Gemeente Amsterdam

Versie 1.0
Januari '26

Bias Impact Assessment – Wat is eerlijk?

Template om in te vullen

Een systematische manier om zelfstandig een weloverwogen en bewuste eerste keuze te maken of de altijd aanwezige bias acceptabel is en een algoritme burgers en/of medewerkers eerlijk behandelt.

Doel van dit document

Mensen en algoritmes[1] kunnen bias vertonen, wat wordt gedefinieerd als een 'systematisch verschil in behandeling van mensen of groepen in vergelijking met anderen'. Dit kan leiden tot ongelijke behandeling, zoals discriminatie. Hoewel bias niet volledig kan worden voorkomen, kunnen de juiste maatregelen helpen om de impact ervan te beperken.

Met het Bias Impact Assessment (BIA) voer je vroegtijdig en gestructureerd het gesprek over bias. En maak je weloverwogen een bewuste eerste keuze. Hiermee bepaal je hoe de toepassing verder wordt ontworpen en ontwikkelt wordt. Dit is een iteratief proces en bedoeld voor hoog-risico AI-systemen en impactvolle algoritmes en AI-systemen. Vanuit drie invalshoeken – ethisch, juridisch en technisch – maakt de BIA je bewust van het risico op discriminatie. Als (proces)eigenaar van het algoritme of AI-systeem bepaal je zelfstandig of (het risico op) de bias acceptabel is.

Verdeeld over de drie invalshoeken begeleiden we je in ongeveer 15 vragen naar jouw conclusie rondom bias. Per invalshoek krijg je toelichting over het doel, meer informatie hoe je tot een conclusie kunt komen en geven we enkele voorbeelden ter inspiratie.

Hoe dit document in te vullen

1. Beantwoord de << rood gemarkeerde teksten >>;
2. Bepaal met de proceseigenaar en als projectteam of (het risico op) de bias acceptabel is;
3. De proceseigenaar (bv. lijndirecteur) accordeert dit document en zet een handtekening;
4. Bewaar het document bij de overige project documentatie. Voeg dit Bias Impact Assessment toe bij bestuurlijke voordrachten en andere relevante communicatie met belanghebbenden.

Met wie contact opnemen bij vragen

Bij opmerkingen en vragen over dit document kan er contact opgenomen worden met het centrale algoritmeteam via algoritmes@amsterdam.nl.

---

[1] Een algoritme is een set van regels en instructies die een computer geautomatiseerd volgt bij het maken van berekeningen om een probleem op te lossen of een vraag te beantwoorden. Algoritme is de verzamelnaam, waar ook AI-systemen onder vallen. Zie de SAM-pagina Algoritmes voor meer informatie.

Inhoudsopgave

Bias Impact Assessment – Wat is eerlijk? 1

Inhoudsopgave 2

1 Conclusie Bias Impact Assessment 3
Algemene uitgangspunten om onwenselijke bias te voorkomen 4

2 Achtergrond Bias Impact Assessment 5
2.1 Wat is bias? 5
2.2 Waarom dit Bias Impact Assessment en hoe maak je een bewuste keuze? 6

3 Ethische Bias Impact Assessment: willen wij dit? 9
3.1 Hoe helpt dit algoritme de stad Amsterdam, haar bewoners, ondernemers en bezoekers? 9
3.2 Welke (mogelijke) negatieve gevolgen ervaren (groepen) personen door het gebruik van de gekozen variabelen in het algoritme? 11
3.3 Conclusie Ethische Bias Impact Assessment 14

4 Juridische Bias Impact Assessment: mag het? 15
4.1 Direct en indirect onderscheid tussen (groepen) personen is rechtvaardig 16
4.2 Persoonsgegevens worden op een rechtvaardige en behoorlijke manier gebruikt 19
4.3 Conclusie Juridische Bias Impact Assessment 21

5 Technische Bias Impact Assessment: kan het? 22
5.1 Wat vinden we belangrijk om later te toetsen? 23
5.2 Sluit het algoritme en de invoer goed aan op de mensen en situaties waarvoor ze bedoeld zijn? 25
5.3 Geschiktheid en eerlijkheid zijn gedefinieerd om later te kunnen toetsen 27
5.4 Conclusie Technische Bias Impact Assessment 29

Bijlage 1. Definities 30

1 Conclusie Bias Impact Assessment

Mensen en algoritmes kunnen bias vertonen, wat wordt gedefinieerd als een 'systematisch verschil in behandeling van mensen of groepen in vergelijking met anderen'. Dit kan leiden tot ongelijke behandeling, zoals discriminatie. Hoewel bias niet volledig kan worden voorkomen, kunnen de juiste maatregelen helpen om de impact ervan te beperken.

Dit Bias Impact Assessment (BIA) helpt om een weloverwogen en bewuste eerste keuze te nemen of (het risico op) de bias acceptabel is en het algoritme burgers en/of medewerkers eerlijk behandelt.

Vinden we (het risico op) de aanwezige bias in dit specifieke algoritme acceptabel om door te gaan met de ontwikkeling?

Baseer de conclusie op de uitkomsten van de ethische- (paragraaf 3.3), juridische- (paragraaf 0) en technische (paragraaf 5.4) bias-analyses. Indien er sprake is van een toepassing met een combinatie van verschillende type algoritmes, toets deze algoritmes afzonderlijk en maak een afweging voor de toepassing als geheel.

Algoritme of AI-systeem		<< Vul in welk algoritme is beoordeeld >>
Conclusie Bias Impact Assessment	**Ethisch**	<< Kopieer de conclusies uit hoofdstuk 3 Ethische Bias Impact Assessment >>
	Juridisch	<< Kopieer de conclusies uit hoofdstuk 4 Juridische Bias Impact Assessment >>
	Technisch	<< Kopieer de conclusies uit hoofdstuk 5 Technische Bias Impact Assessment >>
	Integrale conclusie	Alle invalshoeken beschouwend, vinden we (het risico op) de bias in dit specifieke algoritme << wel / niet >> acceptabel om door te gaan met de ontwikkeling.
Vervolgacties/maatregelen		<< Beschrijf vervolgacties in het algemeen en per invalshoek >>
		<< Indien reeds bekend, beschrijf welke gedetailleerdere bias-analyses later in het proces gedaan zullen worden >>

Ondertekend door:

Naam	<< Vul in >>
Functie	<< Vul in >>
Directie	<< Vul in >>
Plaats en datum	<< Vul in >>

Handtekening proceseigenaar:

<< Vul hier de handtekening in >>

_____

Algemene uitgangspunten om onwenselijke bias te voorkomen

Door onderstaande wettelijke en beleidsmatige uitgangspunten na te leven wordt onwenselijke bias zoveel mogelijk voorkomen:

- ✓ Wees je bewust dat een algoritme zonder bias niet mogelijk is. Een mens-gedreven proces zonder bias is ook niet mogelijk. Alle algoritmes en mensen hebben vooroordelen en maken fouten. Zoek zowel naar bias in het algoritme en de data zelf, als in de mensen en processen om het algoritme heen.
- ✓ Wees je bewust van de mogelijk grote impact bij de inzet van algoritmes. Ook bij simpele beslisbomen. Denk bijvoorbeeld aan recente Nederlandse schandalen met discriminatie bij de inzet van algoritmes rondom de kinderopvangtoeslag ("toeslagenaffaire"), opsporen van fraude met studiebeurzen, antispieksoftware en onrechtmatig gebruik van 'fraudescorekaarten'[2] .
- ✓ Zet een multidisciplinair team in, dit vergroot de kans dat verschillende soorten bias herkend worden. De samenstelling van het team verschilt per project en context. Zorg dat bij het BIA op zijn minst een proces eigenaar/projectleider of product owner en teamleden met data-analyse en juridische kennis betrokken zijn.
- ✓ Laat een mens altijd het uiteindelijke besluit nemen[3], niet het algoritme. Zorg dat een beslissing (zie paragraaf 4.2 voor het verschil met 'besluit') van een algoritme altijd in "normale taal" aan een burger uitgelegd kan worden.
- ✓ Zorg dat het werkproces en werkinstructies zo zijn ingericht dat de menselijke neiging van overmatig vertrouwen op geautomatiseerde systemen beperkt wordt. En dat medewerkers dus kritisch blijven nadenken. Diverse e-learnings kunnen je hierbij op weg helpen, bijvoorbeeld "AI in de gemeente", "Non-discriminatie in algoritmes en data", "de Nationale AI-Cursus" of "de Ethische kant van AI".
- ✓ Wees extra voorzichtig wanneer het algoritme profileert[4], wat wil zeggen dat je persoonlijke aspecten van personen analyseert of voorspelt. Schakel altijd een jurist en/of privacy officer en data analist of data scientist om te helpen met onwenselijke bias voorkomen. En zorg dat je in het werkproces ook willekeurige selectie en signaal-gedreven selectie inzet, naast algoritme profilering-gedreven selectie[5].
- ✓ Gebruik nooit zelflerende-algoritmes ("machine learning" algoritmes) wanneer het algoritme burgers profileert en selecteert voor een bepaald doel. En wees extra voorzichtig met voorspellende algoritmes.
- ✓ Controleer ook na ingebruikname op bias (bv. 1 keer per jaar): sommige bias ontstaat pas als het algoritme in het werkproces draait. En hernieuw de BIA als er een substantiële verandering in het proces/algoritme is.
- ✓ Evalueer periodiek (bv. 1 keer per jaar) of het algoritme nog presteert zoals bedoeld. Stel een protocol vast voor de situatie dat er (een vermoeden van) discriminatie door een algoritme is geconstateerd en pas dit wanneer nodig toe. Een template protocol van het ministerie van BZK is te vinden op Discriminatieprotocol.

---

[2] Zie bijvoorbeeld Algoritmes en discriminatie hebben hoofdrol in privacytoezicht | Autoriteit Persoonsgegevens.
[3] Voor meer informatie, zie de standaard 'Betekenisvolle menselijke tussenkomst risicoprofileringsalgoritmes'
[4] Bijvoorbeeld om iemands beroepsprestaties, economische situatie, gezondheid, persoonlijke voorkeuren, interesses, betrouwbaarheid, gedrag, locatie of verplaatsingen te analyseren of te voorspellen.
[5] Voor meer informatie over profilering zie AVG-regels voor profilering | Autoriteit Persoonsgegevens en voor een praktische standaard rondom profileringsalgoritmes zie Publieke standaard 'Profileringsalgoritmes'

2 Achtergrond Bias Impact Assessment

Waarom

- De gemeente staat voor eerlijke en transparante inzet van algoritmes en AI-systemen ("algoritmes")
- Met het gebruik van algoritmes schenden we soms mensenrechten. Onder andere door onwenselijke bias[1], bv. discriminatie. Dit willen we voorkomen
- Ook willen we onbedoeld negatieve effecten van het algoritme op (groepen) mensen voorkomen

Doel instrument

- Het Bias Impact Assessment instrument biedt handvatten om op een systematische manier, een weloverwogen, bewuste en vroegtijdige eerste keuze te maken of de aanwezige bias acceptabel is en het algoritme eerlijk
- Het instrument is aanvullend aan bv. een Ethische Bijsluiter, DRIA, DPIA en gedetailleerde technische bias-analyses

Voor wie

- *Proceseigenaar:* Epic Owner, B-Lead, Product Owner en/of projectleider
- *Inhoudsdeskundigen voor specifieke hulpvragen:* digitale rechten & ethiek officer, jurist of privacy officer, data analist/-scientist, centrale algoritmeteam

Uitkomst instrument

- Vroegtijdige beslissing of een algoritme:
  1. Eerlijk is;
  2. Eerlijk is en welke aanpassingen doorgevoerd worden; of
  3. Oneerlijk is en stoppen met ontwikkeling/gebruik van het algoritme.
- Weloverwogen en bewuste eerste keuze gebaseerd op Amsterdamse waarden, juridische kaders en technische (on)mogelijkheden
- Indicatie voor gedetailleerdere bias-analyses als het algoritme wordt (door)ontwikkelt en ingezet

Disclaimer en door-ontwikkeling instrument

- Het instrument wordt idealiter in een vroeg stadium gebruikt, zodat het kan helpen om een algoritme verder vorm te geven
- Dit instrument bevat de belangrijkste elementen rondom bias. En het is een eerste versie. Aan de hand van praktijkervaringen wordt dit instrument continu geüpdatet
- Waar relevant wordt verwezen naar gedetailleerde uitleg en instructies, dit instrument is niet bedoeld als "allesomvattend" naslagwerk

[1]Bias is gedefinieerd als 'een systematisch verschil in behandeling van mensen of groepen in vergelijking met anderen'.

