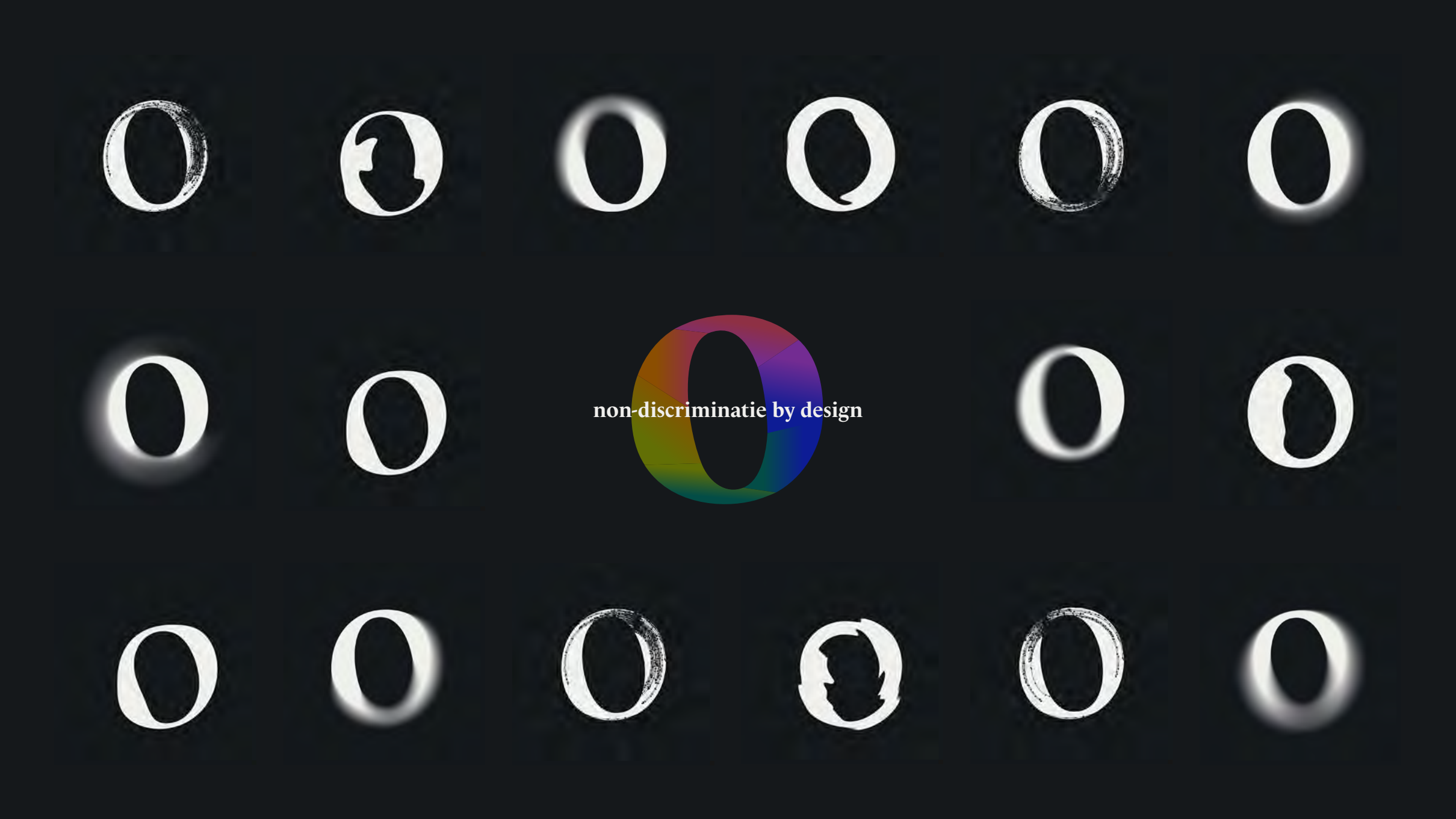
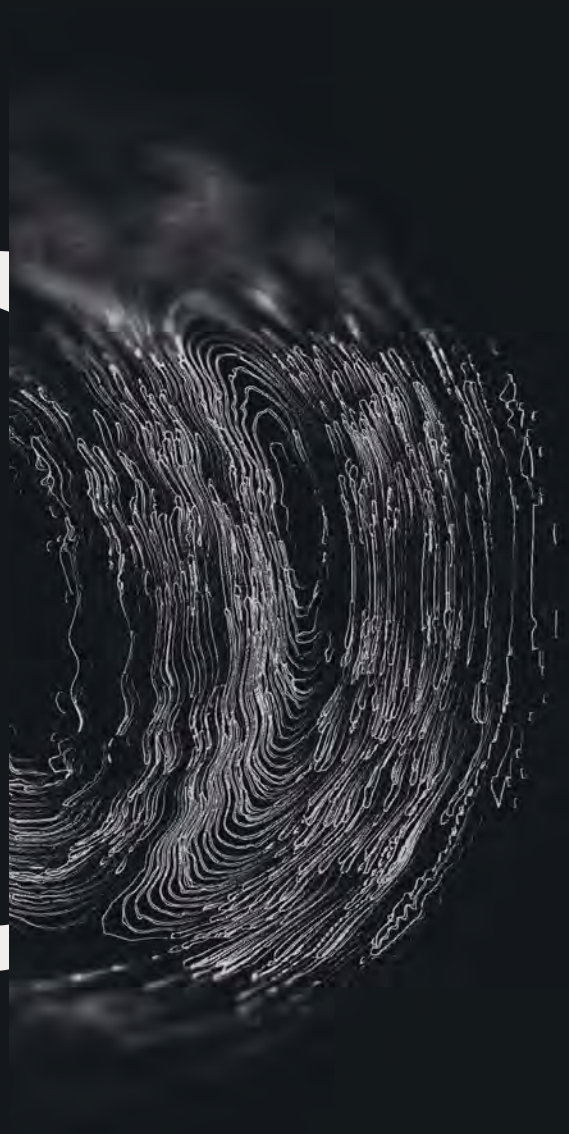

non-discriminatie by design



Over deze handreiking

Dit document legt uit welke vragen en principes leidend zijn bij het ontwikkelen en implementeren van een AI-systeem met het oog op het discriminatieverbod, vanuit zowel juridisch, technisch, als organisatorisch perspectief. Het document is bedoeld voor projectleiders die sturing geven aan systeembouwers, data-analisten en AI-experts. Stel je wilt een AI-systeem zo non-discriminatoir mogelijk maken, waar moet je dan aan denken en welke discussies moet je binnen je team voeren?

Over het ontwerp van deze handreiking

Het visueel concept van dit document is ontwikkeld om de leesbaarheid en herkenbaarheid van de inhoud te optimaliseren. De basis van dit concept is de letter ‘o’ die je door middel van kleur- en vormveranderingen begeleidt bij het lezen van dit document. Op de volgende pagina (de inhoudsopgave) maak je voor het eerst kennis met de verschillende vormen en kleuren. Op pagina 18 en 19 wordt dit nader toegelicht.

Inleiding

Artikel 1 van de Nederlandse Grondwet verbiedt discriminatie: *'Allen die zich in Nederland bevinden, worden in gelijke gevallen gelijk behandeld. Discriminatie wegens godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht of op welke grond dan ook, is niet toegestaan.'* Non-discriminatie is dus de basis van ons rechtssysteem en onze samenleving.

De afgelopen jaren is duidelijk geworden dat ook AI-systemen discriminatoire effecten kunnen hebben. Denk aan een gezichtsherkenningssysteem dat niet goed werkt voor mensen met een donkere huidskleur, een vertaaldienst die stereotyperende teksten genereert, of een CV selectiesysteem dat een ongefundeerde voorkeur heeft voor mannelijke kandidaten. Hoe kunnen we systemen ontwikkelen die zo min mogelijk onbedoeld en ongerechtvaardigd onderscheid maakt tussen groepen mensen?

Het non-discriminatierecht geeft uitgangspunten en vragen en geen ge- of verboden. Vaak is de hoop dat juristen 'gewoon zeggen' wat wel en wat niet mag, maar zo werkt het recht niet. Het recht geeft een standaard, een uitgangspunt of een principe, maar daarop bestaan altijd uitzonderingen. Bovendien zal een rechter altijd rekening houden met de context, of wat juristen noemen 'de omstandigheden van het geval'.

Onderscheid maken op basis van geslacht is bijvoorbeeld verboden, tenzij dit een relevante factor is. Zoekt een castingbureau een actrice om de vrouwelijke hoofdrol op zich te nemen, dan mag het natuurlijk mannen uitsluiten. Sterker nog, positieve discriminatie kan in sommige gevallen geoorloofd zijn; als een organisatie bijvoorbeeld met name mannelijke werknemers heeft, dan mag die besluiten dat bij gelijke geschiktheid aan vrouwen de voorkeur wordt gegeven.

Ook de vraag wat nu precies 'onderscheid maken' is, is niet eenduidig te beantwoorden. Het uitgangspunt in het recht is dat gelijke gevallen gelijk behandeld moeten worden, en ongelijke gevallen ongelijk. Maar wat is precies gelijk en wat niet? Welke factoren zijn relevant voor het

besluit en welke niet? Al dat soort vragen kunnen niet in het algemeen worden beantwoord in deze handreiking, omdat ze afhangen van de context en de vraag hoe een AI-systeem functioneert, waartoe het dient en welke waarborgen er zijn getroffen.

Belangrijk is dat het recht niet alleen uitgaat van een beperkt aantal gronden waarop in principe geen besluiten mogen worden genomen, zoals ras, geslacht of geaardheid, maar dat het ook discriminatie 'op welke grond ook' verbiedt. Dit vergt van een systeembouwer dat die zich goed bewust is op basis van welke groepen het AI-systeem onderscheid maakt en of dat onderscheid te rechtvaardigen is. Is het onderscheid bedoeld of onbedoeld; is het relevant of niet?

De grondwettelijke non-discriminatiebepaling is nader uitgewerkt in verschillende wetten, zoals:

- *De Algemene Wet Gelijke Behandeling*
- *Wet gelijke behandeling op grond van handicap of chronische ziekte*
- *Wet gelijke behandeling op grond van leeftijd*
- *Wet gelijke behandeling van mannen en vrouwen*
- *Wet onderscheid arbeidsduur*
- *Wet onderscheid bepaalde en onbepaalde tijd*

Non-discriminatiebeginselen

Directe discriminatie

Binnen het non-discriminatierecht is niet alleen 'directe discriminatie' verboden, maar ook 'indirecte discriminatie'. Van directe discriminatie is sprake indien een persoon op een andere wijze wordt behandeld dan een ander in een vergelijkbare situatie.

In een vacaturetekst wordt vermeld dat enkel vrouwen in aanmerking komen voor een baan. Dit is directe discriminatie op grond van geslacht, omdat mannen niet mogen solliciteren.

Indirecte discriminatie

Van indirecte discriminatie is sprake indien een ogenschijnlijk neutrale bepaling, maatstaf of handelwijze een groep personen in vergelijking met andere groepen personen bijzonder treft.

> Een systeem dat automatisch advertenties plaatst voor gemeubileerde huurwoningen in Nederland geeft aan dat enkel 'expats' in aanmerking komen voor een huurovereenkomst. Dit zijn weinig problematische huurders gebleken. Uitgaande van de definitie dat een expat iemand is die tijdelijk in het buitenland woont, is dit criterium indirect discriminerend naar nationaliteit: er zullen wel Nederlanders zijn die aan het criterium voldoen, maar over het algemeen zullen expats voornamelijk een niet-Nederlandse nationaliteit hebben. Personen met een Nederlandse nationaliteit worden dus in het bijzonder uitgesloten door zo een criterium.

Directe discriminatie is vaak vrij duidelijk zichtbaar. Zoals gezegd is dit niet per definitie verboden, maar meestal wel. Bij indirecte discriminatie is het moeilijker vast te stellen of hier in een bepaald geval sprake van is en zo ja, of dat gerechtvaardigd is. Het gaat immers om het gebruik van kenmerken die indirect verwijzen naar een beschermde grond: proxies. Omdat voor AI-systemen deze indirecte vorm van discriminatie het belangrijkste is volgen hierna enkele voorbeelden per grond om een idee te geven van waar je aan moet denken.

Op grond van

Burgerlijke staat

Doelstelling

Ik wil een eerlijke verdeling van schaarse woningen verzorgen.

