig mogelijk algoritmes te ontwikkelen, met oog voor fairness, bias en transparantie. In lijn met het data science en technologische karakter van de afdeling, wordt met name ingezet op technische instrumenten en strategieën om algoritmes verantwoord in te kunnen zetten. De geschetste bias- en fairnesstoets (zie: Appendix B) laat onder andere zien dat de analisten bezig zijn met een kritische bevraging van de algoritmes, dat er biastoetsen worden uitgevoerd en ethische afwegingen worden gemaakt.

Met de Adviescommissie en de Expertgroep Analytics wordt gewerkt aan het ontwikkelen van checks & balances, maar ook – en wellicht belangrijker – aan de praktijk van algemene digitale-ethische vaardigheden voor de hele organisatie. Belangrijk hierbij is de aandacht voor en bewustwording van mogelijk ongewenste effecten van algoritmes bij de delen van de organisatie die in contact staan met burgers. Eén van deze mogelijk ongewenste effecten is het bestaan van restbias. De doelstelling moet altijd zijn om (rest)bias te minimaliseren. De methodologie kent echter zijn beperkingen. Hier gaan we in deze adviesvraag op in, waarna per thema handvatten aangereikt worden om onwenselijke effecten te beperken.

De Adviescommissie maakt onderscheid tussen een aantal verschillende scenario's binnen DF&A die een rol spelen bij de adviesvraag. Deze worden hier kort beschreven en waar nodig worden deze apart behandeld in het advies.

Adviezen voor verschillende organisatieniveaus

Adviezen die van toepassing zijn voor DF&A, en adviezen die van toepassing zijn op de bredere context van de Belastingdienst of het ministerie.

Een goede Data&AI governance zorgt ervoor dat checks and balances worden gewaarborgd. Gegeven de rol van DF&A als aannemer van opdrachten met aansluitvoorwaarden binnen de Belastingdienst, is het ontwikkelen, implementeren en uitvoeren van een goede governance structuur niet iets waar DF&A alleen verantwoordelijkheid voor kan nemen. Te allen tijde moet dit een gedeelde verantwoordelijkheid zijn van alle partners in de keten. In dit schrijven komt naar voren dat er een duidelijke Data&AI governance nodig is. Sommige van deze adviezen hebben daarom niet louter betrekking op de adviesvrager, maar op de Belastingdienst in bredere context.

Verschillende fases in gebruik van AI met verschillende risiconiveaus

Adviezen met betrekking tot algoritmes in ontwikkeling en algoritmes in productie, met hoog of laag risico (volgens de classificering in de AI Act).

DF&A maakt een onderscheid in het toetsen van algoritmes die ontwikkeld worden en algoritmes die al in gebruik zijn. Hier sluit de AI Act bij aan. Deze vraagt expliciet om een impact assessment voordat hoog-risico algoritmes in gebruik worden genomen (pre-deployment). Het is noodzakelijk om te inventariseren welke bestaande algoritmes onder de definitie van hoog-risico vallen. Er kan een prioritering gemaakt worden van de algoritmes met de meeste impact op de burger of de samenleving. Voor hoog-risico algoritmes die al in gebruik zijn (post-deployment) is op termijn een impact assessment nodig. Een dergelijk assessment moet periodiek herhaald worden omdat omstandigheden kunnen wijzigen en risico's kunnen veranderen. Hoewel er wettelijk minder voorwaarden en vereisten zijn voor algoritmes met laag risico, is een evaluatie nog steeds van waarde. Het juridisch kader is niet het enige wat acties bepaalt, maar juist ook het morele kader. De dienst zal een positie moeten creëren en innemen met betrekking tot de evaluatie van algoritmes die niet onder hoog risico vallen, omdat deze algoritmes ook impact hebben en potentieel leed kunnen veroorzaken. Hier is governance (e.g., monitoring, proactief handelen en korte feedback loops) van belang.

Verschillende categorieën van algoritmes

Adviezen die betrekking hebben op machine learning en op simpelere rule-based modellen.

Hoewel machine learning onder een vergrootglas ligt wat betreft bias en fairness, zijn deze vragen net zo relevant voor regel-gebaseerde systemen. Deze systemen, waarbij regels vaak door mensen worden opgesteld, kunnen eveneens problemen met bias en fairness bevatten. De manier waarop deze regels tot stand komen, kan subjectieve beslissingen en vooroordelen weerspiegelen. Daarom beperken de vraagstukken rondom bias en fairness zich niet alleen tot statistische modellen en machine learning, maar omvatten ze ook traditionele, regel-gebaseerde benaderingen. In de context van DF&A is het van belang om bij zowel algoritmische als regel-gebaseerde systemen kritisch te evalueren hoe beslissingen tot stand komen en om mechanismen te ontwikkelen die zorgen voor transparantie, rechtvaardigheid en het minimaliseren van (rest)bias.

3.2 Minimaliseren van (rest)bias: de methodologie kent beperkingen

De methodologie kent zijn beperkingen – dit is een inherent aspect van het werken met data en algoritmes. Hoewel technische analyse belangrijk en noodzakelijk is voor het detecteren en mitigeren van bias, is dit niet voldoende: volledige toetsing kan niet bereikt worden, discriminatie kan op voorhand niet volledig worden uitgesloten, en het is niet mogelijk om de bias in de uitkomsten van een algoritme volledig te verklaren en te mitigeren.

Eerst wordt een deel van de bias gedetecteerd, en de restbias heeft betrekking op dat deel van de gedetecteerde bias dat niet verwijderd kan worden. Hoeveel er verwijderd kan worden en welke impact dat heeft, hangt sterk af van de omvang en aard van de gedetecteerde bias. Over het deel dat niet gedetecteerd is, valt weinig te zeggen. Het vaststellen van een enkele universele norm voor acceptabele restbias is niet haalbaar, maar een nauwkeurige evaluatie van wat acceptabel is in de specifieke context, en wat het risico acceptabel zou maken, is wel haalbaar en noodzakelijk.

Gegeven dat restbias nooit volledig verwijderd kan worden, zou de nadruk moeten liggen op het inrichten van governance aan de hand van principes die de kans op mogelijke discriminatie in de ontwikkeling en gebruik van algoritmes minimaliseert. In de praktijk past men vaak de principes van proportionaliteit en noodzakelijkheid toe. Dit betekent dat elke vorm van (rest)bias strikt noodzakelijk moet zijn voor het bereiken van een legitiem doel en dat de mate van inbreuk proportioneel moet zijn aan het nagestreefde doel. Kortom, de focus moet meer liggen op het continu monitoren van het algoritme en op het inbouwen van mechanismen in het proces om eventuele schade te minimaliseren.