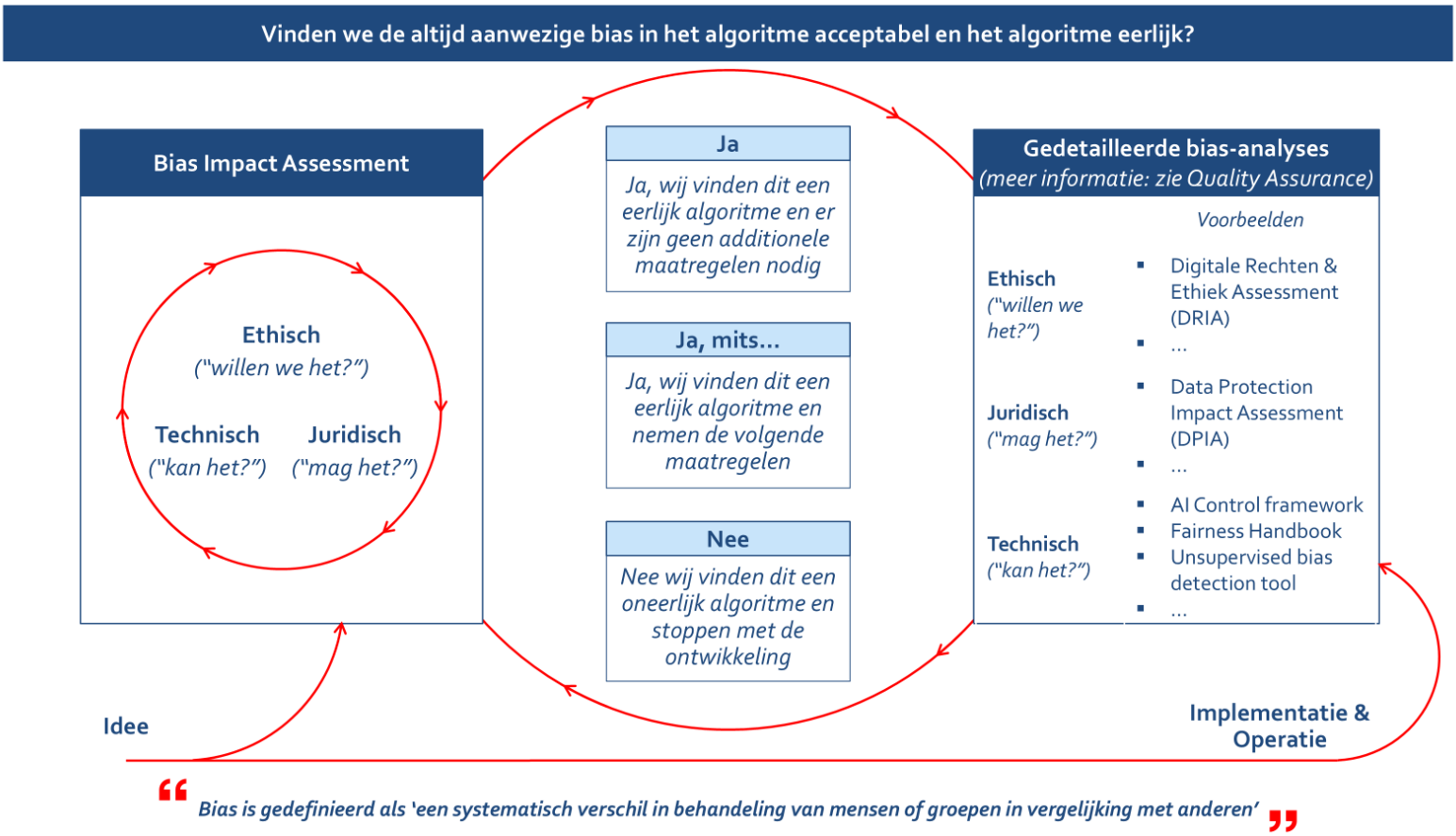
2.1 Wat is bias?

Bias is een systematische vertekening in hoe we informatie waarnemen, verwerken of beslissingen nemen. Een voorbeeld is dat in sommige landen bepaalde boetes worden gerelateerd aan iemands inkomen: des te hoger iemands inkomen, des te hoger de boete. Dit zorgt voor bias. In dezelfde situatie hebben mensen met een lager inkomen een voordeel en worden mensen met een hoger inkomen benadeeld. Deze bias is echter een bewuste keuze en wordt in deze landen eerlijk gevonden.

Bias is er altijd en komt niet alleen voor in algoritmes, maar ook in menselijk handelen. Bias op zichzelf is niet per se slecht. Het kan een bewuste keuze zijn, bijvoorbeeld wanneer we ongelijk investeren in gelijke kansen. Daarmee vergroten we kansengelijkheid door sommige groepen systematisch anders te behandelen dan andere groepen. Deze bias vinden we acceptabel.

In andere gevallen kan de bias wél problematisch zijn, vooral als het leidt tot discriminatie of onbedoelde ongelijkheid. Daarom is het belangrijk om bias te herkennen en bewust te accepteren of corrigeren waar nodig. Dit Bias Impact Assessment instrument helpt daarbij.

2.2 Waarom dit Bias Impact Assessment en hoe maak je een bewuste keuze?



"Bias is gedefinieerd als 'een systematisch verschil in behandeling van mensen of groepen in vergelijking met anderen'"

Ons werk moet steeds efficiënter: meer taken met minder geld. Algoritmes en AI-systemen (verder: "algoritmes") kunnen daarbij helpen. Maar die algoritmes moeten wel eerlijk en transparant ingezet worden. Vaak worden algoritmes ingezet om processen te automatiseren en beslissingen op grote schaal te maken. Dit kan in korte tijd grote gevolgen hebben.

Daarbij is het erg moeilijk om discriminerende effecten te voorkomen. Discriminerende algoritmes zijn bijvoorbeeld meestal niet overduidelijk racistisch of seksistisch. De problemen zijn subtiel, en daardoor moeilijker op te lossen. Discriminatiegronden weglaten als variabele in je algoritme is niet genoeg. Want discriminatie ontstaat ook indirect. En algoritmes kunnen snel een discriminerend effect hebben op grote groepen mensen.

Daarom zijn in juli 2025 enkele uitgangspunten en keuzes gemaakt om (het risico op) de bias acceptabel te houden. Het Bias Impact Assessment is een uitwerking van de uitgangspunten en keuzes en is bedoeld als hulpmiddel om het algoritme en bijbehorende proces te verbeteren. Het Bias Impact Assessment faciliteert een bewuste, weloverwogen eerste keuze of de altijd

aanwezige bias in het algoritme acceptabel is en het algoritme eerlijk. Het instrument is bedoeld voor hoog-risico AI-systemen en impactvolle algoritmes en AI-systemen.[6]

Het instrument kan in verschillende stadia van algoritme-ontwikkeling gebruikt worden – of het nou een idee is of een algoritme wat reeds (ingekocht en) in productie is. Des te eerder de Bias Impact Assessment gebruikt wordt, des te eerder besloten kan worden om bijvoorbeeld het algoritme anders in te richten, processen aan te passen of geen algoritme te gebruiken.

Het Bias Impact Assessment behandelt de ethische, juridische en technische invalshoek. Kort door de bocht: uit een juridische bias-analyse volgt of het algoritme voldoet aan diverse wetgeving en uit een technische bias-analyse krijg je inzicht in de mate van bias, in de vorm van een getal of percentage. Op die conclusies kun je baseren dat een algoritme *niet* acceptabel is, bijvoorbeeld door onrechtmatig gebruik van bepaalde variabelen. Maar je kunt niet zeggen of (het risico op) de bias *wel* acceptabel is. Dat is namelijk geen exacte wetenschap. Wat eerlijk is, hangt af van de situatie en kan verschillen ("case-by-case"), of veranderen over tijd (bij dezelfde casus). Dit kan alleen worden vastgesteld door menselijk (af)wegen, wat samenkomt met de toevoeging van de ethische bias-analyse.

Het Bias Impact Assessment bevat de belangrijkste elementen uit de drie invalshoeken om een goede eerste beslissing te nemen om het algoritme al dan niet in te zetten. Voor een volledig beeld rondom bias zijn gedetailleerdere analyses nodig. Bijvoorbeeld door een technische bias-analyse waarin gemeten wordt of de bias binnen acceptabele grenzen valt. Exacte (detail)analyses hangen van de context af.

Overleg binnen het projectteam welke gedetailleerde bias-analyses nodig zijn. Bijvoorbeeld aan de hand van de risico classificatie van het algoritme, waarbij een hoger risico algoritme meer analyse vereist. Het projectteam kan hier zelf met hulp van een jurist/privacy officer en een data analist/data scientist een beslissing over nemen en vastleggen. In dit instrument wordt waar relevant naar de belangrijkste additionele analyses verwezen.

De drie invalshoeken worden in aparte hoofdstukken behandeld voor de bruikbaarheid en leesbaarheid. Het is echter nadrukkelijk de bedoeling om een integrale afweging te maken. De vragen zijn bedoeld om te inspireren en het projectteam aan het denken te zetten. Daarom werkt het goed om iteratief door het Bias Impact Assessment te lopen: pas waar nodig het algoritme en proces aan en evalueer daarna wat het effect is op eerdere beslissingen. Ook als er iets verandert in het algoritme of het proces.

Bekijk deze Bias Impact Assessment niet als losstaand instrument:

- ✓ Zorg dat je aan alle Quality Assurance eisen per project/algoritme voldoet: Quality Assurance - Kwaliteitsborging - Startpagina. De maatregelen zijn aanvullend aan dit Bias Impact Assessment. Zo bepaal je bijvoorbeeld aan de hand van Amsterdamse Tada

[6] Bepaal hier de risico classificatie van het algoritme: 04a. Risico Classificatie algoritme - Tamtam

waarden de morele kans van technologie of een project in de Ethische Bijsluiter. En ga je in een DPIA specifiek in op privacy risico's bij (hoog-risico) gegevensverwerking. Dit Bias Impact Assessment bevat additionele vragen en onderwerpen, waarbij specifiek gefocust wordt op bias, het algoritme en de context.