Problematische proxy

Ik programmeer het systeem zo dat personen die een tweepersoonshuishouden voeren korting krijgen op de reguliere huur ten opzichte van personen die een eenpersoonshuishouden voeren voor hetzelfde type woning.

Mogelijk gevolg

Personen die behoren tot de groep 'gehuwd en/of geregistreerd partnerschap' zullen vaker een tweepersoonshuishouden voeren dan personen die behoren tot de groep 'niet gehuwd en/of geen geregistreerd partnerschap'. Deze laatsten worden door deze maatregel in het bijzonder getroffen.

Geslacht	**Doelstelling**	**I**k zoek fitte werknemers voor zwaar werk.
	Problematische proxy	**O**p basis van het huidige personeelsbestand blijkt dat iedereen langer is dan 1.70.
	Mogelijk gevolg	Ik laat het algoritme CV's hier op selecteren. **V**rouwen zijn gemiddeld kleiner en zullen dus minder snel door het algoritme worden geselecteerd.
Geslacht	**Doelstelling**	**I**k wil als verzekeraar zoveel mogelijk mijn risico's beperken met betrekking tot mijn arbeidsongeschiktheidsverzekering.
	Problematische proxy	**H**et algoritme geeft punten aan personen met beperkte risico's, zoals zij die in de afgelopen vijf jaar onafgebroken hebben gewerkt.
	Mogelijk gevolg	**V**rouwen hebben vaker werkonderbrekingen, bijvoorbeeld als zij kinderen krijgen, en komen als gevolg daarvan minder snel in aanmerking voor een verzekering.
Geloofs- of levensovertuiging	**Doelstelling**	**I**k wil een beveiligingssysteem toepassen op basis van gezichtsherkenning.
	Problematische proxy	**I**k stel strenge kledingvoorschriften op met als doel het hoofd en gezicht vrij te houden zodat het systeem het gezicht goed kan analyseren.
	Mogelijk gevolg	**P**ersonen met hoofd- of gezichtsbedekking vanwege religieuze redenen worden door de voorschriften in het bijzonder getroffen.
Handicap/ chronische ziekte	**Doelstelling**	**I**k wil mijn lening alleen verstrekken aan personen die deze zullen terugbetalen.
	Problematische proxy	**P**ersonen met een uitkering hebben vaak geen stabiel inkomen en komen niet in aanmerking.
	Mogelijk gevolg	**P**ersonen met een handicap/chronische ziekte zijn vaak oververtegenwoordigd in de groep van personen met een uitkering en worden hierdoor in het bijzonder getroffen.

Homo- of hetero-seksuele gerichtheid	**Doelstelling** **Problematische proxy** **Mogelijk gevolg**	**I**k verkoop erotische waar en wil op PRIDE dag een korting aanbieden. **I**k geef al mijn klanten die in het verleden veel erotiek met homoseksuele inslag hebben gekocht een kortingsbon van €5. **P**ersonen met een heteroseksuele gerichtheid kopen over het algemeen minder vaak erotische waar met homoseksuele inslag, waardoor zij niet snel in aanmerking voor de kortingsbon zullen komen.
Leeftijd	**Doelstelling** **Problematische proxy** **Mogelijk gevolg**	**I**k wil ervoor zorgen dat mensen mij een aantrekkelijke werkgever vinden door een verhuiskostenvergoeding aan te bieden. **D**e verhuiskostenvergoeding is hoger als je een eigen huishouding voert dan als je geen huishouding voert. **P**ersonen onder de 30 voeren minder vaak een eigen huishouding dan personen boven de 30. Jongeren krijgen dus vaker een lagere vergoeding.
Nationaliteit	**Doelstelling** **Problematische proxy** **Mogelijk gevolg**	**I**k wil mijn gemeubileerde woningen in Nederland alleen kortdurend verhuren en alleen aan goede huurders. **I**k laat mijn algoritme alleen ‘expats’ selecteren om in aanmerking te komen voor een huurwoning. **P**ersonen met een Nederlandse nationaliteit zullen doorgaans geen ‘expat’ zijn en deze categorie wordt hierdoor benadeeld.
Nationaliteit	**Doelstelling** **Problematische proxy** **Mogelijk gevolg**	**I**k wil goede werknemers aantrekken. **I**k laat het algoritme CV’s selecteren op basis van de diploma’s die worden uitgegeven door een Nederlandse universiteit. **P**ersonen met een Nederlands universitair diploma zullen vooral Nederlanders zijn, hierdoor komen niet-Nederlanders minder snel in aanmerking voor een baan.

Politieke gezindheid	**Doelstelling**	**I**k wil graag een harmonieuze arbeidsomgeving.
	Problematische proxy	**B**ij potentiële sollicitanten laat ik een algoritme het internet afzoeken om te bezien of zij vaak aanwezig zijn bij demonstraties. Deze worden niet geselecteerd voor een interview.
	Mogelijk gevolg	**P**ersonen met een sterke politieke overtuiging worden mogelijk geweerd uit het personeelsbestand.
Ras/etniciteit	**Doelstelling**	**I**k wil graag spontane mensen aantrekken voor de door mij uitgezette vacatures.
	Problematische proxy	**I**k laat een algoritme video-inzendingen van sollicitanten beoordelen op spontaniteit. Ik train het algoritme op mijn huidige werknemers. Deze zijn echter voornamelijk wit.
	Mogelijk gevolg	**H**et algoritme is minder goed in staat om de gewenste kwaliteiten te herkennen bij personen met een andere huidskleur.

Belangrijk is dat indirecte discriminatie niet altijd verboden is, namelijk als het onderscheid 'objectief gerechtvaardigd' kan worden. Dit betekent dat er sprake moet zijn van een legitieme reden om het onderscheid te maken, dat het maken van onderscheid proportioneel is en dat er geen minder ingrijpende middelen ter beschikking staan om hetzelfde doel te bereiken. Een taaleis gesteld in een vacature, zoals beheersing van het Nederlands, kan indirect discriminerend zijn. Over het algemeen zullen vooral Nederlanders Nederlands spreken, waardoor vooral niet-Nederlanders worden uitgesloten. Maar een taaleis kan toch gerechtvaardigd zijn, bijvoorbeeld als het gaat om een baan met veel contact met klanten die Nederlands spreken.

Een en ander kan als volgt schematisch worden samengevat:
(zie volgende pagina)

1 - Bewustwording

Is er bij mijn doel, design of uitkomst sprake van mogelijk 'verdacht' onderscheid?

- Burgerlijke staat
- Handicap/chronische ziekte
- Geslacht (incl. genderidentiteit)
- Godsdienst
- Leeftijd
- Levensovertuiging
- Nationaliteit
- Politieke gezindheid
- Ras/etniciteit
- Seksuele gerichtheid

→ Het door mij gebruikte algoritme dat acceptatievoorwaarden toets geeft een lagere waardering voor personen die langdurig arbeidsongeschikt zijn (geweest).

→ Mijn sollicitatiealgoritme wordt getraind op succesvolle CV's. Bij mij werken alleen mannen en niemand is onder de 18.

→ Ik wil een AI systeem bouwen dat die personen met een dubbele nationaliteit in mijn data er uit filterert en aanmerkt voor extra controle.

2 - Onderscheid?

Leidt dit tot benadeling?

→ Personen met een handicap/chronische ziekte worden mogelijk uitgesloten van mijn dienst.

→ De recruiter krijgt mogelijk geen CV's van vrouwen of personen onder de 18 onder ogen.

→ De groep wordt extra gecontroleerd en ondervindt daarvan negatieve consequenties.

3 - Kan ik mijn keuze rechtvaardigen?

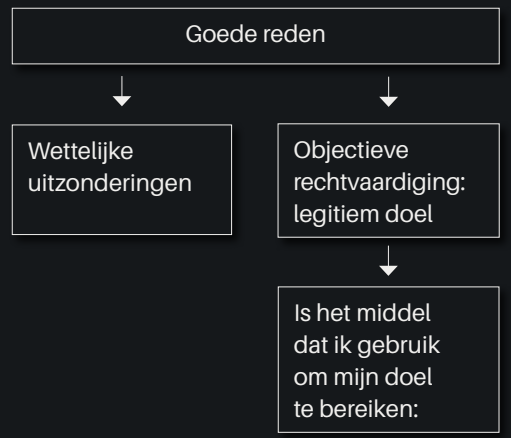
Heb ik een goede reden voor het gemaakte onderscheid?

Voorbeeld:

Sollicitatiealgoritme

Ik zoek personen voor een specifieke, gevaarlijke functie. Personen onder de 18 jaar mogen dit werk wettelijk niet verrichten.

Ik mag echter geen vrouwen weigeren voor de functie, en het algoritme kan er toe leiden dat deze CV's systematisch worden ondergewaardeerd. Ik moet hiervoor corrigeren.



1. Passend
- *Geschikt* om het legitieme doel te bereiken (draagt het bij aan de verwezenlijking ervan)?
- *Consistent* (vrij van innerlijke tegenstrijdigheden?)
- *Coherent* (bezien in de context waarbinnen de maatregel opereert)?

2. Noodzakelijk
Subsidiariteitsbeginsel: zijn er minder ingrijpende maatregelen waarmee het doel even effectief kan worden bereikt?

3. Evenredig
Evenredigheidsbeginsel: is er sprake van een redelijke afweging van de (belangen besloten in de) nagestreefde doelen en de belangen die door toepassing van het algoritme worden aangetast?

Doel van de handreiking

Besluitvormingsprocessen

Deze handreiking heeft ten doel te helpen bij het inrichten van het data-management, het bouwen van een algoritme en het ordenen van processen om tot besluiten te komen. Het gaat in deze handreiking dus om het proces dat voorafgaat aan een besluit, terwijl procedures in discriminatiezaken vaak gaan over concrete besluiten en de uitkomst daarvan. Zo'n besluit wordt vervolgens getoetst op drie vragen:

- Ten eerste: (1) is het besluit tot stand gekomen op basis van een van de verboden gronden (directe discriminatie) of;
- Ten tweede: (2) heeft het besluit onevenredig nadelig effect op categorieën personen met bepaalde kenmerken (indirecte discriminatie); en
- Ten derde: indien (1) of (2), is dat legitiem?

Deze handreiking vertaalt deze *ex post* toets (oftewel een toets nadat een beslissing is genomen) naar *ex ante* zorgvuldigheidsnormen (oftewel normen die voorafgaan aan een beslissing). Hieruit volgen twee beperkingen. Allereerst kan niet worden gezegd dat als aan de principes uit deze handreiking wordt voldaan er nooit een discriminatoire uitkomst kan zijn; daarop moet altijd ook *ex post* worden getoetst. Daarnaast is het omgekeerde ook niet waar; als de in deze handreiking vervatte principes niet worden gevolgd betekent dat niet dat de discriminatieregels zullen worden overtreden. Als een AI-systeem wordt gebruikt voor het voorspellen van astrologische processen, dan zal er immers niet snel een discriminatie-aspect mee gemoeid zijn.

Uiteraard moet worden bedacht dat besluitvormingsprocessen die geen gebruik maken van AI nu ook vaak gebiased zijn, maar dat er weinig kennis en data over dit soort bevooroordeeldheid is. Het werken met AI is dan ook niet per se een risico in termen van discriminatie, het is juist ook een kans om processen neutraler en rechtvaardiger te maken en beter in beeld te krijgen welke groepen worden getroffen door bepaald handelen, beleid of besluiten. Het gevaar met AI is wel dat als er discriminatie plaatsvindt, het een structureel probleem wordt dat zit ingebakken in het systeem en zodoende zeer grote gevolgen kan hebben.

Omdat het discriminatierecht zelf weinig *ex ante* regels bevat, wordt in deze handreiking naast dit rechtsgebied ook gekeken naar het privacy- en gegevensbeschermingsrecht, statistische principes, organisatorische handvatten en technologische best practices. Samen bij elkaar vormt dat een pakket aan principes voor het bouwen van een non-discriminatoir AI-systeem. De relevante vragen en uitgangspunten worden in deze handreiking weergegeven aan de hand van fasen die zijn gebaseerd op, maar niet precies hetzelfde zijn als, het CRoss-Industrie StandaardProces voor DataMining (*CRISP-DM).*

Vier uitgangspunten zijn leidend voor de hele handreiking en zullen in alle fasen in acht moeten worden genomen: 1. Diversiteit, 2. Context, 3. Controleerbaarheid en 4. Evaluatie.