De additieve maatregelen om eventuele onwenselijke effecten te beperken en risico's te mitigeren, zijn toe te passen op verschillende niveaus, die als volgt terugkomen in de onderstaande hoofdstukken van dit advies:

- Kritisch op methode en proces van bias- en fairnesstoetsing (3.3)
- Wisselwerking met de doelgroep; impact en monitoring (3.4)
- Governance en het afwegen van waarden en belangen (3.5)
- Praktische belemmeringen en mogelijke oplossingen (3.6)

Ten slotte bevat Appendix A concrete voorbeelden om het advies handen en voeten te geven.

3.3 Kritisch op methode en proces van bias- en fairnesstoetsing

3.3.1 Beoordeling van behandelaren en 'human in the loop'

Het streven moet zijn om zo foutloos mogelijk te werken. Gegeven dat volledige toetsing niet altijd mogelijk is en dat het risico op bias nooit compleet uitgesloten kan worden (dan zou elk geval afzonderlijk beoordeeld moeten worden, waarvoor de aantallen veel te groot zijn), vraagt dit om aandacht voor de mogelijkheid van ongewenste effecten. Daartoe is een bepaalde scepsis nodig van de uitvoerende medewerkers om de resultaten van algoritmes niet als "foutloos", en "zonder vooroordeel" te lezen. Het is essentieel dat de "human in the loop" niet slechts symbolisch is, maar competent genoeg om mogelijke – door het algoritme veroorzaakte – discriminatie te herkennen. Daarnaast dient deze

'human in the loop' over de capaciteit en het mandaat te beschikken om adequaat te kunnen reageren en moet deze ondersteund worden door processen om dit mogelijk (en liefst makkelijk) te maken. Algemene vaardigheden omtrent data en AI literacy, ethiek, en aandacht voor de sociale impact zijn hiervoor uiterst relevant.

De beoordeling (annotatie) van de data door de behandelaar speelt een belangrijke rol; in ieder geval bij het ontwikkelen en trainen van het algoritme (1) alsook bij het beoordelen van de output van het algoritme (2).

Voor het ontwikkelen van een bruikbaar algoritme zouden er in het toepassingsdomein ruim voldoende data en annotaties aanwezig moeten zijn (kwantiteit) die door experts als valide worden beoordeeld (kwaliteit). Aan deze evaluatie moet voldoende tijd en aandacht besteed worden en de evaluatiecriteria en beoordeling moeten vooraf worden vastgelegd. De opdrachtgever, met inhoudelijke expertise, zou de eerste evaluatieronde kunnen doen van de beoordeling van de behandelaren. De opdrachtgever levert de documentatie van de peer review aan bij DF&A en licht deze toe tijdens het interdisciplinair gesprek. De tweede evaluatieronde wordt dan door DF&A gedaan, waarbij biastoetsing plaatsvindt. De uitkomsten hiervan worden weer gezamenlijk besproken. Dit geeft de kans om via statistisch onderzoek verborgen vooroordelen in menselijke beslissingen te ontdekken en aanpassingen te doen voordat het algoritme verder wordt ontwikkeld en getest. Een algoritme moet minstens zo goed presteren als een behandelaar dat doet, maar idealiter is het een holistisch proces met als doel om de dienstverlening als geheel te verbeteren.

Daarnaast moet een proces worden opgezet om de effectiviteit van het algoritme steeds opnieuw te evalueren. Hier kan het beoordelingskader algoritmen (Algemene Rekenkamer) worden ingezet. Een zogenaamd 'functioneringsgesprek voor algoritmen' evalueert de documentatie van geconstateerde fouten en aanpassingen van een algoritme en bepaalt of het werken met het algoritme nog steeds gepast en verantwoord is. Signalen over mogelijke fouten die het algoritme veroorzaakt moeten worden geïnventariseerd en geëvalueerd. Deze signalen kunnen komen van de behandelaren, helpdesken, kritisch publiek zoals gedupeerden, of journalisten. Externe audits voor high-risk algoritmes geven de mogelijkheid van een onafhankelijk perspectief en feedback. Zie ook Appendix A.

3.3.2 Geen gouden standaard, dus continu monitoren

Het idee van een gouden standaard impliceert dat er een uitkomst of handelingswijze is die onfeilbaar is. Er is echter niet één allesomvattende standaard op basis waarvan gecheckt kan worden of alle beslissingen (menselijk of algoritmisch) correct zijn. Bovendien kan een gouden standaard ook veranderen door de tijd, bijvoorbeeld omdat wetgeving verandert, de maatschappelijke opinie zich in een andere richting ontwikkelt, of nieuwe kennis wordt opgedaan. Enige terughoudendheid met het gebruik van het idee "gouden standaard" is daarom geboden.

Gegeven de fluïde aard van de gouden standaard en de verschillende risiconiveaus van de domeinen, moet goed gekeken worden naar wat het referentiekader is voor de kwaliteit van het algoritme en hoe dit getest wordt. Bijvoorbeeld: presteert het algoritme beter dan het huidige systeem of de mens in dezelfde situatie? Wordt alleen een retrospectieve dataset gebruikt of wordt er in een live omgeving getest, bijvoorbeeld via een A/B test of een randomized controlled trial om het effect van het algoritme te testen? Deze beslissingen moeten goed gedocumenteerd en onderbouwd worden in de Model Card (zie ook: Appendix A).

Het blijft noodzakelijk om machine learning modellen continu te monitoren en bij te stellen om te zorgen dat ze rechtvaardig blijven presteren en geen schade veroorzaken. Omdat de gouden standaard veranderlijk en "always under construction" is, moet er bij ingebruikname voortdurend gemonitord worden of het algoritme nog naar behoren functioneert. Een hybride model voor monitoring, zoals

voorgesteld door Livingston Slosser et al. (2023)[3], kan hierbij als inspiratie dienen. In grote lijnen is hun voorstel om algoritmes altijd in gebruik te nemen in combinatie met menselijke interactie.

Andere governance modellen zijn ook mogelijk. Belangrijk is dat er ruimte is voor een veranderende gouden standaard en dat er continu gemonitord wordt of het systeem naar behoren blijft functioneren. Vanuit technisch perspectief kan gedacht worden aan active learning[4] en drift monitoring[5] om ervoor te zorgen dat de uitkomst van een algoritme blijft voldoen aan de kwaliteitsstandaarden, en om te identificeren of er een verschuiving van de gouden standaard aan de orde is. Het model moet frequent geüpdatet worden om de kwaliteit over tijd te kunnen garanderen. Transparantie en documentatie over hoe de kwaliteit van het model getoetst wordt (acceptatiecriteria, context, methode), is cruciaal.