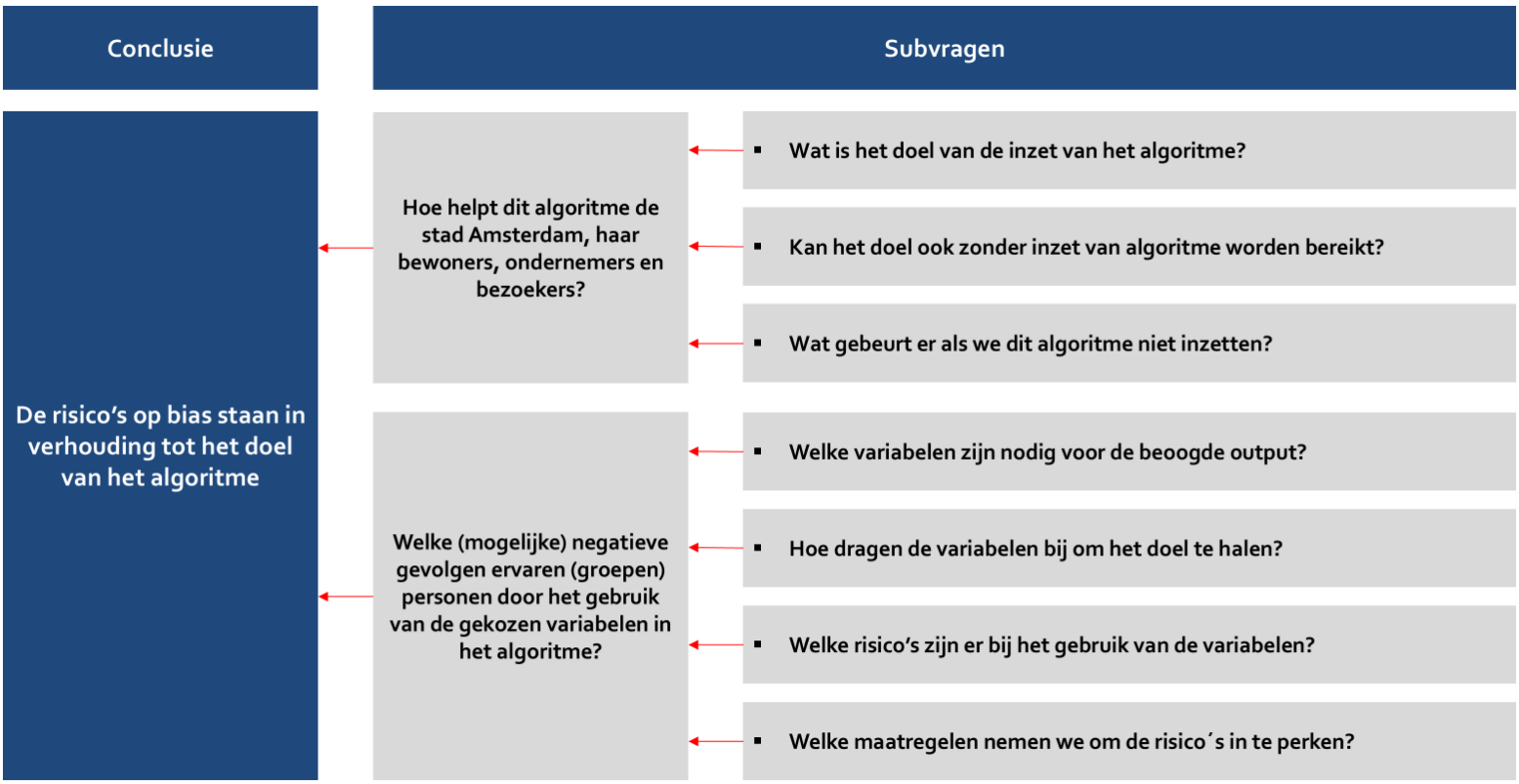
- ✓ Gebruik je een AI-systeem? Volg dan de principes uit de vastgestelde Amsterdamse AI Visie, waarbij we de mens centraal stellen en betrouwbare en toekomstbestendige AI-systemen inzetten. Zie de Amsterdamse Visie op AI | Innovatie.

- ✓ Zie de SAM-pagina rondom Algoritmes voor meer achtergrondinformatie. Het centrale algoritmeteam kan daarbij ook helpen, te bereiken via algoritmes@amsterdam.nl.

- ✓ Indien gewenst, is hulp bij specifieke vragen beschikbaar voor iedere invalshoek via:

    - Ethisch: Digitale rechten en Ethiek SAM-pagina of het digitale rechten team (digitalerechten@amsterdam.nl).
    - Juridisch: Jouw jurist of privacy officer.
    - Technisch: Jouw data analist/data scientist.
        - Voor detailvragen over bias in AI-systemen: het AI Enablement team, via Generieke Voorzieningen.
    - Overige vragen: centrale algoritmeteam (algoritmes@amsterdam.nl).

3 Ethische Bias Impact Assessment: willen wij dit?



3.1 Hoe helpt dit algoritme de stad Amsterdam, haar bewoners, ondernemers en bezoekers?

Bij de Ethische Bias Impact Assessment staat de vraag centraal: willen wij dit? Deze vraag moet je stimuleren om bewust te zijn van de effecten van het algoritme en om weloverwogen het algoritme in te zetten. Het stimuleert je om na te denken over het doel van het algoritme en welke variabelen moeten helpen om het doel te behalen, welke mogelijke risico's hierbij horen en welke maatregelen genomen kunnen worden om risico's in te perken.

Wellicht is al een Ethische Bijsluiter gedaan voor het project, waarbij gefocust is op ethische en verantwoorde digitalisering. Dit Bias Impact Assessment bouwt hier op voort, met verdiepende vragen en onderwerpen waarbij specifiek gefocust wordt op het algoritme en de context.

Voorbeeld: casus Mokumse Bos

Om concrete voorbeelden te geven, gebruiken we in de Bias Impact Assessment steeds dezelfde fictieve casus uit het "Mokumse Bos": een algoritme dat boswachters helpt om dieren op camerabeelden in het Mokumse Bos te herkennen, labelen en snel overzicht te geven welke dieren gespot zijn. Het uiteindelijke doel is het onderzoeken of zeldzame dieren uitsterven. En effectiever plagen te kunnen bestrijden.

Gebruik de voorbeelden ter inspiratie, het is geen volledige opsomming van mogelijkheden.

3.1.1 Wat is het doel van de inzet van het algoritme?

Denk aan: Welk probleem wordt opgelost? Of hoe wordt de huidige situatie verbeterd door het gebruik van het algoritme? Haalt het algoritme menselijke bias weg?
Heb je de Morele Kans van het project in een Ethische Bijsluiter bepaald? Gebruik deze dan als basis.

Voorbeeld: casus Mokumse Bos

Het algoritme helpt boswachters door foto's van wildcamera's te analyseren en een voorstel te doen welke diersoort is vastgelegd. Hierdoor krijgen boswachters snel een overzicht welke dieren waargenomen zijn, op basis van duizenden foto's. Dat scheelt handmatig labelen van iedere foto, waar momenteel geen tijd voor is en waardoor boswachters meer tijd hebben voor hun kerntaken.

<< Vul in >>

3.1.2 Kan het doel ook zonder inzet van het algoritme worden bereikt? Of met een minder complex algoritme?

Denk aan: zijn er alternatieve werkwijzen (zonder algoritmes) die eerlijker zijn? Hoeveel tijd, middelen en kosten zijn nodig voor een niet-algoritmische aanpak? Wat zijn de risico's of beperkingen van het niet gebruiken van een algoritme? Kunnen we in plaats van een complex algoritme (bv. machine learning) ook simpele algoritmes (bv. rekenregels) gebruiken?

<< Vul in >>

3.1.3 Wat gebeurt er als we het algoritme niet inzetten?

Denk aan: Zijn er cruciale processen of doelen die vertraging oplopen of niet worden bereikt als het algoritme niet wordt ingezet? Ontstaat er een extra werklast of inefficiëntie voor medewerkers of systemen als het algoritme niet wordt gebruikt?

<< Vul in >>

3.1.4 Hoe helpt dit algoritme de stad Amsterdam, haar bewoners, ondernemers en bezoekers?

Denk aan: welke (groepen) personen ondervinden positieve/negatieve effecten door de inzet van het algoritme? Wie is er gebaat bij de inzet van het algoritme? Wat heeft de Amsterdammer of ambtenaar eraan?

<< Vul in >>

3.2 Welke (mogelijke) negatieve gevolgen ervaren (groepen) personen door het gebruik van de gekozen variabelen in het algoritme?

Het algoritme wordt met een reden ingezet en heeft een duidelijk doel (paragraaf 3.1). Om dat doel te bereiken heeft het algoritme variabelen nodig om te functioneren. In de context van een algoritme zijn variabelen gegevens of waarden die het algoritme verwerkt, opslaat of aanpast. Bijvoorbeeld het aantal inwoners of type woning. Variabelen fungeren als een soort opslagplaats waarin informatie tijdelijk wordt opgeslagen, zodat het algoritme berekeningen kan uitvoeren.

Variabelen spelen een cruciale rol bij het uitvoeren van berekeningen en het verwerken van gegevens. Het gebruik van variabelen brengt echter risico´s met zich mee. Verkeerd geïnterpreteerde variabelen kunnen leiden tot ongewenste gevolgen. Dit kan gebeuren als variabelen onvolledig zijn, vooroordelen bevatten of niet representatief zijn voor de werkelijkheid.

Voorbeeld: casus Mokumse Bos

In het Mokumse Bos is een konijnenplaag. De konijnen tasten de begroeiing en bomen aan in het bos. De konijnen hebben geen natuurlijke vijand en moeten worden gevangen om op een andere plek te worden uitgezet. Door het grote aantal konijnen, besluiten de boswachters om een algoritme in te zetten. Het algoritme gebruikt beelden van wildcamera´s om de locaties van konijnen te traceren zodat de boswachters effectief vallen kunnen zetten.

De algoritmes zijn getraind op beelden van konijnen uit het Mokumse Bos. Als variabelen gebruikt het algoritme de oren en voeten van konijnen, die onderscheiden zich in het bijzonder van andere veel voorkomende dieren zoals reeën en wilde zwijnen. De boswachters zetten vallen, maar vangen niks. Na lang zoeken komen de boswachters erachter: het algoritme heeft hazen verward met konijnen. Hazen lijken op konijnen, maar hebben grotere oren en voeten en een andere leefomgeving.