1. Diversiteit

De vragen die in dit document staan kunnen niet in isolement worden beantwoord. Het document probeert een brug te slaan tussen de juridische wereld en de technische wereld. Terwijl juristen gewend zijn aan hele algemene uitgangspunten zoals ‘wees transparant’ en ‘vermijd discriminatie’ moeten bij het bouwen van een AI-systeem concrete keuzes worden gemaakt. Welk deel van het proces moet bijvoorbeeld transparant zijn en voor wie moet het proces begrijpelijk zijn: de systeembouwer zelf, een rechter of een leek? Hetzelfde geldt voor discriminatie: de keuze om de ene bias te vermijden betekent vaak dat een andere bias de kop op steekt.

Het is daarom belangrijk om het team dat aan een AI-project werkt zo divers mogelijk samen te stellen. Dat geldt voor de verschillende expertises en professionele achtergronden van mensen, maar ook voor de persoonlijke achtergrond binnen het team. Let daarbij op etniciteit, geslacht, geaardheid, culturele en religieuze achtergrond, leeftijd en andere aspecten die bij de concrete toepassingsfunctionaliteit van het AI-systeem relevant kunnen zijn.

2. Contextgevoeligheid

AI-systemen hebben uiteindelijk invloed op de praktijk, vaak binnen een specifieke context. Daarom is het belangrijk om al bij het ontwerp van het systeem specifieke domeinkennis binnen het team te hebben. Als AI-systemen zonder deze praktijkkennis worden ontwikkeld ontstaat er vaak een mismatch met de realiteit. Dat heeft niet alleen nadeligegevolgen voor de effectiviteit, maar kan ook ongerechtvaardigde discriminatie in de hand werken, bijvoorbeeld omdat de bouwers van een AI-systeem niet hebben gezien dat een bepaald datapunt een proxy is voor een van de gronden die in het anti-discriminatierecht een rol spelen.

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 de besparing in kosten en middelen worden ingezet voor extra capaciteit bij het diagnosticeren van vrouwen zonder een AI-systeem? Etc.

3. Controleerbaarheid

Het proces en de stappen die worden gezet moeten inzichtelijk, systematisch en controleerbaar zijn. Daarom is het belangrijk dat alle keuzes goed worden gedocumenteerd en verantwoord. Dit zorgt er voor dat achteraf kan worden gecontroleerd of er fouten zijn gemaakt en het systeem later beter kan worden geüpdatet. Het uitgangspunt is daarnaast dat onderdelen in het proces zo concreet mogelijk worden gedocumenteerd, zodat alle stappen herhaalbaar

en verifieerbaar zijn. Bedenk daarbij dat het proces transparant moet zijn voor verschillende doelgroepen, die elk weer verschillende vormen van documentatie vragen: de burger/belanghebbende, de toezichthouder en collega-systeembouwers die een *second opinion* doen.

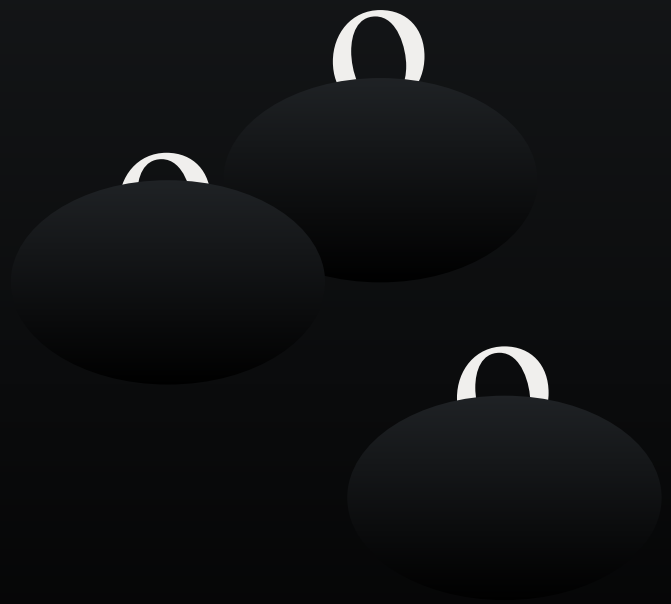
Belangrijk is om te zorgen voor interne en externe controle op de procesonderdelen. Dat kan bijvoorbeeld door een *second opinion* te vragen van externe experts, door testen te doen op dezelfde data met een ander algoritme of door het proces twee keer te doorlopen met een andere *fairness*definitie.

4. Evaluatie

In deze handreiking zijn zes stappen onderscheiden en lineair gepresenteerd, terwijl het bouwen van een AI-systeem in werkelijkheid een iteratief proces is. Van stap 4 ga je soms terug naar stap 2 en dan weer naar stap 3 en soms begin je bij elementen uit stap 5 en bekijk je pas daarna de vragen uit stap 1, om maar iets te noemen. Bovendien kan het zijn dat er al data zijn verzameld, een AI-systeem wordt gekocht van een externe partij of de probleemdefinitie al vast staat. In zulke gevallen kan het voorkomen dat niet alle stappen hoeven te worden doorlopen of in een andere volgorde.

De laatste van de zes stappen in deze handreiking is de evaluatiestap; het evalueren van het systeem is echter iets wat permanent moet gebeuren. Omdat het AI-systeem bovendien een lerend systeem is dat voortdurend in ontwikkeling is moet er constant worden gemonitord of aan alle voorwaarden en non-discriminatieprincipes is voldaan. Deze handreiking is dus nadrukkelijk niet bedoeld als een lijst met vragen die simpelweg afgevinkt kunnen worden. Het is opgesteld om op een gesystematiseerde en bewuste wijze keuzes te maken over hoe non-discriminatieprincipes in AI processen kunnen worden ingebed. Heb je eenmaal alle stappen doorlopen dan ben je er nog niet; omdat het systeem constant verandert moet je je constant bewust zijn van de verschillende vragen in de verschillende fasen in dit document als het systeem eenmaal loopt.

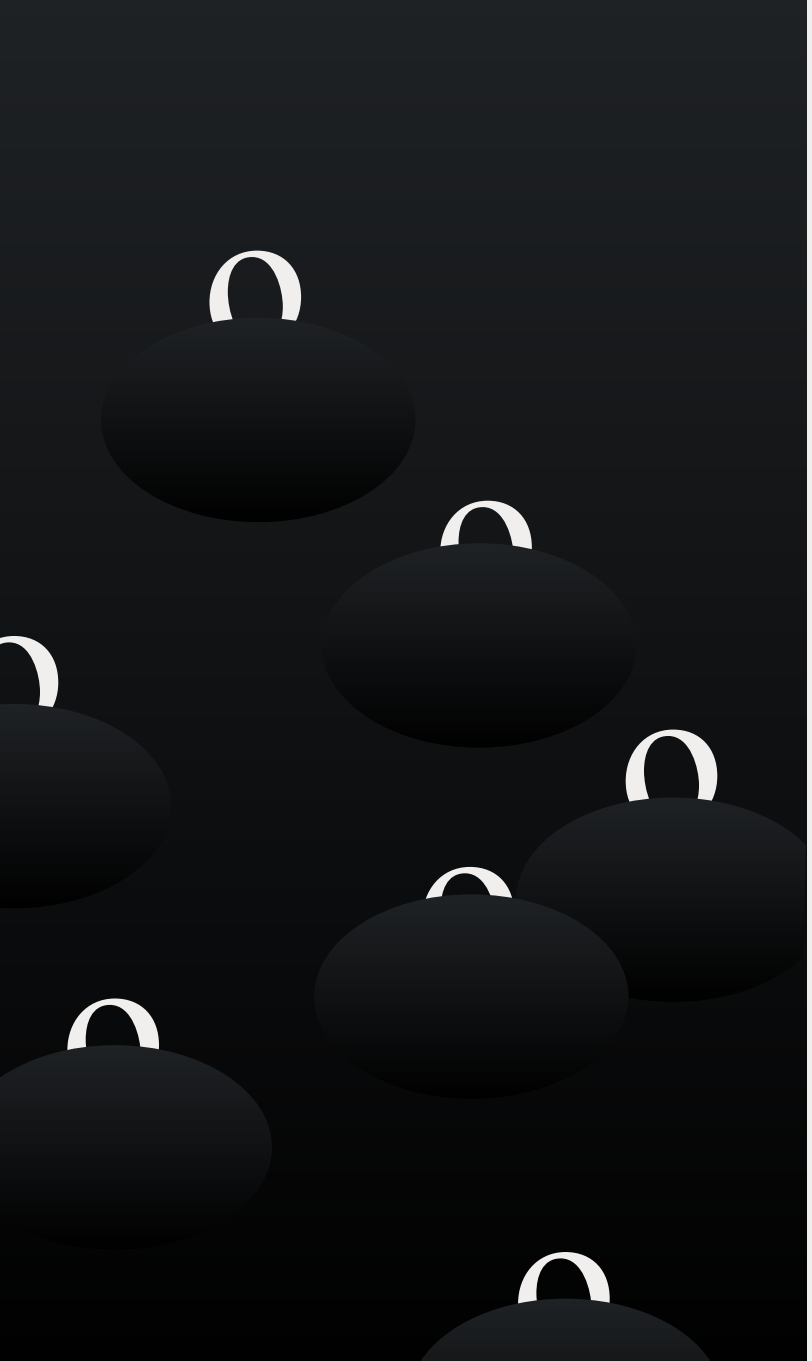
Aan de slag!



Tot slot. Deze handreiking is geschreven voor de projectleider van een AI-systeem en heeft ten doel de projectleider de juiste fasen te laten onderscheiden, de juiste mensen op juiste momenten bij elkaar te zetten en hen de juiste vragen te laten stellen. Als de technische experts en de data-analisten samen met de juristen en de functionaris gegevensbescherming aan tafel zitten, aangevuld met relevante stakeholders, domeinexperts en data stewards, zijn de vragen die in dit document zijn vervat leidend voor de discussie. De handreiking moet dan ook met name worden gezien als document dat dit vraaggesprek kan faciliteren en dat er voor zorg kan dragen dat alle relevante vragen in de juiste fasen van het project worden gesteld.

Ook kan deze handreiking dienen voor opdrachtgevers van AI-systemen, ofwel om vooraf offrerende partijen te vragen aan te geven hoe zij rekening zullen houden met de diverse punten uit deze handreiking, ofwel om tijdens het proces mee te kijken en op relevante punten aanwijzingen te geven, ofwel om achteraf te controleren of een opgeleverd product aan alle relevante voorwaarden voldoet. De tekst van deze handreiking is echter primair geschreven op teams die zelf AI-systemen bouwen.

Er is gekozen om de drie belangrijkste aspecten van het project, het juridische, technische en organisatorische onderdeel, apart te bespreken. Dit onderscheid is gemaakt, zodat de lezer duidelijk voor ogen heeft wat er op deze vlakken gedaan moet worden bij het ontwikkelen en implementeren van een AI-systeem. Deze drie aspecten staan niet los van elkaar, maar vullen elkaar aan en zullen in samenhang moeten worden bekeken. Het is dus geen “pick-and-choose-model”. Alle drie de componenten moeten in het AI project ingebed worden, aangezien ze elkaar onderling versterken. Bijvoorbeeld, in de datavoorbereidingsfase is het belangrijk om de juridische categorieën die tot directe en indirecte discriminatie kunnen leiden te betrekken in het technische proces wanneer categorieën worden gekozen ten aanzien van het aggregatie-