3.4 Wisselwerking met de doelgroep; impact en monitoring

3.4.1 Voorzie belanghebbenden van informatie op een transparante manier

Verschillende belanghebbenden hebben verschillende informatiebehoeften. Daarom is het nodig om in begrijpbare taal uit te leggen hoe de algoritmes werken die impact op burgers hebben. Om onnodige discussies te voorkomen over de kleine kans van sociale impact door restbias, moet worden benadrukt dat er een effectieve bezwaarprocedure is voor het geval er fouten optreden. Ook is het belangrijk dat onafhankelijke derden regelmatig de werkwijze en het beheer van de algoritmes controleren, en dat er zowel vóór als na gebruik met enige regelmaat impact assessments en audits (door een externe partij) plaatsvinden.

3.4.2 Creëer een feedback-loop van doelgroep naar organisatie

Een mogelijke restbias heeft een negatieve impact op de samenleving. Burgers die hierdoor getroffen zijn merken dit vaak eerder dan de organisatie. Een oplossing hiervoor kan zijn om een feedback-loop op te zetten van de doelgroep naar de uitvoerende organisatie. Het is belangrijk om de reacties van de doelgroep te begrijpen en te zien als mogelijke signalen van ongewenste effecten van het algoritme. Deze signalen moeten worden verzameld, onderzocht en gedocumenteerd, en zo nodig ingezet worden voor een verbetering van het proces. Het betrekken van verschillende stakeholders, inclusief de groepen die mogelijk getroffen worden door de bias, bij het ontwerp en de evaluatie van algoritmes kan helpen om verschillende perspectieven en zorgen te integreren en aan te pakken.

De handreiking Non-Discriminatie by Design[6] die in 2021 is opgesteld in opdracht van het Ministerie van Binnenlandse Zaken geeft hier ook handvatten voor: "*Daarnaast is het belangrijk om in een vroegtijdig stadium belanghebbenden bij het ontwerp te betrekken. Stel dat een AI-systeem met 80% accuraatheid een bepaalde diagnose kan stellen, terwijl artsen dat slechts met 60% kunnen, maar het nadeel van het systeem is wel dat het is getraind op data met betrekking tot mannen en dus een veel hogere foutmarge heeft ten aanzien van vrouwen. Dat betekent niet dat het per definitie verboden is om zo'n systeem in te voeren, maar wel dat het belangrijk is om al in een vroegtijdig stadium in gesprek te gaan met patiëntenorganisaties over de inrichting van het ontwikkel- en besluitvormingsproces en het mitigeren van potentiële problemen. Kunnen er alsnog data worden verzameld die het beeld compleet maken? Moet het AI-systeem alleen worden toegepast ten aanzien van diagnoses bij mannen? Kan een deel van*

[3] Slosser, J. L., Aasa, B., & Olsen, H. P. (2023). Trustworthy AI: a cooperative approach. *Technology and Regulation*, *2023*, 58-68. https://doi.org/10.26116/techreg.2023.006

[4] Active learning is een leermethode van lerende systemen waarbij het systeem voor gevallen waarvoor het geen of geen zekere uitkomst kan geven, een menselijke gebruiker (idealiter een expert) vraagt om deze gevallen te beoordelen en van een label of uitkomst te voorzien. Deze gevallen met de antwoorden worden daarna door het systeem gebruikt om te leren wat het antwoord had moeten zijn.

[5] Drift in deze context betekent een verandering die niet was voorzien. Het kan zijn dat het model, de data of de relatie tussen variabelen in de data verandert over tijd. Door dit goed in de gaten te houden ('monitoring') kan er op tijd ingegrepen worden als het model minder goed gaat functioneren of presteren door drift.

[6] https://open.overheid.nl/documenten/ronl-3f9fa69c-acf4-444d-96e1-5c48df00eb3c/pdf.

de besparing in kosten en middelen worden ingezet voor extra capaciteit bij het diagnosticeren van vrouwen zonder een AI-systeem?" (p.13). Deze handvatten sluiten ook aan bij par. 3.6.1.

3.4.3 Leg scenario's voor aan het publieke domein

Door middel van Design Thinking of een andere gestructureerde methode kunnen oplossingen, innovatieve aanpakken en kosten-batenafwegingen worden besproken. Aan de hand van voorbeelden kunnen scenario's voorgelegd worden aan een diverse groep van wetenschappers, maatschappelijke organisaties en burgers. Zo bepaalt men samen wat wenselijk is. Dit zou bijvoorbeeld ingezet kunnen worden wanneer de wens bestaat om extra data, waaronder (bijzondere) persoonsgegevens, te gebruiken voor een volledige analyse van (de oorzaak van) bias (zie: par. 3.6.1). Deze scenario's zouden een goed startpunt kunnen zijn voor een dialoog met het publieke domein.

3.5 Governance en het afwegen van waarden en belangen

3.5.1 Governance commitment: Ethics by Design

Een mogelijke vervolgstap is om aan de methodiek van DF&A een governance check toe te voegen waarbij de opdrachtgever zich committeert bij de start van een ontwikkeltraject. Dit wordt ook wel Ethics by Design genoemd.

Ethics by Design houdt in dat ethische overwegingen en morele waarden worden geïntegreerd in de ontwerpfase van diensten, producten, systemen en architecturen, in dit geval specifiek gericht op bias en fairness (zie: Appendix A). Tijdens dit proces is communicatie van cruciaal belang. Regelmatig overleg met alle belanghebbenden om consensus te bereiken voordat er naar de volgende fase gegaan wordt, is belangrijk. Multi-level governance wordt toegepast om ervoor te zorgen dat er op elk niveau van het project toezicht en verantwoording is. Dit zorgt ervoor dat de uiteindelijke operationalisering van het systeem gebaseerd is op een breed gedragen akkoord en voldoet aan ethische standaarden.

3.5.2 Governancestructuur inrichten

Gezien de technische expertise die aanwezig is bij DF&A en de goede connecties die de dienst heeft in de organisatie, ziet de Adviescommissie een belangrijke rol weggelegd voor DF&A als vliegwiel voor een integraal perspectief op de inzet van algoritmes binnen de Belastingdienst. Door aan opdrachtgevers en andere actoren proactief en duidelijk te communiceren wat de grenzen zijn van de techniek en waar mogelijke risico's op de loer liggen bij ingebruikname, kan er effectieve governance worden ontwikkeld die sommige van deze geïdentificeerde risico's kan mitigeren. Hoewel technische analyse belangrijk en noodzakelijk is voor het detecteren en mitigeren van bias, is dit niet voldoende voor volledig risicomanagement. Het is belangrijk om op het niveau van de Belastingdienst een feedbacksysteem in het governance model op te nemen. Afhankelijk van de situatie moet het mogelijk zijn om bijvoorbeeld strikte menselijke controle te hebben, waarbij het systeem alleen ondersteuning geeft voor het maken van beslissingen. Dit sluit aan bij par. 3.3.2.