Soms kun je een directe variabele niet makkelijk meten, zo is 'sociale cohesie' bijvoorbeeld niet direct te meten. Daarvoor zou je bijvoorbeeld het 'aantal buurtinitiatieven' of 'hoeveelheid vrijwilligerswerk' kunnen nemen. Dit noemen we proxy variabelen. Bij proxy variabelen is de kans op indirecte/onbewuste discriminatie groter, omdat het verband tussen de gemeten waarde en de conclusie niet altijd logisch is. Dit is het principe "correlatie is nog geen causaliteit": dat er een verband is, betekent nog niet dat de variabele een goede *verklaring* is voor het effect. Wees daarom extra kritisch bij het gebruik proxy variabelen. Voorbeelden van proxy variabelen zijn:

Variabele	Als proxy voor…	Waarom de variabele niet geschikt is
Arbeidsduur	Promotiekansen	Vrouwen werken gemiddeld minder/korter, onder andere vanwege zorg voor kinderen, waardoor indirecte discriminatie een reële kans is omdat ze in dit geval minder kans maken op een promotie.
Aantal arrestaties	Criminaliteit	Waar je controleert, vind je meer. Als je in de ene wijk meer controleert, krijg je meer arrestaties. Dat wil niet per se zeggen dat de criminaliteit daar relatief hoger is.
Aantal en prijs auto's	Opleidingsniveau	Hoewel meer en dure auto's een indicatie kunnen zijn van inkomen, zegt het niet per se iets over het opleidingsniveau. Je kunt "aantal en prijs auto's" dus niet gebruiken om opleidingsniveau te bepalen om zo bijvoorbeeld gerichte onderwijsprogramma's aan te bieden.
Afstand tussen twee postcodes	Kans op fraude uitwonende studiebeurs	Culturele verschillen zorgen ervoor dat studenten met een migratieachtergrond dichter bij hun ouders blijven wonen – dit zegt niets over de kans op fraude.
Meldingen vieze straten	Sociale status wijk	Afval op straat kan ook andere oorzaken hebben, bijvoorbeeld omdat er door toerisme onevenredig meer mensen in een bepaalde wijk zijn waardoor wijkbewoners onevenredig hard geraakt worden door bijvoorbeeld meer controles en dus meer boetes.
Tijdstip van aanvragen	Indicatie van fraude	Mensen met onregelmatige-/nachtdiensten doen niet op "normale" tijdstippen een aanvraag en worden vaker ten onrechte onderzocht op fraude.
Aantal verkochte ijsjes	Aantal mensen dat verdrinkt	Meer mensen verdrinken als er meer ijs wordt verkocht. Maar een betere variabele om verdrinkingen te voorspellen is temperatuur: met mooi weer zijn er meer mensen op het strand, dus er verdrinken ook meer mensen. En er wordt meer ijs verkocht, maar er is geen causaal verband tussen aantal verkochte ijsjes en verdrinken.

Bedenk met het team welke (proxy) variabelen gebruikt worden en waarom. Bedenk welke mogelijke nadelige impact het gebruik van de variabelen heeft, zoals mogelijke discriminatie. Definieer kwetsbare groepen en besteed extra aandacht aan de mogelijke negatieve impact op hen. En bepaal de maatregelen die moeten worden genomen (of zijn genomen) om de nadelige impact te verlagen. Gebruik de tabel op de volgende pagina als template.

Voor een generatieve AI-systeem is het vaak lastig om de gebruikte variabelen te bepalen. Bepaal mogelijke nadelige impact dan op basis van andere factoren, bijvoorbeeld de informatie die je in het model stopt, het proces eromheen, de "temperatuur" (hoe creatief het model mag zijn) die ingesteld wordt, prompts die gedaan worden, bias op basis van de uitkomsten, etc. Bestaande technische standaarden vanuit het AI Enablement team kunnen daarbij als inspiratie dienen.

In een gedetailleerdere bias-analyse zal een volledige analyse moeten worden gedaan op de (proxy)variabelen aan de hand van een gedegen datastudie. Hoofdstuk 5 is daarbij het begin, waarna jouw data analist / data scientist of het AI Enablement team verder kan helpen.

Verschillende werkvormen zijn mogelijk, het is aan het team welke methode de voorkeur heeft. Hierbij enkele suggesties:

1. Workshops of interactieve sessies: Print de tabel op de volgende pagina en gebruik post-its om individueel proxy variabelen, mogelijke nadelige impact en maatregelen te bepalen.
2. Ethische 'serious game': aan de hand van dillema's de groep keuzes laten tussen verschillende variabelen die gebruikt kunnen worden.
3. Krantenkoppen spel: bepaal per variabele welke krantenkop zou verschijnen in het Parool als a) het algoritme precies doet wat het doel is en b) het algoritme (onbedoeld) onwenselijke bias heeft.

3.2.1 Welke (mogelijke) negatieve gevolgen ervaren (groepen) personen door het gebruik van de gekozen variabelen in het algoritme?

Variabele	Waarom variabele gekozen?	Als proxy voor…	Mogelijke nadelige impact (bias), met extra aandacht voor kwetsbare groepen	Te nemen maatregelen
<< Variabele 1 >>	<< Vul in >>	<< Vul in >>	<< Vul in >>	<< Vul in >>
<< Variabele 2 >>	<< Vul in >>	<< Vul in >>	<< Vul in >>	<< Vul in >>
<< Variabele 3 >>	<< Vul in >>	<< Vul in >>	<< Vul in >>	<< Vul in >>
<< Variabele 4 >>	<< Vul in >>	<< Vul in >>	<< Vul in >>	<< Vul in >>
<< Variabele 5 >>	<< Vul in >>	<< Vul in >>	<< Vul in >>	<< Vul in >>

3.3 Conclusie Ethische Bias Impact Assessment

Willen we het algoritme gebruiken gebaseerd op de Ethische Bias Impact Assessment?

Vorm een conclusie gebaseerd op paragraaf 3.1 en 3.2. Maak uiteindelijk een integrale afweging in hoofdstuk 1.

<< Ja/Nee >>, de risico's op bias op (groepen) personen (zie paragraaf 3.2) staan << wel/niet >>, in verhouding tot het doel van het algoritme (zie paragraaf 3.1).

We gaan << wel/niet >> door met het algoritme en nemen onderstaande maatregelen:

- **<< Vul maatregelen in >>**

4 Juridische Bias Impact Assessment: mag het?

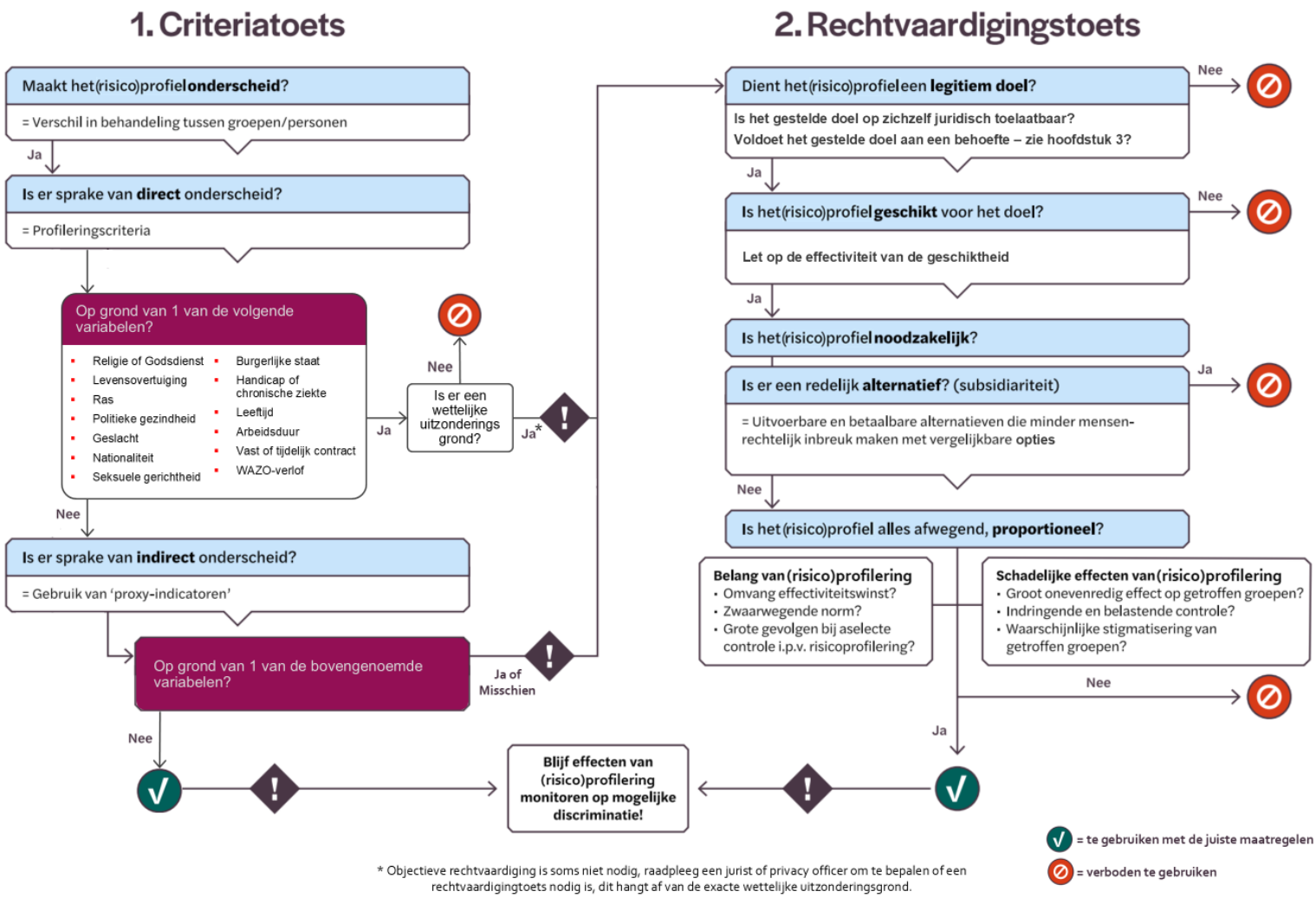


Bij de Juridische Bias Impact Assessment staat de vraag centraal: mag het? Deze vraag moet je stimuleren om bewust te zijn van rechtvaardig onderscheid tussen (groepen) personen en bij het gebruik van persoonsgegevens in het algoritme. Hierbij bouw je voort op de Ethische Bias Impact Assessment.

Belangrijk is om te onderzoeken of het algoritme direct of indirect onderscheid tussen (groepen) personen maakt. En of dat onderscheid wel rechtvaardig is. Bijvoorbeeld vanuit de Algemene wet gelijke behandeling. Daarnaast bepaal je welke persoonsgegevens je gebruikt en of dit mag volgens geldende wetgeving zoals de Algemene verordening gegevensbescherming (AVG). Wees je bewust dat je ook indirect gebruik kunt maken van persoonsgegevens, volgens zogenoemde proxy variabelen (zie ook paragraaf 3.2).

De gemeente is in veel gevallen wettelijk verplicht een risicoanalyse te maken met betrekking tot de bescherming van persoonsgegevens bij bepaalde processen en/of producten. Zo'n risicoanalyse wordt ook wel een DPIA (*Data Protection Impact Assessment*) genoemd. Daarin breng je de privacyrisico's in het proces in kaart en neem je maatregelen om risico's te verkleinen. De Juridische Bias Impact Assessment richt zich ook op persoonsgegevens, maar focust specifiek op de werking in het algoritme en bias. Beide hulpmiddelen zijn dus aan elkaar gelinkt en vullen elkaar aan om een goed beeld te krijgen van privacyrisico's en bias.