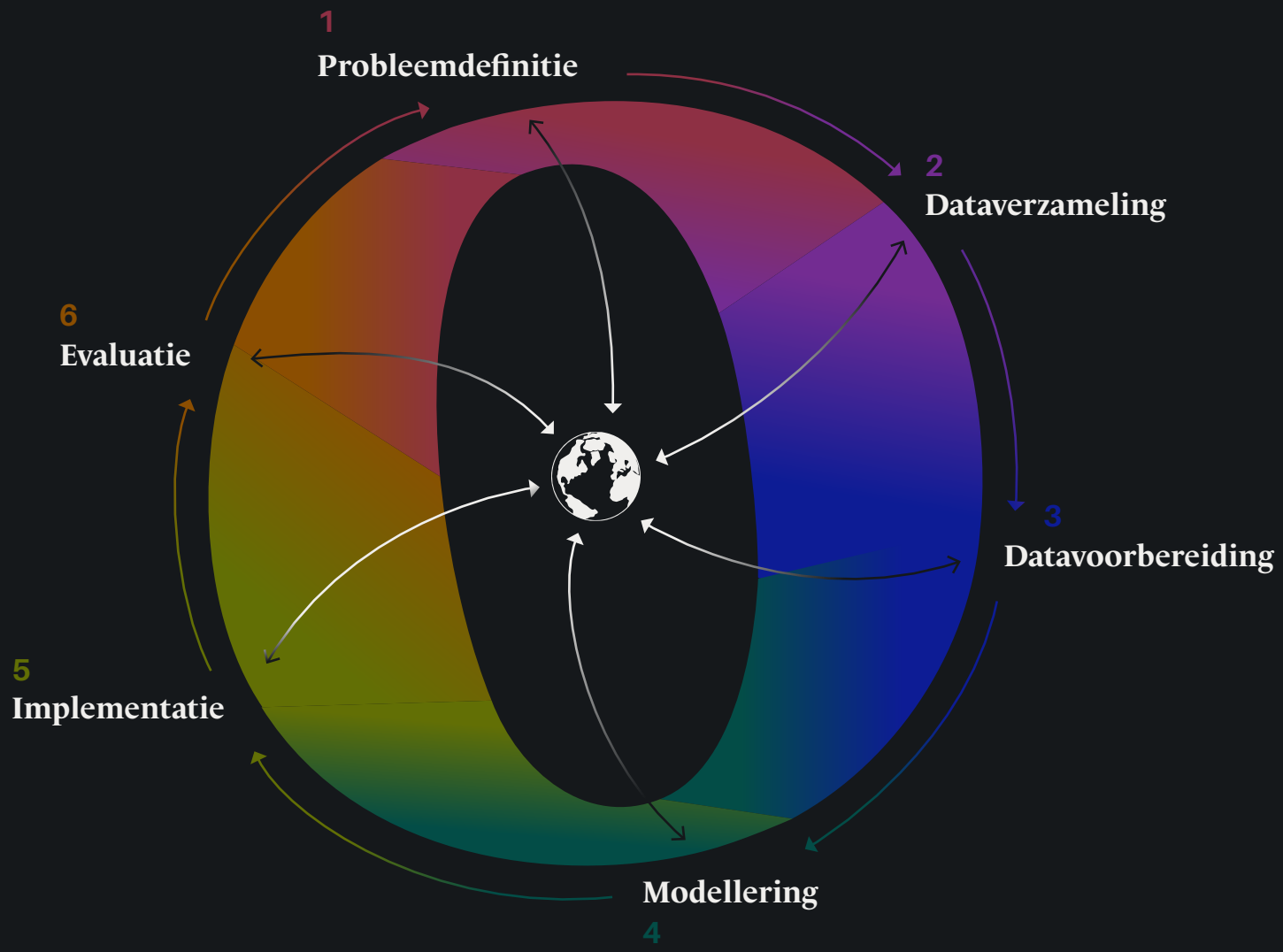
proces. Vervolgens moeten er op organisatorisch niveau duidelijke afspraken zijn over hoe ontdekte fouten worden gecorrigeerd en eventueel (wanneer van belang) gedeeld met andere organisaties.

Voor elk van de drie basiscomponenten, bevat de handreiking vragen en tips hoe ervoor te zorgen dat discriminatie zoveel mogelijk vermeden wordt. Afhankelijk van de context van het AI-project kan de ene vraag of opmerking relevanter zijn dan de andere. Aangezien dit een generieke handreiking is met als doel toepasbaar te zijn in verschillende contexten, is dit iets waar gebruikers zelf rekening mee dienen te houden. Om toch enigszins inzicht te geven in hoe verschillende type toepassingen tot verschillende aanpakken leiden, zijn er per fase voorbeelden opgenomen uit drie domeinen: de arbeidsmarkt, de strafrechtketen, en de medische context. In drie hypothetische casus worden niet alle, maar wel een aantal vragen uitgewerkt. De casus zijn gekozen ter illustratie; het betreft vrij tot de verbeeldingsprekende toepassingen, omdat daar het duidelijkst is welke morele, juridische en technologische vraagstukken met AI zijn gemoeid. De uitwerkingen geven een illustratie van wat een organisatie zou kunnen neerleggen bij de verschillende vragen; het zijn geen best practices.

Je vindt in dit document dus zes fasen en per fase:

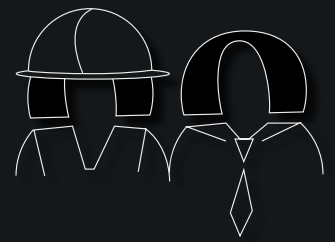
- **D**e belangrijkste vragen die je team moet doornemen;
- **E**en voorbeeld van hoe je een aantal van die vragen zou kunnen uitwerken in drie hypothetische casus;
- **E**n een uitwerking van de belangrijkste vragen voor de juridische, technische en organisatorische kant van het AI-systeem.

Non-discriminatie by design



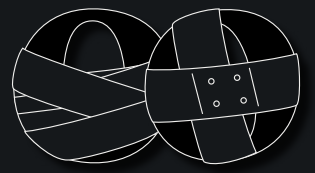
fase	Strategie		
1 - Probleem-definitie	Doel & noodzaak	Impact	Succes criteria
2 - Data-verzameling	Doel & noodzaak	Datakwaliteit	Dataopslag
3 - Data-voorbereiding	Inclusie & exclusie	Integratie & aggregratie	Labelen
4 - Modellering	Pre-modellering	Model (selectie)	Test
5 - Implementatie	Praktijktest	Aanpassing model	Toepassing
6 - Evaluatie	Evaluatie-voorbereiding	Evaluatie	Actiepunten

Voorbeeld



Arbeidsmarkt

Strafrechtketen



Medische domein

Principe

Juridisch

Technisch

Organisatorisch

Het uitgangspunt voor het ontwerp van dit document is gebaseerd op drie verschillende manipulaties van de letter ‘o’. De ‘o’ representeert het subject dat slachtoffer is van een bias. Het ontwerp is tot stand gekomen door een vrije interpretatie van prominente (maar niet uitsluitende) knelpunten in de verschillende domeinen waarin we de principes classificeren.

De beeldtaal is ontwikkeld ter ondersteuning van de tekst en wordt zowel functioneel als esthetisch ingezet.

Dit concept wordt volgens de volgende manipualties toegepast:

Ontwerpprincipes; een visuele vertaling van de vorm op basis van:

Juridisch: Een eenzijdig of onscherp perspectief.
Technologisch: Een onvolledige of selectieve dataset.
Organisatorisch: Een sturende definitie van succes.

Juridisch

Technologisch

Organisatorisch

fase
1 - Probleemdefinitie

Doel & Noodzaak

Impact

Succes-criteria

Belangrijke punten in deze fase:

- Scherp in beeld krijgen wat het probleem en de onderliggende aannames zijn.
- Het bepalen en helder formuleren van het doel en de noodzaak van het AI systeem.
- Er moet worden gekeken naar de invloed die het systeem zal hebben op de omgeving en de mensen in deze omgeving (wat verandert er voor wie en wat heeft dat voor gevolgen voor de relaties tussen mensen).
- In kaart brengen van de stakeholders en de groepen die onderscheiden worden in de probleemdefinitie.

Doel & noodzaak

1. Wat is het probleem en hoe gaat AI helpen het probleem op te lossen?

2. Is het noodzakelijk om met AI te werken of kan het probleem ook zonder een AI-systeem worden geadresseerd?

3. Welke groepen worden onderscheiden in de probleemdefinitie(s) en waarom?

4. Welke veronderstellingen over de verschillende groepen liggen ten grondslag aan de formulering van het probleem en het doel van het systeem?

5. Zijn de verschillende belanghebbenden daarin gehoord?

Impact

6. Moeten er voor dit project meer data worden verzameld en verwerkt dan reeds beschikbaar zijn binnen de organisatie en welke gevolgen heeft dat voor burgers?

7. Welke impact heeft het systeem op burgers en de maatschappij ten positieve en ten negatieve?

8. Wordt het systeem gebruikt om informatie te verkrijgen, om besluiten voor te bereiden of om zelfstandige besluiten te nemen en welke gevolgen heeft dat voor de mate waarin AI bepalend zal zijn in de praktijk?

9. Welke procedures zijn er voor belanghebbenden om een beslissing aan te vechten?

10. Wat is er bekend over aanwezige discriminatie/bias in de bestaande processen? Kan de invoering van het AI-systeem hier een positieve impact op hebben, al is het maar de bestaande bias verminderen?

Succescriteria

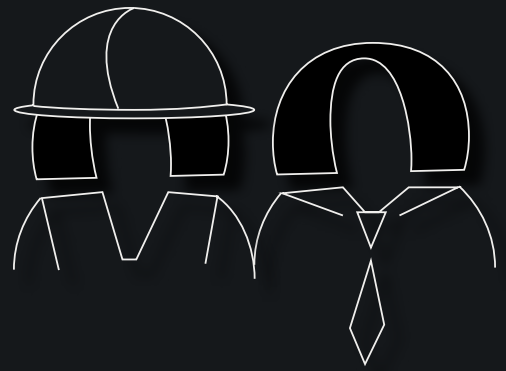
11. Wat zijn de financiële, computationele en organisationele kosten voor dit systeem en welke kosten zouden er zijn als er voor een niet-AI gedreven oplossing zou worden gekozen?

12. Wanneer is het AI-systeem een succes, bijvoorbeeld bij welk percentage van effectiviteit, en wanneer moet deze benchmark zijn gehaald, bijvoorbeeld na 1 maand of 2 jaar?

13. Welk percentage in foutnegatieven en foutpositieven is acceptabel en waarom?

14. Wat is de gekozen definitie van *fairness* en waarom?

15. Wat betekenen de verschillende succescriteria voor verschillende groepen waarop het systeem (mogelijk) een impact heeft?



Het probleem is dat het handmatig beoordelen van brieven veel tijd kost en manuele selectie biased is (human bias). Dit systeem dient om een voorselectie te maken van sollicitatiebrieven. Het heeft ten doel het proces efficiënter en minder biased te maken door een prioritering (ranking) te geven van de brieven op basis van een aantal voorgedefinieerde categorieën.

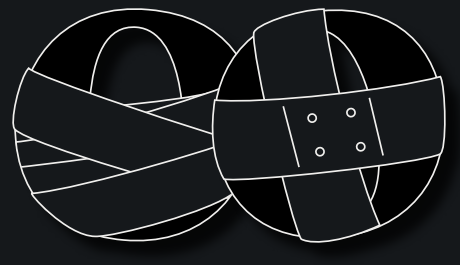
Voor de training van dit systeem wordt gebruik gemaakt van reeds binnengekomen sollicitatiebrieven. Het systeem maakt automatisch beslissingen. Foutpositieven betekent een minder efficiënt systeem, foutnegatieven betekent geen baankans voor een gekwalificeerde sollicitant.

Succes bij 20% kostenreductie en 5% minder onterechte afwijzing t.o.v. het bestaande proces. Moet zich uiterlijk na 1 jaar hebben bewezen. Maximaal foutpositief: 40% (t.o.v. de mensen die worden uitgenodigd voor een gesprek) en foutnegatief: 2% verschil (t.o.v. de mensen die zouden worden uitgenodigd voor een gesprek door manuele selectie).

Het probleem is dat de politie nu vaak pas na de feiten ter plaatse is. Dit systeem dient om te voorspellen waar en wanneer een overtreding zal plaatsvinden en aldus preventief en gericht patrouilles in te zetten. Dit systeem heeft ten doel het proces kosten-efficiënter, nauwkeuriger en minder biased te maken.

Gebruik van reeds bestaande databases, aangevuld met gescrapte social media data. Het systeem maakt voorspellingen en informeert de corpsleiding. Foutpositieven betekent efficiëntie verlies maar mogelijk ook verstoring van relaties met de buurt of ondermijning van het vertrouwen in het AI systeem. Foutnegatieven betekent bijvoorbeeld minder effectiviteit en ook ondermijning van het vertrouwen van gebruikers in het AI systeem.

Succes bij 10% hogere effectiviteit van politiesurveillances, gemeten aan het aantal arrestaties en staande houdingen. Moet zich na 3 jaar hebben bewezen. Gedurende de 3 jaar worden surveillances uitgevoerd op basis van het bestaande proces en anderen met behulp van het AI-systeem. Die laatsten mogen geen hogere foutpositieven of -negatieven opleveren. Om foutnegatieven te meten wordt een random allocation policy ingevoerd.