3.5.3 Waarden bepalen en afwegen

Verantwoord werken met algoritmes kan worden versterkt door de kernwaarden van goed bestuur te vertalen naar de digitale samenleving (zie: Meijer, Schäfer, Branderhorst, 2019[7]). In iedere context moeten belangen expliciet in kaart gebracht en afgewogen worden. Daarbinnen geldt de vraag: wat is redelijk en billijk? Het handelen moet voortkomen uit de waarden die geprioriteerd worden. Codes of conduct maken de waarden van beroepsorganisaties duidelijk en bieden richting voor de uitvoering. De kernwaarden worden vertaald naar de burgers wanneer behandelaars zich bewust zijn van de mogelijke

[7] Meijer, A. J., Mirko Tobias Schäfer, en Martiene Branderhorst. 2019. Principes voor goed lokaal bestuur in de digitale samenleving: Een aanzet tot een normatief kader. *Bestuurswetenschappen* 73.4 (2019): 8-23. Online: https://tijdschriften.boombestuurskunde.nl/tijdschrift/bw/2019/4/Bw_0165-7194_2019_073_004_003.pdf

ongewenste sociale impact van algoritmes, en er een proces van effectieve compensatie is bij fouten. Naast de competentie om algoritmes effectief in te zetten moet er dus een 'incompetentie-compensatiecompetentie' (de competentie om je eigen incompetentie te compenseren met iets anders - Marquard 1973[8]) worden ontwikkeld. Zie ook Appendix A voor andere tools.

Als de voordelen van het inzetten van het algoritme groter zijn dan de mogelijke risico's, wordt overgegaan tot ontwerp en inzetting van het algoritme. De governance moet zo worden ingericht (door middel van processen, rollen en risicomanagement) dat er gerichte monitoring en reflectie is. Dit kan op verschillende manieren, bijvoorbeeld:

- Het algoritme alleen gebruiken voor de populatie waarvan zeker is dat het goed werkt;
- Duidelijke communicatie naar de gebruikers van het algoritme (als beslisondersteuning) over de doelgroepen waarvoor meer risico of bias wordt verwacht, zodat indien nodig met menselijk toezicht kan worden bijgestuurd. Dit is ook een goed moment om het publieke domein erbij te betrekken en samen een wenselijke consensus te ontwikkelen;
- Monitoring van de werking van het algoritme in de praktijk, met nadruk op de risicogevoelige groepen en de mogelijkheid tot snelle terugkoppeling indien juiste werking niet gegarandeerd kan worden of fouten worden gedetecteerd.

3.6 Praktische belemmeringen en mogelijke oplossingen

3.6.1 Mogelijkheden om (de oorzaak van) bias te detecteren

In sommige gevallen zijn er te weinig data beschikbaar om een volledige analyse te doen van (de oorzaken van) de bias. Als er te weinig data beschikbaar zijn, moet men de risico's van bias afwegen op basis van de impact ervan, bijvoorbeeld door middel van het inschatten van de *probability* en *severity* van de impact. Hierbij is het belangrijk om de doelgroep(en) waarop de impact het grootst is te inventariseren, hoewel het niet mogelijk is om alle risico's voor alle subgroepen vooraf in te schatten. Het gaat erom dat redelijkerwijs kan worden voorzien dat een groep benadeeld zou kunnen worden. Hoe groter de mogelijke impact, hoe groter het detailniveau van bias waarnaar gekeken moet worden en hoe meer moeite gedaan moet worden om de juiste data te verkrijgen.

Het is moeilijk om een generieke dataset te ontwikkelen die alle mogelijke contexten afdekt waarin een model wordt ingezet. De complexiteit van bias en fairness in modellen hangt sterk af van de specifieke context waarin het model wordt gebruikt, en deze contexten kunnen sterk variëren. De rol van de behandelaar is cruciaal voor het waarborgen van fairness en het identificeren van bias in modeluitkomsten. Het is belangrijk dat de menselijke controle niet slechts symbolisch is, maar dat de betrokkenen de competentie en het mandaat hebben om discriminerende uitkomsten te herkennen en te corrigeren. Dit pleit voor een hybride aanpak waarbij menselijke beoordelaars samen met technische oplossingen werken om bias te detecteren en te mitigeren. Een meer geschikte aanpak zou zijn om te investeren in een governance-structuur en hybride monitoringssysteem dat ruimte biedt voor continue menselijke evaluatie en contextspecifieke aanpassingen.

Voor algoritmes in gebruik waarvan de mogelijke impact groot is, zou men extra input kunnen vragen aan de gebruiker ter bevestiging, of het handmatig kunnen controleren van uitkomsten voor doelgroepen met een groter risico op bias. Een mogelijke oplossing voor zowel nieuwe als bestaande algoritmes is het geven van een confidence score[9] en het gebruik van een conservatieve grenswaarde in overleg met de opdrachtgever. Deze grenswaarde kan erin voorzien dat er geen definitieve uitkomst wordt gegeven voor risicogroepen, maar alleen een ondersteunend advies met een waarschuwing voor mogelijke bias (zie ook: par. 3.5.2).

[8] Marquard, Odo. 1973. *Abschied vom Prinzipiellen*. Stuttgart: Reclam.
[9] Een confidence score is een cijfer dat aangeeft hoeveel vertrouwen het systeem erin heeft dat de gegeven uitkomst de juiste is.

Transparantie over de beperkingen van de data en het algoritme is in alle gevallen belangrijk (ook die gevallen waarbij er geen data voorhanden zijn om verdere analyse te doen). Deze beperkingen dienen gedocumenteerd en naar de opdrachtgever gecommuniceerd te worden.

3.6.2 Definitie 'fairness'

Fairness in algoritmes kan moeilijk meetbaar zijn wanneer er te weinig data beschikbaar is of wanneer de kwaliteit van data te wensen overlaat. Soms lijkt het praktisch om een alternatieve definitie van fairness te gebruiken om de toepassing van een algoritme mogelijk te maken, ondanks deze beperkingen. Wat 'fair' is, kan verschillen afhankelijk van de context en de waarden van de betrokken gemeenschap of samenleving. Dit onderstreept echter het belang van transparantie en het betrekken van belanghebbenden bij het bepalen wat als 'fair' wordt beschouwd.

In machine learning wordt met verschillende definities van fairness gewerkt die de veelzijdigheid en complexiteit van fairness weerspiegelen. In machine learning worden technieken gebruikt om bias te detecteren, gebaseerd op het concept van het gelijkheidsbeginsel. Dit betekent dat modellen geen ongerechtvaardigd onderscheid mogen maken op basis van kenmerken zoals ras, geslacht, religie of andere beschermde attributen. Deze concepten kunnen statistisch onderzocht worden. Het aanpassen van de definitie van fairness verandert echter niets aan de onderliggende data die mogelijk systematische vooroordelen of blinde vlekken bevatten. Data en statistische methoden zijn beperkt in het omvatten van de volledige reikwijdte van sociale context. Het is cruciaal om te begrijpen dat dezelfde data, zelfs bekeken door de lens van een andere definitie van fairness, dezelfde inherente beperkingen en vooroordelen kunnen bevatten. Daarom leidt het veranderen van de definitie niet noodzakelijkerwijs tot bruikbare of rechtvaardigere uitkomsten.