4.1 Direct en indirect onderscheid tussen (groepen) personen is rechtvaardig[7]



Figuur 1 Aangepaste versie van toetsingskader tegen discriminatie van het College voor de Rechten van de Mens

In processen en algoritmes maken we vaak onderscheid tussen (groepen) personen. Bijvoorbeeld om hen sneller te helpen door ze naar een afdeling te sturen die hun probleem waarschijnlijk kan oplossen. Dit heet profilering – want je maakt een (risico)profiel van (groepen) personen – wat is toegestaan onder bepaalde voorwaarden. Om te toetsen of het (risico)profiel voldoet aan de non-discriminatienormen, moeten twee stappen worden doorlopen (zie bovenstaande Figuur 1):

[7] Voor meer informatie en toelichting, met name rondom risicoprofilering op grond van ras en nationaliteit, zie Nieuw toetsingskader tegen discriminatie door risicoprofilering | College voor de Rechten van de Mens. Figuur 1 is aangepast om ook te voldoen aan andere situaties dan risicoprofilering en andere variabelen.

1. De criteriatoets: maakt de (risico)profilering direct of indirect onderscheid op grond van een aantal beschermde variabelen?
2. De rechtvaardigingstoets: is het doel van het onderscheid legitiem, en de profilering daartoe geschikt, noodzakelijk en proportioneel?

Direct onderscheid op basis van bovenstaande variabelen is verboden, tenzij er een wettelijke uitzondering is. Indirect onderscheid is eveneens verboden, tenzij er een goede reden voor is (objectieve rechtvaardiging). Objectieve rechtvaardiging betekent dat gebruik van de variabele proportioneel is, dat wil zeggen dat er een legitiem doel is, het risicoprofiel geschikt is voor het doel, noodzakelijk is en er geen redelijk alternatief is. Om te bepalen of persoonsgegevens rechtvaardig gebruikt worden moet onder andere het doel met het middel vergeleken worden, net als in de Ethische Bias Impact Assessment. Neem daarom ook de eerder geformuleerde doelen uit vraag 3.1.1 mee. Soms is het nodig om indirect onderscheid exact te meten, dit kan het beste in de gedetailleerde juridische en technische bias-analyse. Vervolgens kun je met de extra informatie controleren of je nog steeds op dezelfde conclusie uitkomt in stap 2. de rechtvaardigingstoets.

4.1.1 Bepaal per onderstaande variabele of er direct of indirect onderscheid gemaakt wordt en of er een objectieve rechtvaardiging voor het maken van onderscheid is

Direct onderscheid op basis van onderstaande beschermde variabelen is niet toegestaan, tenzij er een specifieke wettelijke uitzondering geldt. Indirect onderscheid mag alleen als er een goede reden voor is (objectieve rechtvaardiging). Waarschijnlijk gebruik je geen beschermde variabelen. Maar toch kun je de beschermde variabelen indirect wel gebruiken, door zogenoemde proxy variabelen. Daarom sta je expliciet stil bij de variabelen die je gebruikt en of deze variabelen in jouw algoritme en proces een mogelijke proxy is voor een van de beschermde variabelen. Zie voor meer uitleg paragraaf 3.23.2. Gebruik bovenstaande flowchart in Figuur 1 per (proxy) variabele om te bepalen of er objectieve rechtvaardiging voor het maken van onderscheid is en documenteer de uitkomsten in onderstaande tabel.

Beschermde variabelen	Voorbeelden van mogelijke proxy variabelen *(zie paragraaf 3.2)*	Gebruiken we de beschermde variabele en mogen we direct onderscheid maken?	Mogen we indirect onderscheid maken?	
			Is er indirect onderscheid door proxy variabelen?	Het indirect onderscheid is proportioneel en gerechtvaardigd (is geschikt voor een legitiem doel, noodzakelijk en zonder redelijk alternatief) *(zie figuur 1)*
Religie of godsdienst	Stemgedrag, postcode, foto's waarop religieuze attributen of kleding te zien zijn	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Levensovertuiging	Sociale media activiteit, taalgebruik	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Politieke gezindheid	Stemgedrag, opvattingen over migratie en klimaat, voorkeur voor type media	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Ras	Geboorteland, huidskleur, postcode, genetica	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Etniciteit	Taalgebruik, voornaam of achternaam, laaggeletterdheid, buitenlands familielid	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Geslacht	Voornaam, gezondheidsdata, zwangerschap	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Nationaliteit	Postcode, taalgebruik, verblijfsstatus, hoogte inkomen, kenteken	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Seksuele gerichtheid	Sociale media activiteit, gebruik van specifieke apps	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Burgerlijke staat	Gezinsgrootte, fiscale data (bv. toeslagen)	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Handicap of chronische ziekte	Gebruik hulpmiddelen, verzuimpatronen, gezondheid	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Leeftijd	Schooljaar, pensioengerechtigde status	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Arbeidsduur	Salaris, contracturen per week, sociale klasse	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
Vast of tijdelijk contract	Opleidingsbudget, aanspraak transitievergoeding	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>
WAZO-verlof	Verlofregistratie, deelname aan ouderschapsprogramma's	<< Ja, want er is een wettelijke uitzondering >> << Nee, dus we gebruiken de variabele niet >>	<< Ja / Misschien / Nee >>	<< Ja, want… >> << Nee, dus we gebruiken de proxy variabele niet>>

4.2 Persoonsgegevens worden op een rechtvaardige en behoorlijke manier gebruikt

"*Persoonsgegevens: alle informatie direct over een persoon, ofwel naar een persoon te herleiden is*"

Direct herleidbare persoonsgegevens

- Naam
- Adres
- Postcode
- Woonplaats
- Geboortedatum / leeftijd
- Telefoonnummer
- E-mailadres
- Foto's/beeldmateriaal
- …

Overige persoonsgegevens

- Online (surf)gedrag
- Allergieën
- Personeelsnummer
- IP-adres
- Kenteken
- Social media accounts
- Hobby's
- Aankoopgedrag
- …



VOORBEELDEN

Bijzondere persoonsgegevens

- Ras of etniciteit
- Politieke opvattingen
- Lidmaatschap vakbond
- Seksueel gedrag / seksuele gerichtheid
- Religie / levens beschouwende overtuiging
- Genetische gegevens (zoals DNA)
- Biometrische gegevens (zoals vingerafdruk, stemherkenning en gezichtsscan)
- Gezondheid (is breder dan alleen medische gegevens)
- …

Gevoelige persoonsgegevens

- Strafrechtelijke gegevens
- Kopie ID-kaart / paspoort / rijbewijs
- BSN
- Locatiegegevens
- Financiële gegevens
- Politiegegevens
- …

Naast de beschermde variabelen uit paragraaf 4.1 willen we zorgvuldig omgaan met persoonsgegevens. Je bepaalt welke persoonsgegevens je gebruikt en kijkt per persoonsgegeven in meer detail of het gebruik ervan rechtvaardig en behoorlijk is.

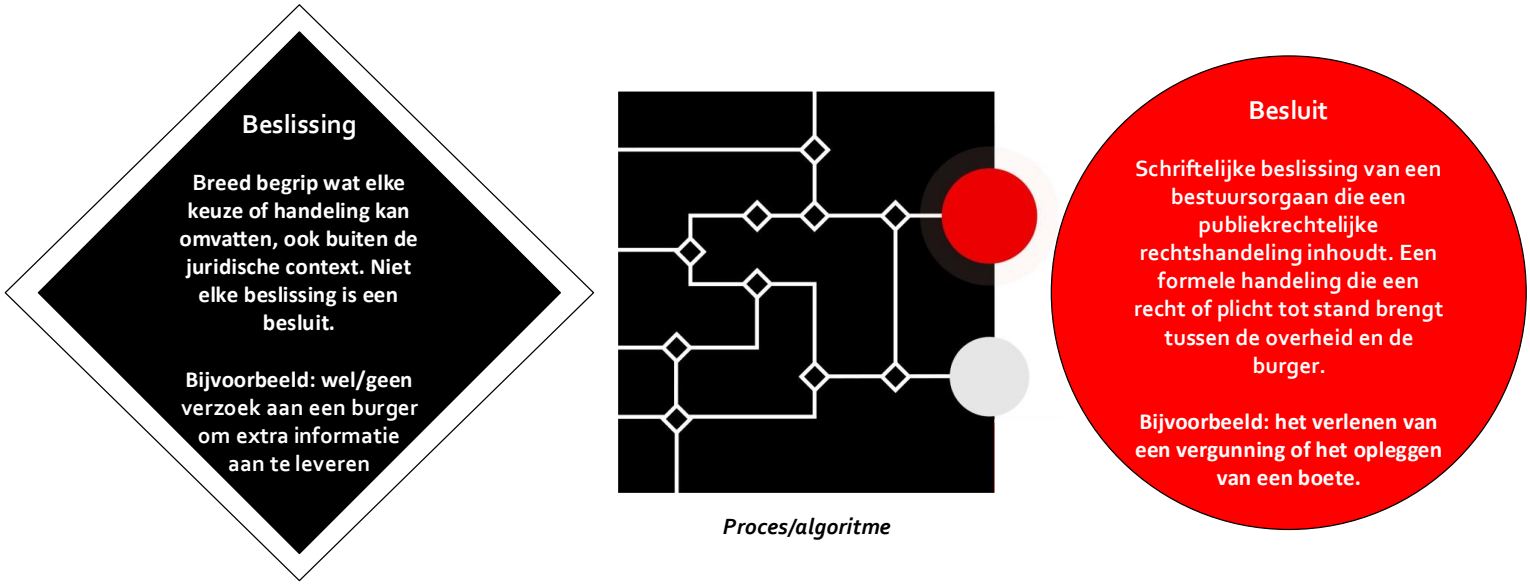
Sommige persoonsgegevens gaan direct over iemand, dit noemen we directe persoonsgegevens. Daarnaast zijn er gegevens die niet direct over iemand gaan, maar die wel naar die persoon te herleiden zijn, dit noemen we indirecte persoonsgegevens. Het gaat dan om gegevens die in combinatie met andere gegevens iets zeggen over een persoon. Zo kan bijvoorbeeld een IP-adres een persoonsgegeven zijn.

4.2.1 Wat is het doel van het gebruik van persoonsgegevens in het algoritme?

Als we persoonsgegevens gebruiken, moet er een duidelijk doel zijn. Persoonsgegevens mogen alleen worden gebruikt als dat echt noodzakelijk is. Wees je bewust dat het gebruik ervan risico's op bias met zich meebrengt. Gebruik waar mogelijk informatie uit de Privacy Quick Scan / DPIA.
Beschrijf per persoonsgegeven het doel daarvan voor dit algoritme. Bekijk kritisch of het doel van het algoritme ook kan worden bereikt zonder het gebruik van bepaalde persoonsgegevens.
Deze analyse is specifieker dan in de DPIA, waar je alleen naar het algemene doel kijkt. Door nu per persoonsgegeven te bepalen of het gebruik doelmatig is, kun je bewuster de keuze nemen om persoonsgegevens wel of niet te gebruiken.