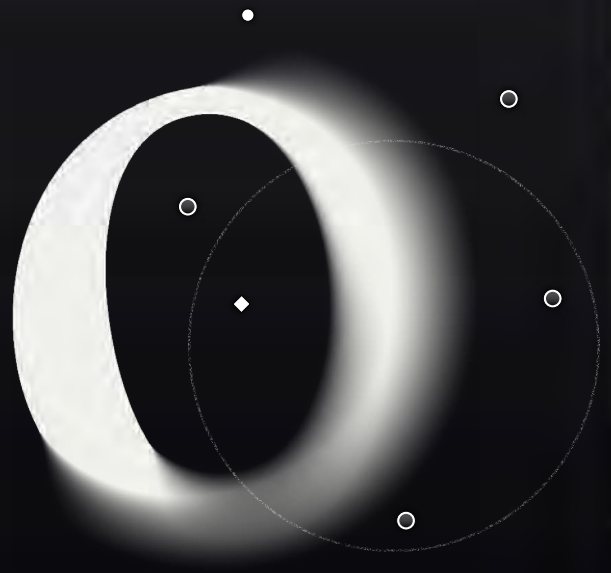


Het probleem is dat het handmatig beoordelen van scans veel tijd kost, duur is en niet altijd accuraat is. Dit systeem dient om betere diagnoses te stellen voor een specifieke vorm van kanker. Dit systeem heeft ten doel het proces kosten-efficiënter en nauwkeuriger te maken.

Data worden verzameld gedurende de trial-fase van dit systeem in verschillende instituten over de wereld. Het systeem maakt automatisch beslissingen bij de duidelijke gevallen en geeft de keuze aan een arts bij een hoge onzekerheidsmarge. Foutpositieven leiden tot angst, foutnegatieven leiden in extremo tot de dood.

Succes bij 10% hogere score van de diagnoses, gemeten ten opzichte van de foutnegatieven die nu bestaan. Kosten mogen niet hoger zijn dan het nu bestaande proces. Er zit geen einddatum aan dit proces, maar bij de constatering van een hogere foutmarge t.a.v. foutnegatieven wordt het project onmiddellijk stopgezet.
Artsen blijven diagnoses uitvoeren zoals nu, daarnaast doet een AI-systeem diagnoses. Een onafhankelijke arts vergelijkt en controleert de resultaten.

Juridisch

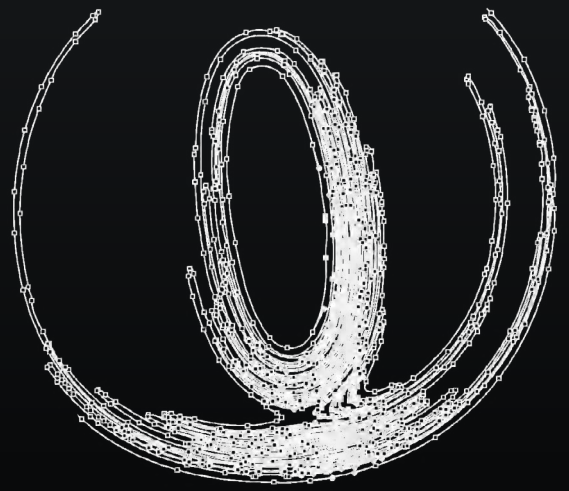


- Er dient vooraf een helder en concreet doel voor het AI systeem te worden gedefinieerd.
- Kleven er aan de opgestelde doelstelling en/of parameters op grond waarvan onderscheid wordt gemaakt discriminatie-aspecten?
- Ga na of er persoonsgegevens worden verwerkt, dat wil zeggen gegevens die kunnen worden herleid tot concrete individuen. Klik hier voor meer *uitleg*.

- Er dient een Impact Assessment te worden uitgevoerd, waarin wordt gekeken naar de impact van het systeem, discriminatie en privacy en hoe die gemitigeerd worden. Als er hoge privacyrisico's blijven bestaan na aanvullende maatregelen dan moet de Autoriteit Persoonsgegevens worden geraadpleegd. Klik hier voor een *model*.

- Juridisch gezien moeten systemen noodzakelijk, proportioneel en subsidiair zijn. Dat betekent dat de nadelen van zo'n systeem niet groter mogen zijn dan de voordelen en dat er geen andere, minder ingrijpende manieren zijn om hetzelfde doel te verwezenlijken.
Klik hier voor *uitleg*.
- Doorloop het schema op pagina 10. Is er sprake van onderscheid, en zo ja, is daar een goede reden voor? Klik hier voor meer *uitleg*.

Technisch



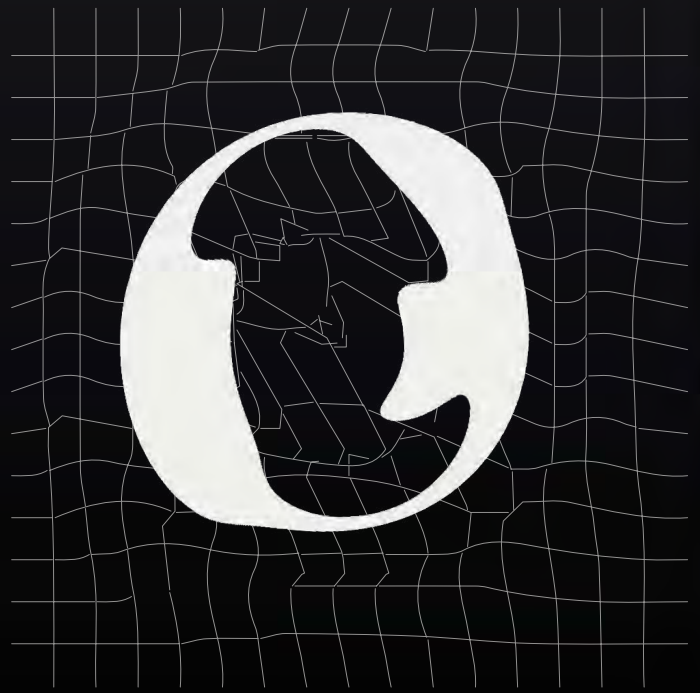
- Beargumenteer en documenteer de keuze voor AI in relatie tot context: bijv. keuze voor doelvariabele, classificatietaak, prestatiedoelen, etc.
- Formuleer en documenteer de logica en het waarom van het AI systeem.
- Beargumenteer en documenteer de keuze voor:

- **Evaluatie metrics:** Representeren de *metrics* de belangen van alle belanghebbenden, ook buiten de organisatie? Denk aan de impact van foutpositieven en foutnegatieven op verschillende groepen. Overweeg om ook een *fairness metric* mee te nemen.
- ***Target* variabele**: Is de *target* variabele een goede maatstaaf voor het concept dat wordt voorspeld of is er sprake van *measurement bias*? In veel gevallen is het concept moeilijk meetbaar en wordt gebruik gemaakt van een *proxy*. In plaats van criminaliteit meten we arrestaties en in plaats van kwaliteit van werknemers meten we de evaluatiescores van hun manager. Het verschil tussen de target en de *proxy* is een vorm van *measurement bias.*
Wanneer de bias verschillend is voor subgroepen, kan dit leiden tot een discriminatoir model.
- **Uitlegbaarheid**: Is het nodig om complexe technieken zoals *deep learning* te gebruiken of volstaat een simpeler uitlegbaar model, zoals een *beslisboom* of *lineaire regressie*? Om te kunnen debatteren over de *fairness* van een model, helpt het om te begrijpen hoe het model tot een beslissing komt.

- Beargumenteer de keuze voor een type AI-systeem: is het een eenvoudige beslisboom, een zelflerend systeem of *deep learning*?
- Waartoe wordt het systeem ingezet: het verkrijgen van inzichten, het voorbereiden van besluiten of het nemen van zelfstandige besluiten?
- Klik hier voor *uitleg*. (het document onder deze link geeft ook een invulling van de algemene beginselen van behoorlijk bestuur bij nieuwe vormen van bedrijfsprocesinrichting die door inzet van algoritmen en data-analyse ontstaan).

- Bepaal de succescriteria ten aanzien van foutpositieven en -negatieven, de effectiviteit en de loopduur van het project.
- Voer een technische evaluatie uit. Een voorbeeld is de Ethics Canvas als ontwikkeld door het Open Data Institute. Klik hier voor de *canvas.*

Organisatorisch



- Loop de volgende vragen langs:
  - Heeft het team de toegang tot de benodigde middelen?
  - Heeft het team de benodigde bevoegdheden?
  - Is alle relevante expertise aanwezig in het team?
  - Is het team zo divers mogelijk?
  - Is een persoon met domeinkennis betrokken bij het project?
- Breng in kaart binnen welk domein het systeem een rol zal gaan spelen en wat de maatschappelijke context is.
- Breng in kaart wie de belanghebbenden zijn en bij wie de voordelen van het systeem en bij wie de nadelen komen te liggen.
- Betrek de relevante belanghebbenden of vertegenwoordigende (burgerrechten) organisaties in een vroegtijdig stadium.
- Documenteer alle stappen en keuzes in het proces en overleg die keuzes zowel binnen het team als met belanghebbenden.
- Bij twijfel, contacteer een externe expert voor een *second opinion.*
- Leg vast wanneer het proces zal worden stopgezet en wat de exitstrategie is.
- Leg vast wie binnen de organisatie welke taken heeft en waarvoor verantwoordelijk is.

fase
2 - Dataverzameling

Doel & Noodzaak

Data-kwaliteit

Data-opslag

Belangrijke punten in deze fase:

- Welke data zijn reeds in het bezit van de organisatie en welke data moeten worden verzameld? Is er een wettelijke grondslag voor het hergebruik van bestaande data of nieuw te verzamelen data voor profiling en/of automatische besluitvorming?
- Welke bias zit er in de data en valt die bias te mitigeren?
- Is er gevaar voor misbruik van de data, ofwel door interne medewerkers of door externe organisaties of hackers, voor discriminatoire doeleinden en zo ja, valt dat risico te mitigeren?

Doel & noodzaak

1. Welke data zijn nodig voor dit project en waarom?

2. In hoeverre zijn deze gegevens al binnen de organisatie beschikbaar en in hoeverre moeten ze van buiten worden gehaald?

3. Is het toegestaan om deze data voor dit project te verzamelen en te verwerken?

Datakwaliteit

4. Welke bias zit er in de data (van binnen, buiten of gecombineerd) en welke consequenties heeft dat?

5. Uit welke context komen de data en wat zijn de aannames die achter de representaties liggen?

6. Zijn de data representatief en zijn alle relevante groepen in gelijke mate vertegenwoordigd?

7. Als verschillende databronnen worden gebruikt, hoe wordt er gezorgd dat deze data compatibel en vergelijkbaar zijn?

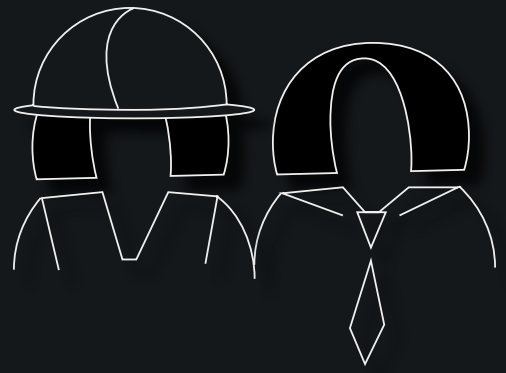
8. Kan het koppelen van data leiden tot proxies en "*disparate impact*"?

Dataopslag

9. Hoe lang worden de gegevens bewaard en hoe?

10. Worden de gegevens veilig en vertrouwelijk behandeld; welke gevolgen zou een datalek hebben voor groepen of categorieën personen?

11. Worden data gedeeld met andere partijen en wat is het gevaar dat die misbruik maken van de data met negatieve gevolgen voor groepen of categorieën personen?



Data van 30 random vroegere sollicitatieprocedures zullen worden beoordeeld op basis van daadwerkelijke uitkomst en de uitkomst als gesuggereerd door het AI-systeem. Data zijn afkomstig van de organisatie zelf. Het gaat hier om secundair gebruik van data dat is geoorloofd vanwege het publieke belang van het voorkomen van sollicitatieprocedures op basis van bewuste of onbewuste vooroordelen.