3.6.3 Tools om bias te minimaliseren

De methode waarop restbias geminimaliseerd wordt, vraagt om dezelfde procedure van kritische reflectie en bias- en fairnesstoetsing als gebruikt wordt bij het detecteren en mitigeren van de oorspronkelijke bias. Het toetsingsproces is dan ook een terugkerend proces, dat per algoritme meerdere keren uitgevoerd zal moeten worden. In een Model Card (of vergelijkbare documentatie) moet elke fase gedocumenteerd worden, waarbij de gevonden biases, toegepaste mitigatiestappen en maatregelen en hun passendheid voor het probleem worden beschreven.

Het gebruik van algoritmes vraagt om een praktijk van "Methodenkritik" of tool criticism[10]; hier zullen analisten het nodige doen om hun methode kritisch te bevragen en afwegingen te maken ten opzichte van de accuraatheid en sociale impact van een algoritme (pre-deployment).

Het mitigeren van bias kan op meerdere manieren. Softwarepakketten bieden vaak algoritmes aan, zoals 'reweighing' of een 'disparate impact remover', om met één druk op de knop bias te verkleinen. Het gevaar hiervan is dat de mitigatie toegepast wordt zonder dat het onderliggende probleem naar boven is gekomen. Daarom is het belangrijk dat ontwikkelaars van AI-modellen en algoritmes die deze pakketten gebruiken (en daarmee ook de mogelijkheid hebben om deze 'bias mitigation techniques' toe te passen), specifieke training krijgen om de bias technieken te leren. Zo krijgen zij een beter begrip van hoe de verschillende mitigatiemethoden werken en in welke gevallen deze effectief zijn.

Het mitigeren van een bias zonder kennis van de onderliggende oorzaak en zonder verificatie van de onderliggende assumpties wordt afgeraden. In bepaalde gevallen kan dit namelijk leiden tot nieuwe biases (voorbeelden hiervan zijn de introductie van de 'race-based corrections' in enkele klinische sensoren en applicaties zoals de spirometer en de eGFR[11]). Indien dit ertoe leidt dat de bias niet verholpen kan worden en fairness niet kan worden gewaarborgd tot op het gewenste kwaliteitsniveau, dan wordt afgeraden om het algoritme in productie te nemen.

[10] Methodenkritik of tool criticism: een kritische benadering van de impact die onze kennistechnologieën hebben op het resultaat.
[11] https://www.pennmedicine.org/news/publications-and-special-projects/penn-medicine-magazine/winter-2021/filtering-bias-out-of-kidney-testing

3.6.4 Juridische beperkingen

Algemeen
De vraag is of het verzamelen van (extra) (bijzondere) persoonsgegevens, mede gelet op de juridische beperkingen, een logische stap is om te zetten in gevallen waarin er te weinig data beschikbaar zijn om bias te achterhalen. Wanneer aanvullende data, waaronder (bijzondere) persoonsgegevens, gebruikt worden voor een volledige analyse van (de oorzaak van) bias, zien we twee scenario's:
(A) De (bijzondere) persoonsgegevens zijn al in het bezit van de Belastingdienst; en
(B) De dienst verkrijgt nieuwe (bijzondere) persoonsgegevens.

Het gebruik van data die al in bezit zijn van de Belastingdienst (scenario A) is alleen mogelijk indien hierbij wordt voldaan aan de toepasselijke wet- en regelgeving. Zonder volledig te zijn wijst de Adviescommissie bijvoorbeeld op het feit dat een beroep op het doelbindingsbeginsel van art. 6 lid 4 AVG veelal niet mogelijk zal zijn gelet op onder meer de verhouding tussen de Belastingdienst en de betrokkene (art. 6 lid 4 onder b AVG) en de aard van de gegevens (art. 6 lid 4 onder c AVG). De Belastingdienst dient steeds te beschikken over een grondslag ex art. 6 AVG waarbij wordt opgemerkt dat overheidsinstanties geen beroep mogen doen op art. 6 lid 1 onder f AVG voor de verwerking van persoonsgegevens in het kader van de uitoefening van hun taken - dit zou anders kunnen leiden tot het oneigenlijk oprekken van bevoegdheden.

Daarnaast moet worden gedacht aan de noodzaak om het verbod om bijzondere persoonsgegevens te mogen verwerken, zoals neergelegd in art. 9 AVG, te doorbreken. Hierbij dient ook de AI Act in acht te worden genomen, waarbij de gehanteerde techniek en wijzigingen van de AI-toepassing in de loop der tijd gevolgen kunnen hebben voor de toepasselijkheid van (bepaalde bepalingen van) de AI Act. Deze juridische aspecten gelden eveneens ten aanzien van het verdergaande scenario waarbij nieuwe data worden verzameld (scenario B).

Zo kan de inzet van CBS-data, mits uiteraard wordt voldaan aan de juridische eisen daaromtrent, een hulpmiddel zijn om de biastoets uit te voeren en zo bias te signaleren en aan te kunnen pakken. Het is daarom aan te raden om in ieder geval steeds te bezien of er een passende dataset beschikbaar is die binnen de juridische randvoorwaarden kan worden ingezet.

Volledigheidshalve merkt de Adviescommissie op dat wat juridisch mag, nog niet altijd moreel wenselijk is. Het gebruiken van data die al in bezit zijn van de Belastingdienst (scenario A) kan worden gezien als een verdedigbare stap indien wordt voldaan aan de juridische voorwaarden (voor hoog-risico systemen). Het verzamelen van beschermde persoonsgegevens (scenario B) lijkt een stap verder te gaan omdat de data nog niet in het bezit is van de dienst. Om de genoemde voorwaarden en de afweging tussen waarden te realiseren kunnen meerdere wegen bewandeld worden: een dialoog met het publieke domein om te achterhalen wat voor de doelgroep wenselijk is; een gestructureerde manier van kosten-baten afweging (zoals bij design thinking); de inzet van de IAMA; of overleg met een auditor.

Geïntensiveerd toezicht Autoriteit Persoonsgegevens
Onlangs is bekendgemaakt dat de Belastingdienst de komende vijf jaar onder geïntensiveerd toezicht van de Autoriteit Persoonsgegevens (AP) zal staan, mede naar aanleiding van recente berichten over de inzet van Risico Analyse Modellen.[12] Dit toezicht houdt onder meer in dat begeleiding wordt geboden bij het duurzaam verbeteren van de bescherming van persoonsgegevens. Nu de AP ook aangewezen is als de algoritmetoezichthouder raden wij DF&A aan om dit toezicht te zien als een kans en actief met de AP in contact te treden over onder meer de juridische aspecten van de bias- en fairnesstoets.