Te gebruiken persoonsgegevens	Doel gebruik persoonsgegeven in het algoritme: waarom gebruik je dit specifieke persoonsgegeven en is dit echt noodzakelijk?
<< Vul in >>	<< Vul in >>
<< Vul in >>	<< Vul in >>
<< Vul in >>	<< Vul in >>
<< Vul in >>	<< Vul in >>
<< Vul in >>	<< Vul in >>

Vaak worden in een proces en algoritme meerdere (impliciete) beslissingen genomen voordat er een formeel besluit volgt



Proces/algoritme

4.2.2 Hoe leg je aan (groepen) personen uit welke beslissingen het algoritme neemt?

Een algoritme neemt vaak meerdere beslissingen om tot een uitkomst te komen. Die beslissingen moeten we betekenisvol kunnen uitleggen aan betrokkenen, zoals de burger. Betekenisvolle verantwoording betekent een transparante uitleg over de logica voor het nemen van de beslissing, zodat voor de betrokkenen duidelijk is hoe de beslissing tot stand is gekomen.

Voorbeeld: casus Mokumse Bos

In een brainstorm om het algoritme breder in te zetten, is het idee opgevat om het monitoringsalgoritme aan te passen om prestaties van boswachters te meten. Als het idee geïmplementeerd wordt, krijgt de manager een mail dat één van de boswachters slecht zou presteren met de vraag om hierover een disciplinair gesprek te voeren met de medewerker. De manager is van mening dat zonder concrete toelichting er geen basis is om een disciplinaire gesprek te voeren. Daarom wordt het systeem aangepast zodat het duidelijk uitlegt waarom de prestaties slecht zijn, waarna de manager zelf kan bepalen of een gesprek nodig is.

<< Vul in >>

4.2.3 Kan een betrokkene redelijkerwijs verwachten dat de persoonsgegevens in het algoritme worden gebruikt? En dat hier een bepaalde mate van bias in zit?

Als een gemiddelde betrokkene (burger/ondernemer/bezoeker) niet redelijkerwijs kan verwachten dat zijn persoonsgegevens in het algoritme worden gebruikt met een bepaalde mate van bias, kan dat een indicatie zijn dat de verwerking onbehoorlijk is. Evalueer of de verwerkingen en de bias behoorlijk en maatschappelijk te rechtvaardigen zijn: als er morgen een krantenartikel over deze verwerkingen en de mate van bias wordt gepubliceerd, zou de gemiddelde betrokkene dan verbaasd zijn of de gemeente onbehoorlijk vinden?

<< Vul in >>

4.3 Conclusie Juridische Bias Impact Assessment

4.3.1 Verwachten we dat we het algoritme mogen gebruiken gebaseerd op de Juridische Bias Impact Assessment?

Vorm een conclusie gebaseerd op paragraaf 4.1 en 4.2. Maak uiteindelijk een integrale afweging in hoofdstuk 1.

Ja, het algoritme kan worden gebruikt in de context van het discriminatieverbod, want we gebruiken alleen (proxy) variabelen waarmee:

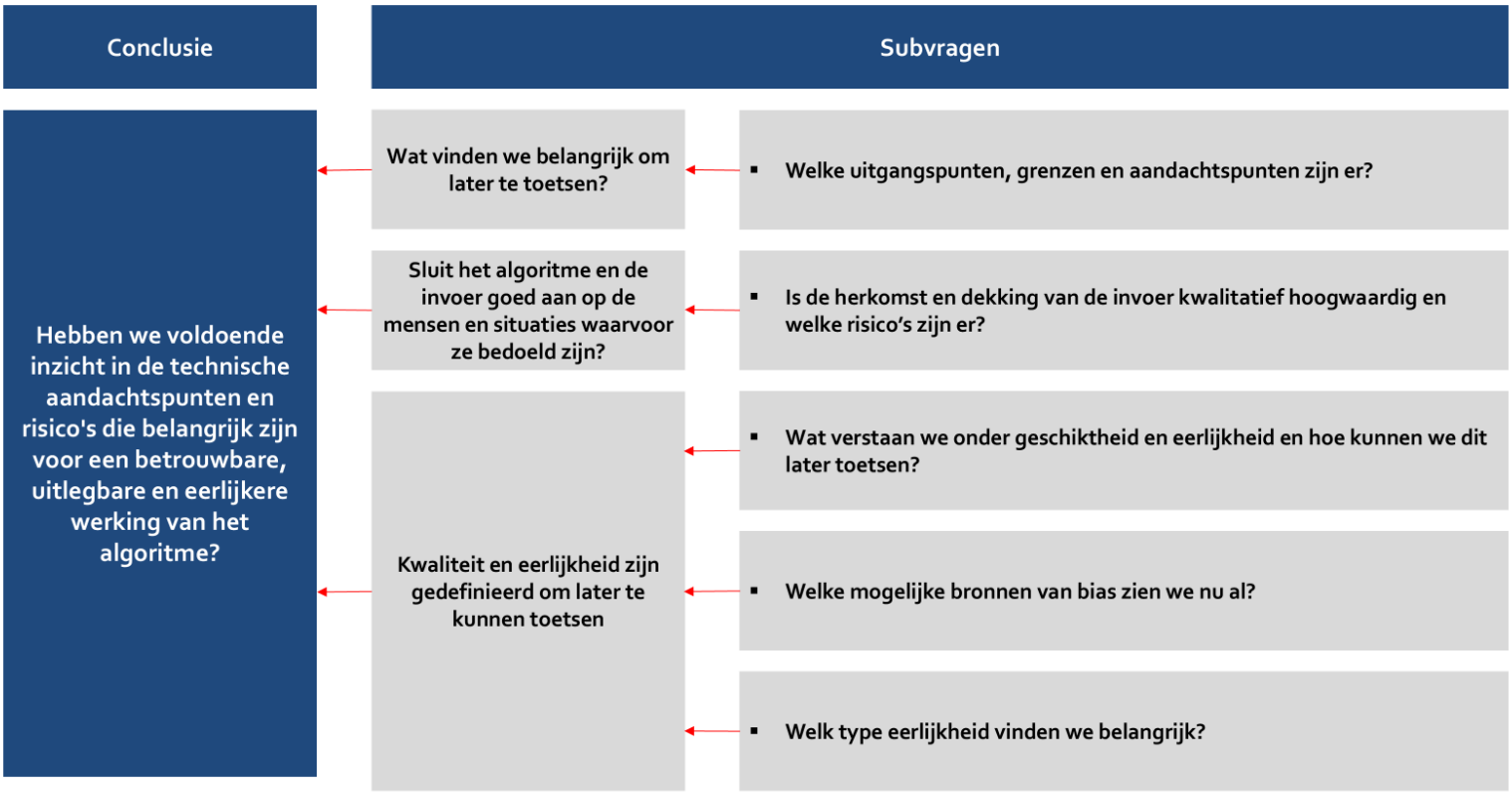
1. **geen direct onderscheid gemaakt wordt en/of wel direct onderscheid gemaakt wordt maar op basis van een wettelijke uitzondering (zie tabel in paragraaf 4.1); en**
2. **indirect onderscheid gerechtvaardigd is en/of er geen indirect onderscheid is (zie tabel in paragraaf 4.1)**

en persoonsgegevens worden rechtvaardig en behoorlijk gebruikt (paragraaf 4.2).

We gaan << wel/niet >> door met het algoritme en nemen onderstaande maatregelen:

- **<< Vul maatregelen in >>**

5 Technische Bias Impact Assessment: kan het?



Bij de Technische Bias Impact Assessment staat de vraag centraal: kan het? Deze vraag moet je stimuleren om bewust te zijn dat het technisch mogelijk moet zijn om de bias te meten en beperken. Hierbij bouw je voort op de Ethische- en Technische Bias Impact Assessment.

Een Technische Bias Impact Assessment kun je niet altijd volledig uitvoeren in de vroege fase van een algoritme. Soms is de data er nog niet, of is het algoritme nog niet ontworpen. Toch kun je juist in deze fase al belangrijke keuzes maken die later bepalen of het systeem eerlijk, betrouwbaar en passend wordt ingezet. In deze Technische Bias Impact Assessment onderzoeken we daarom welke waarden we belangrijk vinden voor het algoritme en/of de data en eventuele regels goed aansluiten op de mensen en situaties waarvoor het systeem bedoeld is. De antwoorden op de hieronder gestelde vragen helpen om in een later stadium zogenoemde "fairness metrics" – vrij vertaald "eerlijkheidsmaatstaven"[8] – vast te laten leggen die technisch getoetst kunnen worden.

[8] Voor meer informatie, zie The Fairness Handbook. Daarnaast zijn er veel (open source) instrumenten beschikbaar, zoals de Unsupervised bias detectie tool die informeert welke afwijkingen nader onderzocht moeten worden door domeinexperts.

Nu bepalen we dus wat eerlijk zou moeten zijn, en in een gedetailleerde bias-analyse testen we of het algoritme dit ook echt doet.

In een gedetailleerdere technische bias-analyse zal, door eventuele data en het algoritme technisch te analyseren, getoetst worden of:

- het model presteert zoals afgesproken (kwaliteit);
- verschillende groepen eerlijk worden behandeld;
- er geen onverwachte nieuwe risico's ontstaan tijdens gebruik;
- oneerlijke uitkomsten tijdig worden herkend en gecorrigeerd; en
- beslissingen uitgelegd en verantwoord kunnen worden.

Deze gedetailleerde technische bias-analyse wordt in de loop van 2026 onderdeel van Quality Assurance[9]. Tot die tijd kan een data analist / data scientist je verder helpen, bijvoorbeeld aan de hand van het AI Control framework.

5.1 Wat vinden we belangrijk om later te toetsen?

Voordat een algoritme wordt ontwikkeld, is het belangrijk dat de proceseigenaar samen met het projectteam bepaalt wat echt belangrijk is voor een eerlijk en verantwoord systeem. Niet in technische termen, maar in gewone taal: wat moet dit algoritme straks goed doen, waar moet op gelet worden en wat verstaan we onder eerlijkheid?

Door dit nu vast te leggen, geeft de proceseigenaar met het projectteam richting aan de keuzes die later gemaakt worden. Het algoritme wordt dan niet alleen gebouwd om "technisch goed te werken", maar ook om eerlijk en verantwoord te werken volgens de waarden die je vooraf belangrijk vindt.

Door eerlijkheid vanuit een technische invalshoek concreet te maken, kan later getoetst worden of het algoritme inderdaad aansluit bij wat het projectteam belangrijk vindt. Om erachter te komen wat eerlijkheid voor het projectteam inhoudt, gaan we eerst na hoe we dit kunnen begrenzen en wat eventuele aandachts- en uitgangspunten zijn.

In deze stap onderzoeken we drie dingen:

1. Uitgangspunten: welke waarden en principes vinden we belangrijk voor eerlijkheid?
2. Grenzen: wat mag het algoritme nooit doen?
3. Aandachtspunten: welke risico's zien we nu al waar later extra op gelet moet worden?