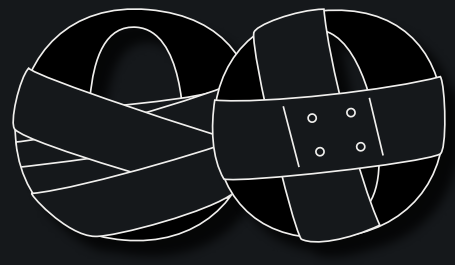
Er zit een grote bias in de bestaande dataset, onder meer op basis van etniciteit en afkomst. De bias in bestaande data is de reden om dit systeem te introduceren. Succes van het AI-systeem zal dus niet worden beoordeeld op basis van de mate waarin de uitkomsten gelijk zijn aan vroegere processen. Een onafhankelijk team zal beoordelen welke van de selectiemethoden de meest rechtvaardige uitkomst oplevert.

Data worden slechts voor de duur van dit project bewaard en daarna verwijderd. Gedurende het project staan de data op een interne cloud. Deze is slechts toegankelijk voor medewerkers van de afdeling werving en selectie en voor de ontwikkelaars van het AI-systeem.

Data over effectiviteit van vroegere surveillances en criminaliteitscijfers zijn beschikbaar binnen de organisatie. Deze data worden gekoppeld aan open data afkomstig van sociale media. De data worden geanonimiseerd en geaggregeerd op groepsniveau. De verwerking vindt plaats op basis van statistische data en niet op persoonsgegevens.

Er zit een historische bias in de dataset van de politie, onder meer ten aanzien van bepaalde wijken, personen met een immigratieachtergrond, lage sociaaleconomische status en mannen. Er zit een bias in de social media die vooral door jongeren worden gebruikt. Doordat sprake is van historische bias die gekoppeld kan worden aan bepaalde wijken, personen met een immigratieachtergrond en mannen is de data niet representatief en neutraal. Deze twee databronnen worden onafhankelijk van elkaar beoordeeld en niet direct gekoppeld. Er zal worden gecorrigeerd op deze bias.

Alle relevante data worden in geanonimiseerde vorm oneindig bewaard omdat ook historische patronen van grote waarde zijn voor toekomstige processen. De data worden binnen een beveiligde ruimte opgeslagen. Slechts leden van een speciale unit hebben toegang tot deze data.

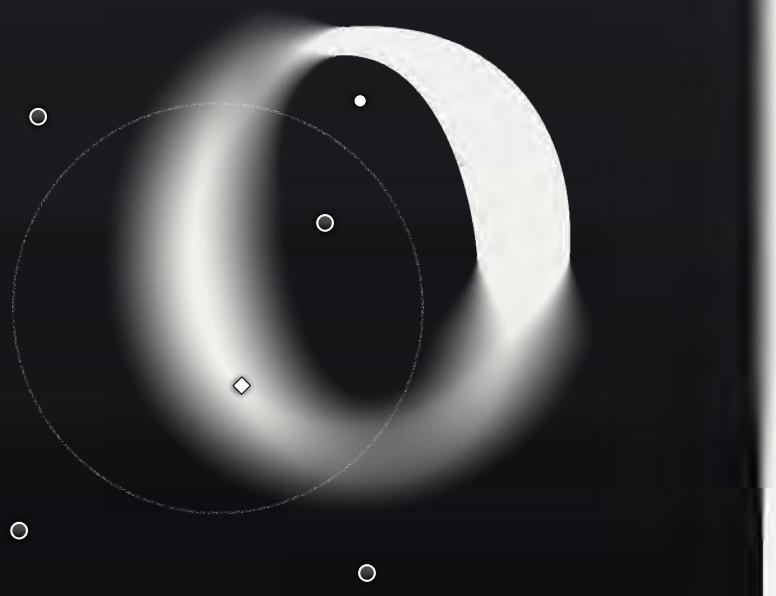


Er zijn zo veel mogelijk data nodig over diagnoses gesteld door artsen en *second opinions* door het AI-systeem. Hoe presteert het AI-systeem ten opzichte van een arts? Er wordt slechts gewerkt met data die patiënten zelf afgeven. Hiervoor geven zij geïnformeerde toestemming.

Er zal een bias zitten in deze patiëntpopulatie, namelijk vrouwen van 50+ met overgewicht. Het AI zal dus leren om met name ten aanzien van deze groep accurate voorspellingen te doen. Dat kan betekenen dat op termijn, het AI-systeem wordt ingezet voor diagnoses bij deze groep en artsen voor andere groepen, al naar gelang de accuratesse daartoe aanleiding geeft. De databronnen worden door een speciale commissie beoordeeld, die expertise heeft op het gebied van de onderzochte ziekte en zal zo de data controleren op onregelmatigheden.

Data worden 20 jaar lang bewaard, om kwaliteitscontroles en assessments mogelijk te maken. Alleen onderzoekers en data-analisten hebben toegang tot de data. De data van de verschillende instituten worden aan elkaar gekoppeld en worden gedeeld in een secured private cloud.

Juridisch



- Beoordeel of er een legitieme grondslag is voor het verwerken van persoonsgegevens, zoals een wettelijke plicht, informed consent of publiek belang. Klik hier voor meer *uitleg*.
- Als er gegevens over iemands ras, geaardheid, gezondheid, strafrechtelijk verleden of andere gevoelige gegevens worden verwerkt, dan is expliciete toestemming, een wettelijke grondslag of een zwaarwegend publiek belang nodig. Klik hier voor *uitleg*.
- Als er gegevens uit andere bron(nen) worden verkregen, dan moet worden geverifieerd of die data op een legitieme wijze zijn verkregen.

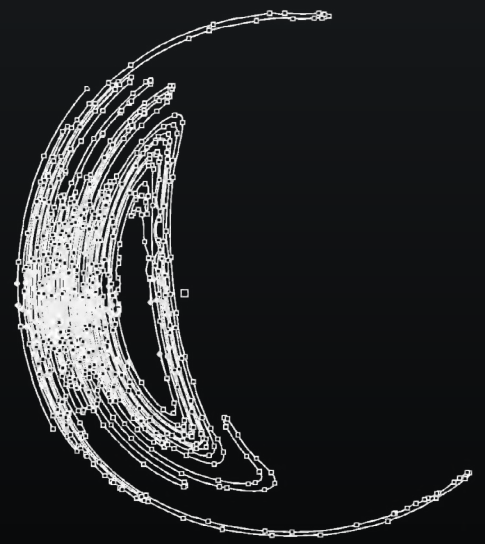
- Data moeten correct zijn, up-to-date worden gehouden en betrokkenen moet de mogelijkheid worden geboden om aanvullende gegevens aan te dragen.
Klik hier voor een *voorbeeldsbrochure.*
- Controleer je data op mogelijke bias ten opzichte van de beschermde gronden. Dit kan bijvoorbeeld met *Aequitas.*

- Als gegevens de EU-grens overgaan, en bijvoorbeeld naar de VS of China worden doorgevoerd, moet de organisatie waarmee de gegevens worden gedeeld zich aan de in de Europese Unie geldende gegevensbeschermingsregels houden. Klik hier voor nadere *uitleg.*

Technisch

- Zorg dat data zo worden verzameld dat ze kunnen worden benut door het AI-systeem.
- Leg daarbij van te voren technische vereisten vast. Beoordeel data op:
  - Verdeling van attributen (bijvoorbeeld het doelattribuut van een voorspellingstaak)
  - Relaties tussen paren of een klein aantal attributen
  - Resultaten van eenvoudige aggregaties
  - Eigenschappen van significante subpopulaties
- Beoordeel data op mogelijk aanwezige bias aan de hand van o.a.:

Technisch



- Verdeling van de doelvariabele in verschillende groepen. Ongelijkheid in de samenleving is vaak ook terug te zien in op zich neutraal verzamelde data.
Voorbeeld: Wanneer de proportie ‘positieven’ voor vrouwen anders is dan voor mannen, kan dit een teken zijn van historische bias.
- Relaties tussen paren of een klein aantal attributen (features): leidt dit tot proxies?
Voorbeeld: postcode kan een proxy zijn voor etniciteit.
- Verdeling van attributen en representatie van relevante subpopulaties. Representatie bias ontstaat wanneer delen van de inputruimte zijn onder- of overgerepresenteerd.
Voorbeeld: wanneer een training dataset van gezichtsherkenningssoftware weinig (d.w.z., lage representatie) foto’s bevat van donkergetinte gezichten (d.w.z., andere verdeling van attributen) bestaat het risico dat het systeem slechter zal werken voor deze groepen.
- De aannames achter de data: meten we wel wat we willen meten?
Voorbeeld: Zijn verkoopcijfers wel een goede proxy voor de verkoopkwaliteiten van werknemers?

- Zorg voor de technische veiligheid en vertrouwelijkheid van de data.
- Zorg dat de data geëncrypteerd en gecompartementaliseerd zijn.

Organisatorisch

- Zorg dat data zo neutraal en objectief mogelijk worden verzameld en dat dit zo transparant en controleerbaar mogelijk geschiedt. De wijze waarop data worden verzameld, hoe, door wie, waar en met gebruikmaking van welke technieken kunnen bepalend zijn voor de neutraliteit en betrouwbaarheid van de verkregen data.
- Leg dit proces nauwgezet vast en bewaar te allen tijde gegevens over hoe de dataset tot stand is gekomen, om zo de processen controleerbaar en herhaalbaar te maken. Doe dat ook voor de data die via derden zijn verkregen.

- Controleer op ontbrekende waarden, nauwkeurigheid en representativiteit en of de verdeling voor verschillende groepen gelijk is. Leg verschillen uit en overweeg mitigerende maatregelen.
- Overweeg om nieuwe gegevens te verzamelen of om het doel van het project te herzien: ga dan terug naar fase 1 (probleemdefinitie).
- Faciliteer een gesprek over de controle op ontbrekende waarden, nauwkeurigheid en representativiteit tussen verschillende betrokkenen (bijv. data scientists, AI experts, product owner, project leider, bevoegd gezag).

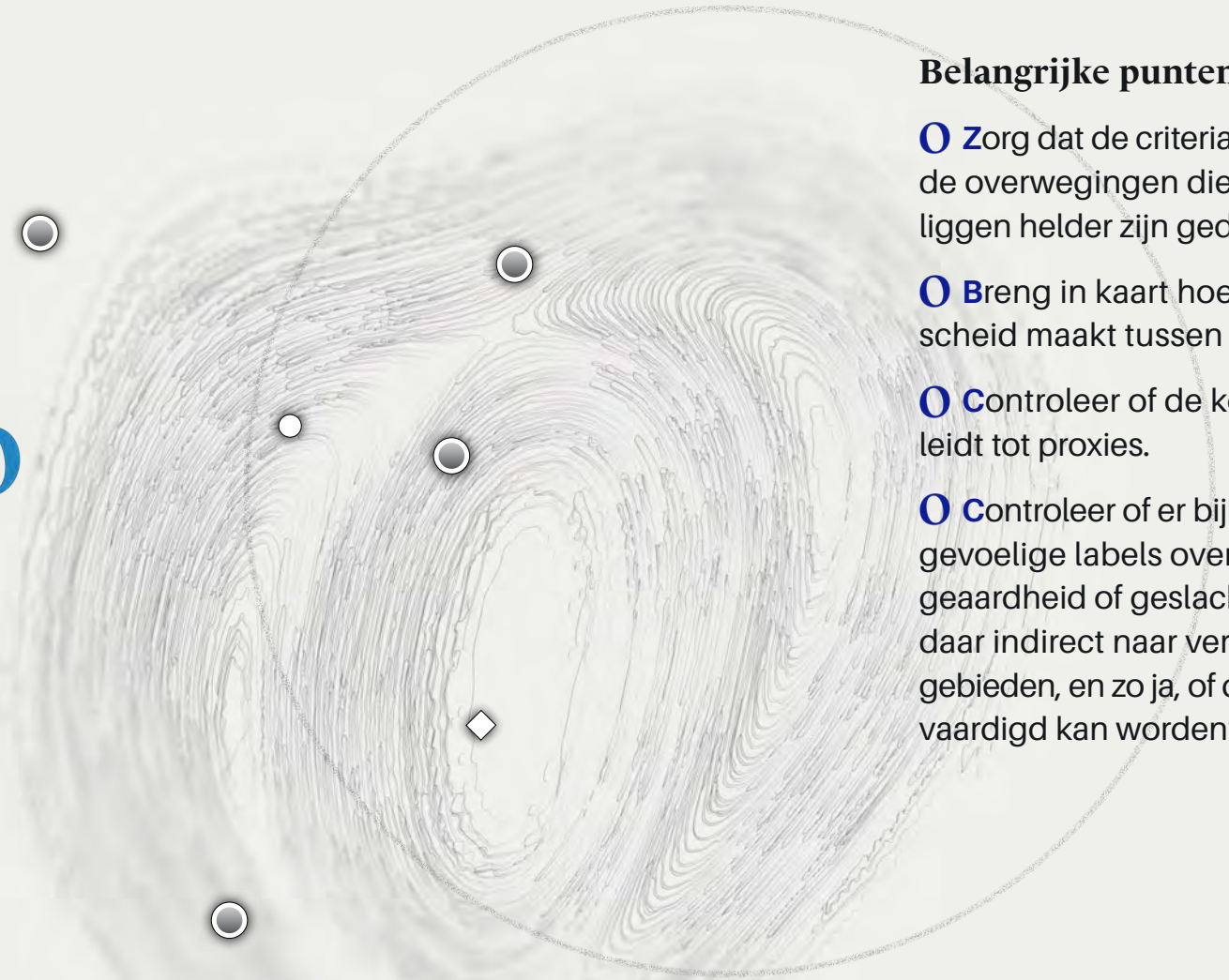
- Zorg voor organisatorische veiligheid en vertrouwelijkheid van de data. Zorg dat data alleen met dubbele authenticatie in te zien zijn door een geselecteerd aantal werknemers of betrokkenen.

fase
3 - Datavoorbereiding

Inclusie & exclusie

Integratie & aggregatie

Labelen



Belangrijke punten in deze fase:

- Zorg dat de criteria voor de dataselectie en de overwegingen die daaraan ten grondslag liggen helder zijn gedocumenteerd.
- Breng in kaart hoe de dataselectie onderscheid maakt tussen verschillende groepen.
- Controleer of de koppeling van data leidt tot proxies.
- Controleer of er bij het labelen van de data gevoelige labels over bijvoorbeeld etniciteit, geaardheid of geslacht zitten en/of labels die daar indirect naar verwijzen, zoals postcodegebieden, en zo ja, of dat logisch is en gerechtvaardigd kan worden?