Opslag en gebruik nieuwe data
De opslag en het gebruik van nieuwe data vragen aandacht voor de cybersecurity requirements. Het introduceren van een nieuwe dataset met sensitieve informatie of het samennemen van bestaande databronnen kan namelijk gepaard gaan met nieuwe, digitale veiligheidsrisico's. Hier moet zo goed mogelijk tegen beschermd worden door het inrichten van een veilige omgeving waarbij goed wordt

[12] Brief staatssecretaris Van Rij d.d. 28 mei 2024.

nagedacht over cybersecurity[13]. Het gebruik van Datasheets of vergelijkbare documentatie (en opslag van deze sheets op een toegankelijke plek, gekoppeld aan de data zelf) is van groot belang om transparantie over het doel, het gebruik en de safeguards van (het gebruik van) de data te documenteren. Zie ook Appendix A.

[13] Cybersecurity zijn de activiteiten die erop gericht zijn om computersystemen te beschermen tegen bedreigingen en aanvallen van buitenaf, zoals ongeoorloofde toegang en schade aan of diefstal van data.

Appendix A: Opzet voor verantwoord ontwikkelen en gebruik algoritmes

Deze appendix geeft enkele voorbeelden hoe verantwoorde datapraktijken manifesteren in het werk van ontwikkelaars. Deze opsomming is niet uitputtend en kan verder worden aangevuld en aangepast.

Voor de onderbouwing van algoritmes zijn drie elementen nodig die naar een breed publiek gecommuniceerd kunnen worden: a) motivatie, b) beschrijving werkwijze, en c) beschrijving governance. Deze kunnen op verschillend detail-niveau voor diverse doelgroepen worden aangeleverd.

a) Motivatie: betreft de reden waarom gekozen is voor dit proces; belangrijk om de waardenpropositie (resultaat reflectie value-sensitive design, DEDA[14], of deel 1 van IAMA; zie boven) te benadrukken. Onderdeel hiervan is ook het overwegen of machine learning inderdaad geschikt is (is er data om van te leren en zijn er (complexe) patronen die geleerd moeten worden?) en of er alternatieven zijn;
b) Beschrijving werkwijze algoritme: uitleg wat het algoritme doet en hoe;
c) Beschrijving governance: uitleg hoe het algoritme beheerd wordt en wat de mogelijkheden zijn om feedback te geven of bezwaar aan te tekenen.

In de afgelopen jaren is een reeks praktijken ontwikkeld om de uitleg van algoritmes te verbeteren, toezicht te verbeteren en verantwoordelijkheid te waarborgen.

1. Ontwikkeling en motivatie

- Value-sensitive design ontwikkeling
  Gedurende de ontwikkeling van een algoritme kan worden gekozen voor een proces van value-sensitive design ontwikkeling. Dit is een gedocumenteerd reflectieproces, gericht op de waarden die een algoritme moet bevatten of mogelijk kan aantasten en op het mitigeren van mogelijke risico's (denk aan processen zoals DEDA of CODIO[15]).

- Gezamenlijk design thinking proces
  Een gezamenlijk design thinking traject met het publieke domein om te komen tot breed gedragen wenselijke consensus. Het betrekken van diverse stakeholders, inclusief de groepen die potentieel beïnvloed worden door de bias, bij het ontwerp- en evaluatieproces van algoritmes kan helpen om verschillende perspectieven en zorgen te integreren en aan te pakken.

- Waardekaart
  Een waardekaart creëren die per context wisselend ingevuld kan worden in hetzelfde raamwerk (zoals de waardekaart voor ruimtelijke planning[16] en het UWV Kompas Data Ethiek[17]).

- Framework voor waardenafweging
  Een gestructureerde en beargumenteerde afweging maken tussen de impact op verschillende waarden en het kiezen van de meest gunstige oplossing. Een voorbeeld van een framework en stappenplan hiervoor is bijvoorbeeld genoemd in *An Ethics Framework for Big Data in Health and Research (2019) – Diagram 1*[18] (dit paper is afgestemd op toepassing in de zorg maar eenzelfde aanpak of een variatie zou hier gebruikt kunnen worden).

[14] De Ethische Data Assistent, zie: https://deda.dataschool.nl/
[15] Zie: Code Goed Digitaal Openbaar Bestuur
[16] https://www.argumentenfabriek.nl/media/1898/10056-ik-waardenkaart-ro-s.pdf
[17] https://www.uwv.nl/imagesdxa/kompas-data-ethiek_tcm94-442573.pdf
[18] Xafis, Vicki, et al. "An ethics framework for big data in health and research"; Asian Bioethics Review 11.3 (2019): 227-254.

- Datasheet
  Een datasheet (Gebru et al 2021[19]) helpt bij het beschrijven van de dataset en stelt derden in staat de nodige informatie in te winnen.

2. Implementatie en werkwijze algoritme

- Model card
  Een model-card beschrijft het model, mogelijke valkuilen en dual-use aspecten, levert informatie over bias etc. (Mitchel et al. 2019[20]).

- Impact assessments
  Voor de implementatie leveren impact assessments een indicatie op van de mogelijke ongewenste gevolgen voor verschillende stakeholders, handvatten voor de implementatie en onderhoud en het mitigeren van risico's (denk aan IAMA of de Plot4AI ResponsibleAI QuickCheck[21]).

- Audits
  Audits (intern of extern) leveren een verificatie en documentatie van een verantwoorde werkwijze van het algoritme (zie bijvoorbeeld de AI Auditing checklist van de European Data Protection Board[22]).

- Iteratief functioneringsgesprek
  Gedurende het gebruik vindt iteratief een "functioneringsgesprek" plaats om geconstateerde fouten en veranderingen in het algoritme te documenteren, alsmede mogelijke bezwaren en de behandeling daarvan.

3. Governance

- Ethics by Design: governance check voor ontwikkeling
  Ethische principes worden vertaald naar concrete ontwerpvereisten om transparantie, rechtvaardigheid en verantwoord gebruik van technologie te waarborgen. In plaats van ethische kwesties achteraf aan te pakken, worden ze vanaf het begin meegenomen in het ontwikkelingstraject.
  Concreet worden de volgende stappen doorlopen.
  1. Conceptual Engineering, waar de doelen en context van het project duidelijk gedefinieerd worden.
  2. Requirements Engineering, waar specifieke eisen en criteria vastgesteld worden voor fairness en bias mitigatie.
  3. Ontwikkelingsfase, waarin de technische oplossingen ontworpen en geïmplementeerd worden.
  4. Evaluaties en tests, na elke ontwikkelingscyclus, om ervoor te zorgen dat de systemen voldoen aan de gestelde eisen.