[9] Quality Assurance - Kwaliteitsborging

5.1.1 Welke uitgangspunten, grenzen en aandachtspunten zijn er?

Uitgangspunten zijn waarden en principes die het projectteam belangrijk vindt voor eerlijkheid in de context van dit project. Denk hierbij aan het doel uit vraag 3.1.1 en aan de groepen personen die door het algoritme geraakt worden. Grenzen zijn de afspraken die het projectteam vooraf maakt over wat niet acceptabel is. Dat kunnen no-go's zijn of aandachtspunten. Op aandachtspunten moet extra scherp gelet worden tijdens de ontwikkeling, het gebruik van een algoritme en de monitoring, omdat ze een risico vormen voor eerlijkheid of betrouwbaarheid. Besteed extra aandacht aan de eerder gedefinieerde kwetsbare groepen uit vraag 3.2.1.

Voorbeeld: casus Mokumse Bos

Uitgangspunten:

- Het model moet alle diersoorten even goed herkennen, ook zeldzame soorten.
- De beslissingen (bijvoorbeeld categorisering) moeten uitlegbaar zijn aan boswachters.

Grenzen:

- Het algoritme mag nooit zelfstandig vastleggen dat een beschermde soort aanwezig is, ook al betekent dit dat boswachters tijd kwijt zijn aan controle van het algoritme.
- Het algoritme mag niet ingezet worden voor handhaving (bv. om loslopende honden te detecteren), ook al zou dat technisch mogelijk zijn.
- Het model mag nooit een gevaarlijke diersoort als afwezig rapporteren als dat leidt tot risico voor mens of dier.

Aandachtspunten:

- Sommige soorten komen weinig voor in de dataset (bv. otters en vossen).
- De wolf is in opkomst. Hoewel deze in Amsterdam nog niet gespot is, moeten we hier in de toekomst mogelijk rekening mee houden.
- We moeten erop letten dat vrijwilligers en burgers begrijpen dat het AI-systeem alleen ondersteunt, omdat er anders verkeerde verwachtingen ontstaan.

Uitgangspunten: welke waarden en principes zijn belangrijk voor eerlijkheid in de context van dit project? Denk hierbij ook aan het eerder gedefinieerde doel in vraag 3.1.1.

1. De volgende uitgangspunten vinden we belangrijk:
   << Vul in >>

2. De uitgangspunten passen bij het geformuleerde doel in vraag 3.1.1 omdat:
   << Vul in >>

3. Een mogelijke keerzijde hiervan is:
   << Vul in >>

Grens: wat mag dit algoritme nooit doen.

Het algoritme mag nooit << vul grenzen in >>, ook al betekent dit dat << vul in >>.

Aandachtspunt: welke risico's zien we nu al. Denk hierbij ook aan de eerder gedefinieerde kwetsbare groepen uit vraag 3.1.1 en de gebruikte persoonsgegevens uit vraag 4.2.1.

We moeten extra opletten dat << vul aandachtspunten in >>, omdat << vul in >>.

5.2 Sluit het algoritme en de invoer goed aan op de mensen en situaties waarvoor ze bedoeld zijn?

Het is belangrijk dat het projectteam al vroeg in het proces nadenkt over de invoer die het algoritme gebruikt. Dit kan trainings- of productiedata zijn, maar ook parameters of gebruikersprompts. Als die invoer niet goed aansluit op de werkelijkheid van de doelgroep, dan kan het algoritme niet eerlijk of betrouwbaar werken.

Veelvoorkomende problemen zijn bijvoorbeeld:

- bepaalde groepen of situaties komen nauwelijks voor in de data[10];
- de data is voor een ander doel verzameld dan waarvoor het algoritme wordt ingezet;
- de doelvariabele (wat het algoritme probeert te voorspellen of bepalen) sluit niet goed aan bij wat je eigenlijk wilt weten; en
- bij GenAI-modellen: beperkte transparantie over trainingsdata.

Om dit te achterhalen hoef je niet direct diepe technische analyses te doen. Vaak kun je al een eerste indruk krijgen door te kijken welke invoer is gebruikt en hoe die tot stand is gekomen, welke groepen of situaties geraakt worden door het algoritme en of deze voldoende terugkomen in de invoer.

Voorbeeld: casus Mokumse Bos

Als de dataset vooral foto's van overdag bevat, leert het algoritme vooral over dieren die overdag actief zijn. Wellicht zijn juist veel dieren 's nachts actief. Hierdoor leert het algoritme niet goed genoeg over nachtdieren, waardoor het niet goed zal werken voor het algemene doel. In het onderstaande voorbeeld worden naast de terugkerende Mokumse Bos casus (machine learning (ML)) voor de volledigheid ook voorbeelden gegeven als er rekenregels en generatieve AI-algoritmes (GenAI) gebruikt zouden worden.

[10] Indien interne gegevens niet beschikbaar zijn, wordt overwogen om het Centraal Bureau voor de Statistiek (CBS) te benaderen voor een maatwerkverzoek.

5.2.1 Is de herkomst en dekking van de invoer kwalitatief hoogwaardig en welke risico's zijn er?

De onderstaande tabel helpt om systematisch te beoordelen of de invoer van het algoritme goed aansluit op de mensen en situaties waarvoor het bedoeld is. Door per onderdeel kort te beschrijven wat bekend is, waar onzekerheden zitten en welke groepen of situaties mogelijk ontbreken, ontstaat een eerste beeld van de risico's op oneerlijke of onbetrouwbare uitkomsten. Vul in wat op dit moment al bekend is en noteer eventuele aannames.

Toelichting GenAI-systemen: Vaak is niet precies bekend welke trainingsdata gebruikt is. Vul dan vooral in wat je wel weet en maak aannames expliciet (bv. systeem is getraind op publieke tekstbronnen met vooral Westers materiaal).

	Vraag	Voorbeeld: casus Mokumse Bos	In de context van ons algoritme:
Herkomst	Waar komt de invoer vandaan? Wat was het oorspronkelijke doel van die data/regels/prompts? Past dat doel bij het gewenste gebruik van het algoritme?[11]	*ML:* Camerabeelden uit het Mokumse Bos, bedoeld voor ecologisch onderzoek. *Rule-based:* Door boswachters vastgestelde regels om het algoritme het aantal dieren te laten tellen (bijv. alle dieren onder de 30 cm worden als klein geregistreerd). *GenAI:* Teksten uit universitaire dier-onderzoeken, bedoeld om het model mee te trainen.	<< Vul in >>
Dekking	Welke groepen of situaties komen terug in de invoer? Welke zijn onder- of oververtegenwoordigd?	*ML:* Veel eekhoorns en uilen, nauwelijks buizerds en vossen. *Rule-based:* De regels gaan vooral uit van veelvoorkomende soorten zoals konijnen, bijzondere gevallen vallen buiten de gedefinieerde categorieën. *GenAI:* Trainingsmateriaal bevat veel teksten over muizen, ratten en andere kleine dieren, maar weinig over grote zoogdieren. Hierdoor zijn sommige soorten minder goed beschreven.	<< Vul in >>
Risico	Waar zie je nu al dat de invoer mogelijk niet volledig of evenwichtig is, waardoor sommige groepen of situaties minder goed meegenomen worden?	*ML:* Veel beelden uit open velden, weinig uit dicht beboste delen, hierdoor worden dieren in struiken en bomen mogelijk slechter herkend. *Rule-based:* De regels houden geen rekening met uitzonderingen (zoals jonge kleine dieren, of nieuwe soorten in het gebied). Hierdoor worden deze categorieën mogelijk systematisch foutief gelabeld. *GenAI:* Het model kan soorten die weinig voorkomen in trainingsdata slechter beschrijven of fout classificeren.	<< Vul in >>

Schrijf je gezamenlijke conclusie op in één zin.

Onze invoer en uitgangspunten zijn << wel/niet >> geschikt, omdat << vul in >>. We zien het risico dat << vul in >>. We nemen << vul in >> maatregelen zodat de data op de situatie aansluit.

[11] Wil je kunstmatig gegenereerde data? Een voorbeeld van een open source instrument voor het genereren van synthetische data is de Synthetische data generatie tool.

5.3 Geschiktheid en eerlijkheid zijn gedefinieerd om later te kunnen toetsen

Wanneer duidelijk is wat het algoritme moet doen en voor wie, is het belangrijk om vast te leggen wat 'goed' en 'eerlijk' werken betekent. Dit helpt om verwachtingen tussen beleid, uitvoering en techniek helder te krijgen en voorkomt dat achteraf blijkt dat iedereen iets anders bedoelde met "het systeem werkt goed".

In deze fase hoeft nog niet gemeten of bewezen te worden. Het doel is om gezamenlijk te verkennen wat geschiktheid en eerlijkheid betekent in de context van het project, te benoemen welke signalen aangeven dat het systeem niet meer aan die waarden voldoet en af te spreken hoe en wanneer dat later getoetst kan worden.

5.3.1 Wat verstaan we onder geschiktheid en eerlijkheid, en hoe kunnen we dit later toetsen?

	Vraag	Voorbeeld: casus Mokumse Bos	In de context van ons algoritme:
Geschiktheid	Beschrijf wat het systeem volgens jou systematisch goed moet doen om betrouwbaar te zijn.	Het systeem moet alle soorten dieren in 95% van de gevallen correct herkennen onder verschillende licht- en weersomstandigheden.	<< Vul in >>
Eerlijkheid	Beschrijf wat "eerlijk" betekent in jouw context. Voorbeelden: gelijke behandeling van groepen, transparante besluitvorming, gelijke kans op correcte beoordeling.	Het model mag geen diersoorten systematisch overslaan. Ook zeldzame dieren moeten herkend worden.	<< Vul in >>
Signalen	Noteer welke signalen erop zouden wijzen dat het systeem niet meer eerlijk of betrouwbaar werkt.	Als bepaalde diersoorten bijna nooit worden herkend of steeds fout worden geclassificeerd.	<< Vul in >>

5.3.2 Welke mogelijke bronnen van bias zien we nu al?

Bias vindt niet alleen plaats in de data, maar kan ook gebeuren doordat regels, beelden of teksten een bepaald perspectief benadrukken en andere uitsluiten. Kruis aan wat van toepassing is en geef een kort voorbeeld voor dit project (meerdere opties zijn mogelijk).

Mogelijke bron van bias	Voorbeelden	In de context van ons algoritme:
☐ In de data	Historische data waarin bepaalde groepen nauwelijks of niet voorkomen of de labels bevatten vooroordelen.	<< Vul in >>
☐ In de regels of drempelwaarden	Een specifieke inkomensgrens waardoor studenten buiten een regeling vallen.	<< Vul in >>
☐ In de output of de manier waarop deze gepresenteerd wordt	Een taalmodel dat vaker vrouwelijke voorbeelden gebruikt of mannen in een negatief daglicht plaatst.	<< Vul in >>
☐ In de interpretatie door mensen	Medewerkers die de uitkomst van het systeem als waarheid zien en niet meer zelf beoordelen.	<< Vul in >>
☐ In het zelflerende vermogen van het algoritme	Het algoritme leert van de data die het gebruikt. Als die data vooroordelen bevat, kan het algoritme die overnemen en versterken.	<< Vul in >>
☐ Anders, namelijk << Vul in >>	…	<< Vul in >>

5.3.3 Welk type eerlijkheid vinden we belangrijk?

Gebruik de dilemma's hieronder om te bespreken wat eerlijkheid betekent voor je eigen algoritme, doelgroep en situatie. Er zijn geen goede of foute antwoorden, het gaat erom waar het team de nadruk op wil leggen. Dit laat zien hoe verschillende interpretaties van eerlijkheid eruit kunnen zien in de praktijk. Elke keuze is gekoppeld aan concrete vervolgstappen en punten die extra aandacht behoeven in ontwerp, testen en monitoring.