Inclusie & exclusie

1. Welke van de verzamelde data zijn relevant voor het model en waarom?

2. Wat gebeurt er met de data die niet worden gebruikt?

3. Aan de hand van welke criteria wordt de keuze voor dataselectie gemaakt en welke impact hebben die op het onderscheid tussen groepen?

4. Beïnvloedt de keuze voor bepaalde data of databewerkingen de probleemdefinitie?

5. Welke aspecten van het probleem worden buiten beschouwing gelaten?

Integratie & aggregatie

6. Hoe wordt gezorgd dat historische data en nieuw verzamelde data op elkaar aansluiten; zijn de data vergelijkbaar en welke aannames ten aanzien van groepen en categorieën zitten er reeds in de verzamelde data en in de nieuw te verzamelen data?

7. Op welke wijze worden data geaggregeerd en welke gevolgen heeft dat voor de representativiteit van de data?

8. Wat betekent dit voor de representatie van het probleem en de belanghebbenden? Bijv. betekent dit een herformulering van een groep of categorie?

9. Leidt de combinatie van verschillende data tot proxies en zo ja welke?

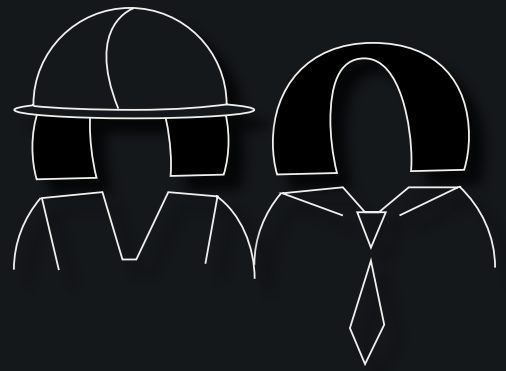
Labelen

10. Hoe worden data gelabeld en waarom?

11. Sluit dit aan bij hoe andere organisaties data labelen en datasets gebruiken waarop het algoritme is getraind?

12. Sluit dit aan bij hoe belanghebbenden/ burgers en domein experts data zouden labelen?

13. Zitten er gevoelige labels over bijvoorbeeld etniciteit, geaardheid of geslacht bij of labels die daar indirect naar verwijzen, zoals postcodegebieden, en zo ja, waarom?



De gegevens worden meegenomen die belangrijk zijn voor de kandidaatselectie: opleiding, ervaring en nevenactiviteiten. Er wordt bij een vacature minimale eisen gesteld aan een sollicitant, deze eisen worden de criteria waaraan het AI-systeem zal toetsen. De gegevens die niet gebruikt hoeven worden, worden bewaard, om later te testen of het systeem beter presteert met meer of andere datapunten.

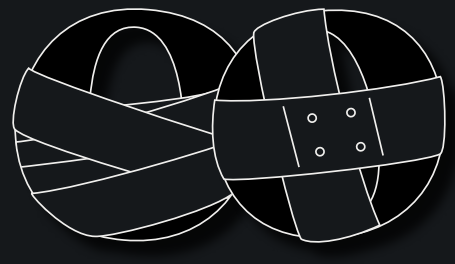
Nieuwe data zullen niet aansluiten op historische data, omdat in de historische data een grote bias zat. Data worden op basis van opleidingsniveau en ervaring (junior, medior en senior) geaggregeerd. Deze vorm van aggregatie heeft geen invloed op het probleem dat hiervoor speelde, omdat hier sprake was van een ongerechtvaardigde bias en deze vorm van aggregatie rechtvaardig is. Wel zal worden bekeken of eventuele bias, ofwel in bestaande data, ofwel in de performance van het AI-systeem varieert voor junior, medior of senior niveau.

Data worden gelabeld op basis van de voorgenoemde kwaliteiten van de sollicitanten, om zo een objectieve keuze te kunnen maken voor een sollicitant. De manier waarop de data worden gelabeld komt overeen met andere organisaties. Er zitten gevoelige gegevens tussen, maar deze spelen geen rol bij de labeling. Wel kunnen de categorieën indirect iets zeggen over groepen; hierop zal worden getest.

Data worden geïncorporeerd die duidelijk maken waar de meeste criminaliteit plaatsvindt: locatiegegevens, informatie over vermeende of geconstateerde overtredingen, tijdsaanduiding, achtergrondinformatie verdachte. Deze keuze ten aanzien van de criteria wordt gemaakt aan de hand van gemelde inbraken/overtredingen en uitgeschreven processen verbaal. Data die niet gebruikt worden, worden verwijderd.

Alle data die gebruikt worden, worden in eenzelfde format gegoten, zodat ze op dezelfde aspecten vergelijkbaar zijn. Bijvoorbeeld: Historische arrestatiedata bevat tijd en plaats, aard van de overtreding en vroegere veroordeling van de overtreder. Ook hier worden alle data uniform gemaakt zodat alle dimensies overeenstemmen en zodoende vergelijkbaar zijn. Data worden geaggregeerd op basis van tijd en plaats, aard van de overtreding, aard en aantal van vroegere veroordelingen van de overtreder.

Data worden gecategoriseerd op basis van wijken/postcodegebieden, om zo in kaart te brengen welke wijken meer aandacht behoeven. Data worden ook gecategoriseerd volgens periode: tijd van de dag, dag in de week, periode in het jaar. Zo kan er via data-anayse worden geleerd dat er bijvoorbeeld meer inbraken plaatsvinden ‘s nachts, ‘s weekends of tijdens de vakantieperiode. Er zitten gevoelige labels tussen, namelijk postcodegebieden, Deze kunnen indirect verwijzen naar bepaalde etnische groepen.



De gegevens worden meegenomen waaruit blijkt dat deze vorm van kanker voornamelijk voorkomt: geslacht, leeftijd, sociaaleconomische status, leefpatroon en ziekteverloop. De gemaakte keuze voor bepaalde data wordt gemaakt op basis van historische data, waarin deze datapunten als meest relevant naar voren kwamen. De gegevens die niet worden gebruikt worden verwijderd, conform de AVG, tenzij er toestemming is gegeven deze gegevens te bewaren of te hergebruiken voor medisch wetenschappelijk onderzoek. Aan de hand van risicogroepen wordt een nadere keuze gemaakt voor dataselectie, hierbij wordt onderscheid gemaakt tussen bepaalde groepen.

Historische data krijgen een bepaald gewicht toegekend, deze historische data worden meegenomen in nieuwe gevallen. Data worden op basis van leeftijd, geslacht en sociaaleconomische status geaggregeerd. In overleg met andere organisaties wordt gekozen voor het gebruik van dezelfde medische en technische termen om zo de data uit verschillende bronnen op elkaar aan te laten sluiten.

Data worden gestructureerd op basis van onder andere leeftijd, etniciteit en geslacht, om zo te zien of deze bepaalde vorm van kanker meer voor komt in bepaalde groepen. Er zitten gevoelige categorieën tussen; deze zijn nodig om in kaart te kunnen brengen in welke groepen een groter percentage deze vorm van kanker heeft.

Juridisch

- Controleer bij de inclusie en exclusie van data welke gevolgen de dataselectie heeft voor de representatie, met name van categorieën die direct of indirect verwijzen naar de beschermde gronden zoals burgerlijke staat, geslacht, godsdienst/levensovertuiging, hetero- of homo seksuele gerichtheid, nationaliteit, politieke gezindheid, ras/etniciteit.
- Controleer bij de integratie en aggregatie van data wederom welke gevolgen de integratie en aggregatie van data hebben voor categorieën die direct of indirect verwijzen naar de beschermde gronden zoals burgerlijke staat, geslacht, godsdienst/levensovertuiging, hetero- of homoseksuele gerichtheid, nationaliteit, politieke gezindheid, ras/etniciteit. Zelfs het samenvoegen van twee neutrale databases kan samen een database met sterke bias opleveren.
- Controleer welke gevolgen het labelen van data heeft voor de categorieën die direct of indirect verwijzen naar de beschermde gronden zoals burgerlijke staat, geslacht, godsdienst/levensovertuiging, hetero- of homoseksuele gerichtheid, nationaliteit, politieke gezindheid, ras/etniciteit.

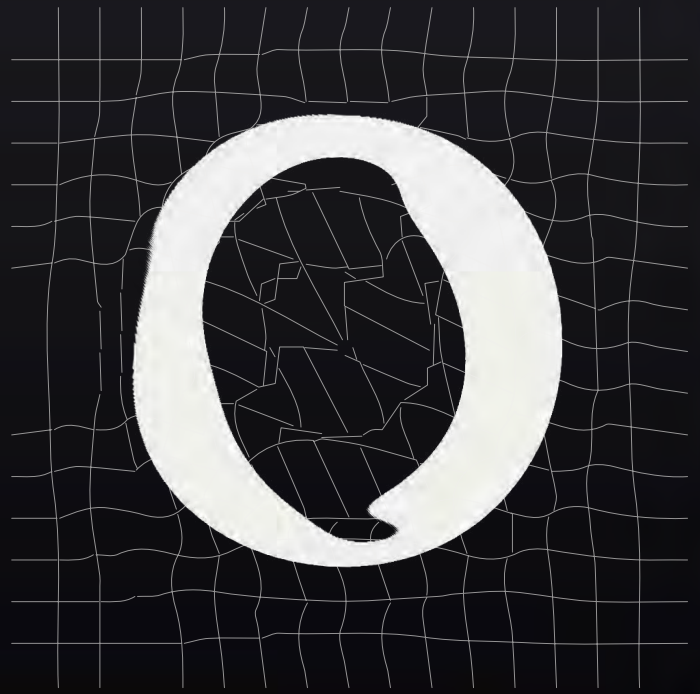
Technisch

- Welke technische beperkingen brengen de data met zich mee? Doe aanpassingen aan het technisch systeem al naar gelang de datakwaliteit dat vereist. Heeft de datakwaliteit gevolgen voor fase 1 (probleemdefinitie)?

- Zijn de data die worden gebruikt op dezelfde wijze verzameld?
- Welke invloed heeft het verschil in de wijze (methodologie, tijd, plaats, etc.) van de data-verzameling op de mogelijkheid van integratie en vergelijkbaarheid van de data?
- Welke categorieën worden gekozen bij het structureren van de data en waarom?
- Is het aannemelijk dat de relatie tussen de features en de target-variabele anders is voor verschillende groepen? Vermijd in dit geval een "*one-size-fits-all*" model; deze werkt in dergelijke gevallen voor geen van de groepen goed, of enkel voor de meerderheidsgroep. *Voorbeeld:* de relatie tussen hemoglobinelevels en diabetes verschilt tussen genders en etniciteiten. Wanneer dit niet wordt meegenomen in het modelleerproces zullen de voorspellingen niet accuraat zijn.