[19] Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets Documentation to facilitate communication between dataset creators and consumers." Communications of the ACM, Vol. 64, no. 12, pp. 86-92. Online: https://cacm.acm.org/research/datasheets-for-datasets/

[20] Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229. Online: https://dl.acm.org/doi/abs/10.1145/3287560.3287596?casa_token=fQD_k__bKmUAAAAA:clWgm12o7tLKe9RdnbXnBImVzVdi8_eQk7xNqH06NCR16RVkeibxgmCwl_CwV0nrMKU173k21Kw

[21] https://plot4.ai/assessments/quick-check

[22] https://www.edpb.europa.eu/system/files/2024-06/ai-auditing_checklist-for-ai-auditing-scores_edpb-spe-programme_en.pdf

- Governance framework
  Een voorbeeld van een AI governancestructuur die helderheid geeft over hoe waarden van een organisatie in relatie staan met 'responsible AI' elementen (waaronder bias) en de bijbehorende risico's) is voorgesteld door het Alan Turing Institute[23,24]. Een dergelijk framework zou gebruikt kunnen worden om de waarden van de organisatie te vertalen naar processen en naar best practices die de kwaliteit van algoritmes kan waarborgen.

- Hybride monitoringsmodel: algoritmes gecombineerd met menselijke interactie (Livingston Slosser, 2023)[25].
  Het hybride model steunt voor schaalvergroting en efficiëntie op het algoritmische onderdeel (track 1) en voor handmatige controle, toezicht en het aanpassen van de gegevensinvoer en -uitvoer op de menselijke input (track 2 en 3). Dit betekent dat een deel van de input via het algoritme loopt, een ander deel via de behandelaren, en dat een derde track met menselijke beoordeling (een deel van de) de output van track 1 en 2 beoordeelt. Hoewel dit model iets inboet op efficiëntie – er is immers geen sprake van volledige automatisering – draagt het bij aan de validatie van het algoritmische systeem dat door de feedback loop van de menselijke beoordelingen ontvankelijk wordt voor contextuele veranderingen (zie ook par. 3.3.2). Bovendien stimuleert dit bij de organisatie en haar werknemers een actieve leerhouding.

Het is handig om dergelijke documenten bij datasets en modellen te voegen. Zo kunnen deze opgeleverd worden bij vragen van toezichthouders, kritisch publiek en stakeholders. Ook borgen de documenten de continuïteit van verantwoord werken met algoritmes, en faciliteren zij een kennistransfer van ontwikkelaars naar gebruikers en beleidmakers. Vertrouwen in de praktijk ontstaat door goede uitleg, documentatie en een effectief proces van bezwaar.

[23] https://www.turing.ac.uk/news/publications/ai-ethics-and-governance-practice-introduction
[24] https://www.turing.ac.uk/news/publications/ai-ethics-and-governance-practice-ai-fairness-practice
[25] Slosser, J. L., Aasa, B., & Olsen, H. P. (2023). Trustworthy AI: a cooperative approach. *Technology and Regulation*, *2023*, 58-68. https://doi.org/10.26116/techreg.2023.006

Appendix B: DF&A Bias- en fairnesstoetsingsproces en -methode

1. Interdisciplinair gesprek

Het toetsingsproces begint bij een interdisciplinair gesprek met relevante interne stakeholders binnen de Belastingdienst (o.a. opdrachtgever van het algoritme, ontwikkelteam van het algoritme, fiscaal-inhoudelijk experts, jurist). Het gesprek wordt ondersteund door een aantal vragen uit de IAMA (onderdeel 1, 'Waarom'). De besproken punten zijn o.a.:

a) Grondrechtelijke risico's voor de burger die van belang zijn in het bedrijfsproces waarin het algoritme geïmplementeerd wordt.
b) De variabelen waarop getoetst dient te worden: kwetsbare populaties en mogelijke proxies voor ontbrekende beschermde variabelen.

Overwegingen bij het interdisciplinair gesprek:

- Vanwege beperkingen in de data (bijvoorbeeld ontbrekende informatie over beschermde persoonsgegevens, over mogelijke proxies en over intersectionaliteit (het behoren tot meer dan één beschermde groep, bijvoorbeeld vrouwelijke gender en het hebben van een migratieachtergrond)) en de onmogelijkheid om compleet te zijn, kunnen wij nooit volledig uitsluiten of bevestigen dat er indirecte discriminatie plaatsvindt.
- In het advies van de Adviescommissie aan Dienst Toeslagen lezen wij: "*Door strategisch te denken over de problemen en de mensen die van de dienst gebruik maken, kan tot een redelijke lijst [van getoetste kenmerken] gekomen worden*". Het is voor ons en onze stakeholders moeilijk in te schatten wanneer de lijst van kenmerken voldoende mogelijke risico's dekt.

2. Biastoetsing

De uitkomsten van het interdisciplinaire gesprek bepalen de keuzes voor de biastoetsing. Voor de biastoetsing worden eerst, op basis van de vastgestelde fairnessdefinitie, relevante metrieken gekozen voor de biasmeting die de risico's voor de burger het beste reflecteert. We kijken hierbij altijd naar de vertegenwoordiging in de selectie van de gedefinieerde groepen (demographic parity) en naar een specifieke metriek die hoort bij de gekozen fairnessdefinitie. De bias van de kwetsbare groepen wordt afgezet tegen de bias van de restgroep.

Vervolgens wordt de biastoets uitgevoerd, de technische meting van hoeveel bias er aanwezig is. De uitkomsten van de biastoets zijn hierbij ook getoetst op statistische significantie.

Overwegingen bij de biastoetsing:

- Vanwege toetsing met metrieken op groepsniveau kunnen wij nooit uitsluiten dat er indirect gediscrimineerd wordt. De uitkomst kan gemiddeld rechtvaardig zijn, maar dat betekent niet dat het algoritme nooit indirect discrimineert.
- Er moet een afweging gemaakt worden tussen de beperkingen van de data t.o.v. de risico's voor de burger in de keuze van een fairness metriek.
- De handmatige beoordeling van de behandelaar wordt als 'waarheid' of 'gouden standaard' gebruikt om de toetsing uit te voeren.
- In het advies van de Adviescommissie aan Dienst Toeslagen lezen wij: "De kern is dat het ervan afhangt wat met de uitkomsten van het model gedaan wordt en wat de consequenties zijn (worden er hiervoor bv. andere mensen minder snel gemonitord, wat als het model het fout heeft?)". Wij vragen ons daarom af of een algoritme los van andere beoordelingsstappen getoetst kan worden. Niet alleen het algoritme zelf maar ook de beoordeling van de behandelaar kan namelijk een bias bevatten en uiteindelijk nadelige effecten voor burgers hebben.

3. Bias oorzaak analyse

Vervolgens wordt er een bias oorzaak analyse uitgevoerd voor de uitkomsten van beide metrieken. Hierbij wordt een steekproef genomen van de groep waarin de bias gevonden is. Er worden door relevante inhoudelijke experts hypotheses geformuleerd en onderzocht voor mogelijke oorzaken van de bias. Vervolgens wordt gemeten hoeveel bias er overblijft als een gevonden oorzaak/oorzaken wordt weggehaald.