Dilemma's (voorbeeld: casus Mokumse Bos)	Keuze in ons project en vervolgstappen
Dilemma 1: Moet het algoritme alle groepen gelijk behandelen, of juist minder vertegenwoordigde groepen compenseren?	Wat past bij het doel van ons algoritme?
Gelijke kans Je wilt een objectief beeld krijgen van hoeveel dieren er in het bos leven en hoe de biodiversiteit is verdeeld. In dit geval is het belangrijk dat het model alle soorten (zoals vos, ree en otter) even goed kan herkennen, maar het is niet erg als zeldzame soorten, zoals de otter, minder vaak worden vastgelegd. Dit geeft een realistisch beeld van de werkelijke populatieverhoudingen in het bos.	☐ **Gelijke kans** **Actie:** Zorg dat het algoritme alle groepen even goed kan herkennen of behandelen. Zorg dat de data evenwichtig is samengesteld, zodat alle categorieën of groepen voldoende vertegenwoordigd zijn tijdens de ontwikkeling van het algoritme. Daarnaast vraagt dit om het testen van de modeluitkomsten: presteert het model daadwerkelijk even goed voor alle groepen? **Aandachtspunt:** Ook als de trainingsdata evenwichtig is, kan het model alsnog verschillende prestaties laten zien voor verschillende groepen. Minder voorkomende categorieën of groepen kunnen ondervertegenwoordigd zijn in de data, waardoor het algoritme ze minder goed leert herkennen of behandelen. Dit moet later expliciet worden gecontroleerd in de gedetailleerde technische bias-analyse.
Gelijke uitkomst Je wilt inzicht krijgen in welke diersoorten extra hulp nodig hebben, bijvoorbeeld om de winter te overleven. In dit geval is het belangrijk dat zeldzame soorten, zoals de otter, voldoende zichtbaar zijn in de resultaten, zelfs als ze minder vaak voorkomen in de data. Dit betekent dat je het model moet aanpassen om deze soorten extra te benadrukken.	☐ **Gelijke uitkomst** **Actie:** Pas het algoritme aan zodat minder vertegenwoordigde groepen een hogere kans krijgen op een correcte herkenning. Denk aan: aangepaste drempelwaarden voor specifieke categorieën, extra trainingsrondes voor kwetsbare groepen, het toekennen van hogere gewichten tijdens training. **Aandachtspunt:** Het benadrukken van minder voorkomende categorieën of groepen kan leiden tot een vertekend beeld van de werkelijkheid. Zeldzame groepen worden zo relatief vaker gedetecteerd dan ze eigenlijk in de werkelijkheid voorkomen. **Toelichting keuze (context, risico, afweging)?** << Vul in >>
Dilemma 2: Hoe gaan we om met onzekerheid in de uitkomsten van een algoritme? Waar ligt de nadruk: fouten vermijden (focus op betrouwbaarheid) of signalen niet missen (focus op volledigheid)?	Waar ligt binnen ons project de focus?
Fouten vermijden Het algoritme geeft alleen waarnemingen door als het vrij zeker is van de herkenning. Bijvoorbeeld als je 95% zeker bent dat er een otter is gevonden (gevolg: je mist er misschien een paar). • **Voordeel:** Minder kans op foutieve meldingen. • **Nadeel:** Mogelijk worden sommige waarnemingen niet doorgegeven, ook als ze relevant zijn.	☐ **Fouten vermijden** **Actie:** Stel strengere drempelwaarden in, voeg menselijke controle toe en communiceer de mate van zekerheid bij elk resultaat. Test het algoritme grondig en stel duidelijke grenzen voor acceptabele foutmarges.
Signalen niet missen Het algoritme geeft zoveel mogelijk waarnemingen door, ook als er enige onzekerheid is. Bijvoorbeeld als je waarnemingen ook doorgeeft bij 60% zekerheid. Dit heeft als gevolg dat de boswachter vaker moet controleren.	☐ **Signalen niet missen** **Actie:** Stel lagere drempelwaarden in om meer resultaten door te geven, bouw extra controlemechanismen in en zorg voor een feedbackloop om het algoritme te verbeteren.

Dilemma's (voorbeeld: casus Mokumse Bos)	Keuze in ons project en vervolgstappen
• **Voordeel:** Meer kansen om beschermde soorten te detecteren. • **Nadeel:** De boswachter moet meer tijd besteden aan het controleren van foutieve meldingen.	**Toelichting keuze (context, risico, afweging)?** << Vul in >>
Dilemma 3: Moeten we iedereen op dezelfde manier behandelen, of vinden we dat we mogen afwijken in bepaalde situaties? In het kort: vinden we uniformiteit of maatwerk belangrijker? **Uniformiteit (iedereen hetzelfde)** Alle beelden worden met dezelfde instellingen en regels verwerkt, ongeacht de omstandigheden, om objectieve vergelijking mogelijk te maken. • **Voordeel:** Consistente en objectieve verwerking, eenvoudiger te implementeren en te beheren. • **Nadeel:** Specifieke of afwijkende situaties (zoals uitzonderlijke gevallen of complexe data) kunnen minder goed worden verwerkt. **Maatwerk (aanpassen bij verschil)** Nachtbeelden of zeldzame soorten krijgen extra controles, aparte modellen of regels, om te voorkomen dat zij gemist worden. • **Voordeel:** Betere prestaties in unieke of uitdagende situaties. • **Nadeel:** Complexere systemen, meer onderhoud en mogelijk minder consistentie tussen verschillende datatypen of situaties.	Waar ligt binnen ons project de focus? ☐ **Uniformiteit** **Actie:** Zorg voor een evenwichtige dataset en test het systeem op verschillende soorten data en groepen om te controleren of de prestaties acceptabel zijn. Monitor consistentie en communiceer dat er geen onderscheid wordt gemaakt. ☐ **Maatwerk** **Actie:** Identificeer kwetsbare groepen en ontwikkel specifieke regels of mechanismen om hen beter te herkennen. Monitor de impact van deze aanpassingen en communiceer transparant over de keuze om extra bescherming te bieden. **Toelichting keuze (context, risico, afweging)?** << Vul in >>

5.4 Conclusie Technische Bias Impact Assessment

Hebben we voldoende inzicht in technische aandachtspunten en risico's om het algoritme door te ontwikkelen?

Vorm een conclusie gebaseerd op paragraaf 5.1, 5.2 en 5.3. Je hebt voldoende inzicht als je bij alle vragen een keuze hebt gemaakt.

<< Ja/Nee >> we hebben voldoende inzicht in de technische aandachtspunten en risico's die belangrijk zijn voor een betrouwbare, uitlegbare en eerlijkere werking van het algoritme.

We gaan << wel/niet >> door met het algoritme en nemen onderstaande maatregelen:

- **<< Vul maatregelen in >>**

Ga terug naar hoofdstuk 1 Conclusie Bias Impact Assessment en maak een integrale afweging op basis van de Ethische-, Juridische- en Technische Bias Impact Assessments of de bias in het algoritme acceptabel is om door te gaan met de ontwikkeling.

Bijlage 1. Definities

	Definitie
Bias	Een systematisch verschil in behandeling van mensen of groepen in vergelijking met anderen.
Directe discriminatie	Een behandeling van een persoon of groep personen op andere wijze dan een ander in een vergelijkbare situatie op grond van een beschermd persoonskenmerk (discriminatiegrond) zonder dat hiervoor een objectieve rechtvaardiging bestaat.
Discriminatiegrond	Juridisch beschermde persoonskenmerken waarop (behoudens objectieve rechtvaardiging) geen onderscheid gemaakt mag worden gemaakt. Voorbeelden van discriminatiegronden zijn 'ras', nationaliteit, religie, geslacht, seksuele gerichtheid, handicap of chronische ziekte.
Ethische Bijsluiter	Een instrument dat teams helpt om waarden en dilemma's expliciet te maken. Hierdoor kan je rekening houden met de rechten, belangen en wensen van alle betrokkenen. Zo creëert de Ethische Bijsluiter transparantie over de morele implicaties van technologie die we inzetten als gemeente. Zie voor meer informatie Quality Assurance.
Etniciteit	Een sociale categorie die verwijst naar het (verondersteld) behoren tot een Sociale of maatschappelijke groep, die gekenmerkt wordt door een gemeenschappelijke nationaliteit, afstamming, religie, taal of culturele en traditionele afkomst en/of achtergrond. Etniciteit valt juridisch onder de discriminatiegrond 'ras'.
Fairness metrics	"Eerlijkheidsmaatstaven" zijn wiskundige hulpmiddelen die worden gebruikt om vooringenomenheid te kwantificeren en te meten hoe eerlijk een algoritme verschillende groepen behandelt.
Indirecte discriminatie	Een ogenschijnlijk neutrale bepaling, maatstaf of handelwijze die personen met een bepaald beschermd persoonskenmerk (discriminatiegrond) in vergelijking met andere personen in het bijzonder benadeelt, zonder dat hiervoor een objectieve rechtvaardiging bestaat.
Nationale afstamming of herkomst	Verbondenheid met een nationale staat, bijvoorbeeld omdat de persoon in kwestie of diens ouders daar is geboren of opgegroeid zijn. Nationale afstamming (of 'herkomst') valt onder de discriminatiegrond 'ras' en is iets anders dan het juridische begrip 'nationaliteit'.
Nationaliteit	Juridische status van een persoon als staatsburger van een staat. Beschrijft derhalve de juridische band tussen een individu en een staat, waar over en weer rechten en plichten aan ontleend kunnen worden.
Profilering	Elke vorm van (geautomatiseerde) verwerking van persoonsgegevens waarbij aan de hand van persoonsgegevens bepaalde persoonlijke aspecten van een natuurlijke persoon worden geëvalueerd, met name met de bedoeling zijn beroepsprestaties, economische situatie, gezondheid, persoonlijke voorkeuren, interesses, betrouwbaarheid, gedrag, locatie of verplaatsingen te analyseren of te voorspellen.
Quality Assurance	Quality Assurance (QA), ofwel kwaliteitsborging, is een gestructureerde aanpak om ervoor te zorgen dat onze digitale producten, diensten en processen voldoen aan relevante wetten, regels en normen én de verwachtingen van de gebruikers. Voor meer informatie, zie: Quality Assurance - Kwaliteitsborging.
Ras	Juridische verzamelterm voor persoonskenmerken als huidskleur, etniciteit en nationale of etnische herkomst. 'Ras' is een discriminatiegrond.
Risicoprofilering	Profilering (zie definitie) die plaatsvindt aan de hand van algemene, verondersteld voorspellende (risico)criteria, die geacht worden te correleren met de kans op overtreding van een bepaalde norm, bij afwezigheid van concrete of specifieke informatie over feitelijke normovertredingen of daders.
Variabele	In de context van algoritmes is een variabele een opslag wat wordt gebruikt om gegevens op te slaan en te manipuleren. Variabelen kunnen verschillende typen waarden bevatten, zoals getallen, tekst of objecten, afhankelijk van het algoritme en de context.