Overwegingen oorzaakanalyse:

- Wij vragen ons af hoe relevant een oorzaakanalyse is van gevonden bias tijdens de ontwikkeling van een nieuw algoritme.
- In sommige gevallen is er te weinig data beschikbaar om een volledige analyse te doen van de oorzaken van de bias. Bijvoorbeeld om te onderzoeken of alle etniciteiten evenveel benadeeld worden of dat het specifiek om één bepaalde etniciteit gaat. Hoe kleiner het detailniveau, hoe meer data benodigd is; dit terwijl de subpopulaties juist kleinere groepen zijn (en dus minder datapunten bevatten).
- De gemeten bias zal nooit volledig verklaard kunnen worden, er zal altijd restbias overblijven, ook als er wel voldoende data beschikbaar is.

4. Fairnessbepaling

Als laatste stap in het toetsingsproces wordt, wederom in een interdisciplinair gesprek (met o.a. fiscaal-inhoudelijk experts, jurist, ontwikkelteam van het algoritme), de fairness van het algoritme bepaald. Hierbij wordt voor de uitkomsten van beide gebruikte metrieken beoordeeld of de gevonden bias (en de gevonden oorzaken) ethisch en juridisch toelaatbaar is, i.e. geschikt, noodzakelijk en proportioneel voor de uitvoering van onze wettelijke taak. Het gesprek wordt ondersteund door een aantal vragen uit de IAMA (onderdelen 2 ‘Wat’, 3 ‘Hoe’ en 4 ‘Mensenrechten’). Indien nodig kunnen de stappen 3 en 4 iteratief worden herhaald.

Overwegingen fairnessbepaling:

- Het is de vraag hoe kan worden bepaald welke restbias acceptabel is, welk percentage toelaatbaar is. Is dit 1%, 10% of 100%? Hierbij staat een bias van 100% voor een verdubbeling van de kans op selectie door het algoritme.
- Daarnaast is een overweging of restbias altijd in het algoritme opgelost moet worden. Misschien is het voldoende als deze geaccepteerd wordt door de eigenaar of als deze in het beoordelingsproces gemitigeerd wordt.

In het advies van de Adviescommissie aan Dienst Toeslagen lezen wij: “*Een bias is niet gelijk aan discriminatie en discriminatie blijkt niet persé uit een algoritme maar kan ook pas blijken uit de impact. Het type fouten kan anders zijn voor verschillende doelgroepen. Uitgedacht moet worden wat de impact daarvan is en in welke mate dit acceptabel is. IAMA kan als een kader behulpzaam zijn.*” Echter blijkt het (zelfs met IAMA als kader) moeilijk in de praktijk om samen met de stakeholders te bepalen welke impact acceptabel is.

Appendix C: Over de Adviescommissie Analytics

Het ministerie van Financiën heeft de onafhankelijke Adviescommissie Analytics (verder: de Adviescommissie) ingesteld voor de Belastingdienst, Dienst Toeslagen, Douane en het kerndepartement om haar kritische blik te versterken en bij te dragen aan een lerende organisatie. Het doel van de Adviescommissie is bijdragen aan een meer verantwoorde omgang met data-analyse, algoritmes, risicomodellen en artificiële intelligentie. De Adviescommissie adviseert in brede zin vanuit de vijf perspectieven[26] die de Algemene Rekenkamer (AR) in haar rapport 'Aandacht voor algoritmes' als toetsingskader heeft meegegeven.

Voordat adviesvragen worden gesteld aan de Adviescommissie, worden deze intern voorbereid door een brede expertgroep, afkomstig van het ministerie van Financiën, de Belastingdienst, Dienst Toeslagen en de Douane. Een adviesvraag kan gericht zijn op hoe data en analytics ingezet kan worden in een nieuwe context met daarbij de nadruk op ethische en sociale aspecten.

Voor wat betreft de werkwijze staat het voeren van dialoog centraal. De Adviescommissie is nadrukkelijk geen onderzoeks- of toetsingscommissie.

De leden van de Adviescommissie betreffen:

- prof. dr. E.H.L. Aarts (Emile) - voorzitter
- prof. dr. S. Bhulai (Sandjai)
- prof. dr. P.J.J. van Geest (Paul)
- prof. dr. M.J. van den Hoven (Jeroen)
- mr. F.C. van der Jagt (Friederike)
- R. van Kan MCI, BSc (Rob)
- prof. dr. E.L.O Keymolen (Esther)
- dr. P. Prüfer (Patricia)
- dr. M.T. Schäfer (Mirko)
- dr. A. van Wissen (Arlette)

Procesverloop beantwoording adviesvraag

Op 5 februari 2024 heeft de adviesvrager de adviesvraag gepresenteerd en toegelicht aan de Adviescommissie met ruimte voor verklarende en verdiepende vragen van de Adviescommissie. Ter vergadering is voor de behandeling van de adviesvraag een subcommissie ingesteld, bestaande uit Sandjai Bhulai, Esther Keymolen, Mirko Schäfer en Arlette van Wissen. De subcommissie is op 7 maart bijeengekomen en n.a.v. de eerste bespreking is op 8 maart aanvullende informatie gevraagd aan de adviesvragers. Dat is 14 maart ontvangen en meegenomen naar de tweede bijeenkomst van de subcommissie op 26 maart. Deze bespreking heeft geleid tot een eerste conceptadvies welke in schriftelijke rondes nader is uitgewerkt en op 15 april is gepresenteerd tijdens de vergadering van de Adviescommissie. Ook is op 21 mei een extra sessie met de adviesvrager geweest om de governance nader toe te lichten.

N.a.v. de subcommissiebijeenkomst op 4 juni is rapporteur Nienke Voshart gestart met het opstellen van het definitief conceptadvies. De tekst is voor een feitelijke check ook voorgelegd aan de adviesvragers.

[26] Te weten: (i) Sturing en verantwoording, (ii) Model en data, (iii) Privacy, (iv) IT General Controls (ITGC) en (v) Ethiek. Zie verder Algemene Rekenkamer, Rapport 'Aandacht voor algoritmes', 14 januari 2021

Reacties van de subcommissie op het definitief concept én de reactie van de adviesvragers heef geleid tot een versie die op 8 juli is besproken met de leden van de Adviescommissie en is de leden de mogelijkheid geboden nog schriftelijke input aan te leveren.

Het tweede definitieve conceptadvies is besproken tijdens de bijeenkomst van de Adviescommissie op 7 oktober 2024 met de insteek: *comply or explain*. De resultante was *comply* waarmee het conceptadvies door de Adviescommissie definitief is vastgesteld en aan de adviesvrager is aangeboden.