- Controleer op measurement bias. De target variabele (bijv. beoordeling van een manager) meet een bepaald construct (bijv. kwaliteit van de medewerker). De kwaliteit van de target variabele kan echter verschillen tussen groepen, wat kan leiden tot bevooroordeelde data.
- Houd er ook rekening mee dat de granulariteit en kwaliteit van data tussen verschillende groepen kan verschillen en dat elke classificatie een versimpeling is van de werkelijkheid.

Organisatorisch



- Leg nauwgezet de keuzes vast ten aanzien van inclusie en exclusie. Welke rationale is er gekozen voor de inclusie of exclusie van de data en waarom?
- Welke gevolgen heeft de kwaliteit en representativiteit voor de werkzaamheid van het systeem en de gestelde succescriteria? Verzamel eventueel meer gegevens als blijkt dat bepaalde groepen over- of onder gerepresenteerd zijn: ga dan terug naar fase 2.

- Datasets, analysemethoden en beslissingen worden gekozen met het oog op objectiviteit; fouten worden vastgelegd en direct gecorrigeerd. Als ze van belang zijn voor andere organisaties worden die onmiddellijk gedeeld.
- Methodes, procedures, definities en classificaties worden consistent toegepast; deze worden zo veel mogelijk gestandaardiseerd tussen organisaties, om vergelijkingen en controle mogelijk te maken.

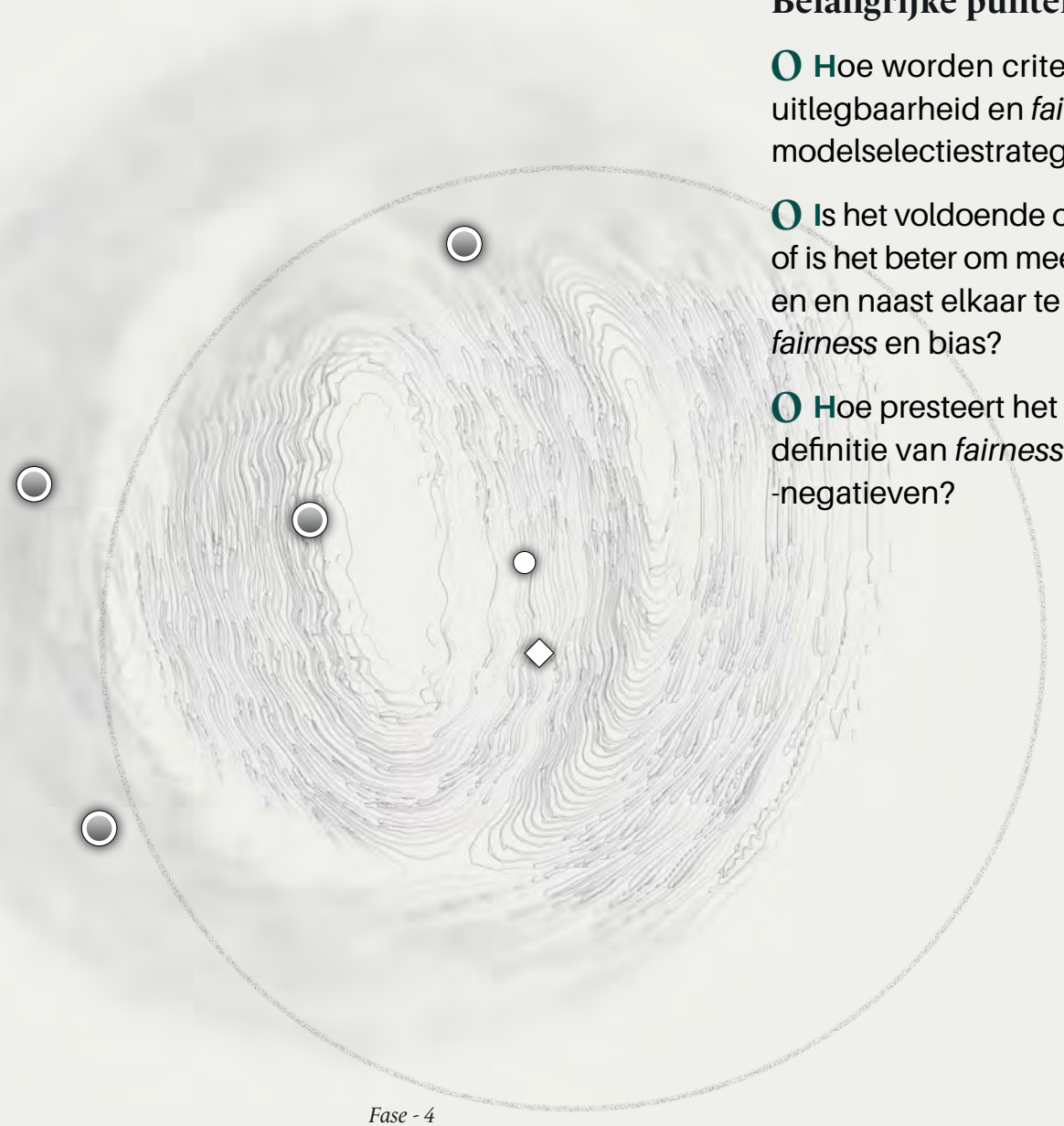
- Controleer de gevolgen van het labelen op de kwaliteit en representativiteit en de werkzaamheid van het systeem.
- Verzamel eventueel meer gegevens als blijkt dat bepaalde groepen over- of onder gerepresenteerd zijn: ga dan terug naar fase 2.

fase
4 - Modellering

Pre-modellering

Model (selectie)

Test



Belangrijke punten in deze fase:

- Hoe worden criteria op het gebied van uitlegbaarheid en *fairness* vertaald naar de modelselectiestrategie?
- Is het voldoende om één model te bouwen, of is het beter om meerdere modellen te bouwen en naast elkaar te leggen om te toetsen op *fairness* en bias?
- Hoe presteert het model op de gekozen definitie van *fairness* en foutpositieven en -negatieven?

Pre-modellering

1. Welk algoritme wordt er gekozen en waarom?

2. Welk modeltype wordt er nagestreefd en waarom?

3. Hoe worden criteria op het gebied van uitlegbaarheid en *fairness* vertaald naar de modelselectiestrategie?

Model(selectie)

4. Welke parameters worden er voor het model gekozen en waarom?

5. Is het voldoende om één model te bouwen, of is het beter om meerdere modellen te bouwen en naast elkaar te leggen?

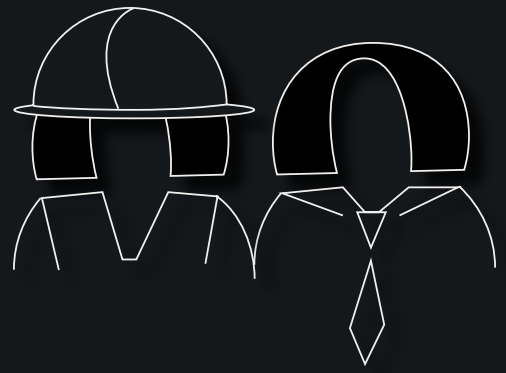
6. Is het model gebaseerd op bestaande modellen en waarom wel of niet?

Test

7. Hoe presteert het model op effectiviteit?

8. Hoe presteert het model op de gekozen *fairness*definitie(s)?

9. Hoe presteert het model op de gekozen succescriteria ten aanzien van foutpositieven en foutnegatieven?



Er wordt gekozen voor *supervised* algoritmen, er worden regels meegegeven waaraan het systeem moet toetsen en op basis van scores van eerdere kandidaten, die zijn aangenomen, wordt een keuze gemaakt voor een nieuwe sollicitant. Er wordt gekozen voor een *decision tree*, omdat dit type het simpelste en voor dit doeleinde het meest efficiënte model is.

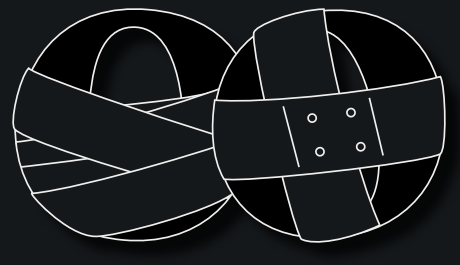
Er wordt gebruik gemaakt van slechts één model, namelijk een bestaand model van een externe partij, dat in het verleden reeds zijn effectiviteit heeft bewezen in soortgelijke toepassingen.

De effectiviteit van het model blijkt uit de ratio hoeveel sollicitanten die volgens het systeem als kanshebbend worden gecategoriseerd, ook door een extern panel als zodanig worden gezien. Datzelfde geldt voor een aselecte steekproef voor door het systeem afgewezen kandidaten.

Een *supervised* lerend algoritme wordt gebruikt om op basis van tijd en plaats te kunnen voorspellen wanneer er een bepaald soort overtreding zal plaatsvinden. Het model genereert een faire uitkomst wanneer daaruit blijkt dat geen van de verdachte discriminatiegronden disproportioneel benadeeld worden, direct of indirect.

Er wordt gekozen om meerdere modellen te testen, gegeven de gevoelige materie. Het model wordt intern ontwikkeld, gegeven de gevoelige materie en toepassingsgebied.

Het model presteert met een effectiviteit van 70%. Het model presteert overeenkomstig de definitie van *fairness*. Het model presteert binnen de aanvaardbare range van foutpositieven en foutnegatieven.

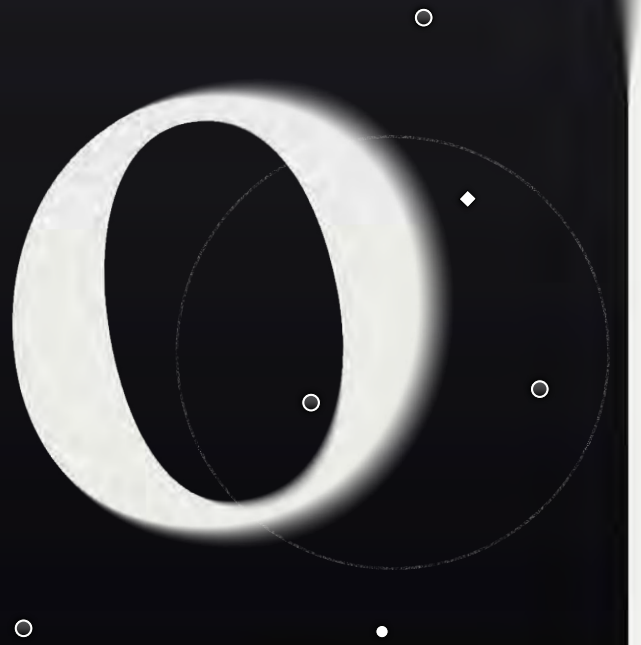


Er wordt gekozen voor een zelflerend algoritme; er worden regels meegegeven waaraan het systeem moet toetsen en op basis waarvan in historische gegevens bepaalde patronen worden herkend. Uiteindelijk zijn efficiëntie en correctheid van de voorspellingen belangrijker dan de begrijpelijkheid en controleerbaarheid van het AI-systeem. Toch is juist in de medische sector (begrijpelijke) informatie voor patiënten van essentieel belang. Daarom is begrijpelijkheid voor de patiënt een minimumvoorwaarde voor het model.

Er wordend rie modellen ontwikkeld en naast elkaar gelegd om te testen op effectiviteit, nauwkeurigheid en bias. Deze worden ontwikkeld door drie aparte onderzoeksteams binnen de aan het experiment verbonden organisaties. De functionaliteit wordt door een extern panel gevalideerd.

De effectiviteit van het model blijkt uit de ratio hoeveel mensen die volgens het systeem als *true positive* worden aangemerkt, ook daadwerkelijk *true positive* zijn en mensen die als *true negative* worden aangemerkt, ook *true negative* zijn. Deze assessment wordt gedaan door een team van onafhankelijke artsen.

Juridisch



- Wordt er gekozen voor een model gebaseerd op causaliteit of correlatie? Wees bewust van het feit dat het juridische domein is gebaseerd op causaliteit; juridische uitlegbaarheid en juridische rechtvaardiging kunnen vaak niet worden gevonden in statistische correlaties. Ga dus na of bij een systeem gebaseerd op correlatie, de uitkomsten kunnen worden gekoppeld aan causaliteit op individueel niveau. *Deep-learning* systemen met black box elementen die besluiten nemen zijn bijna altijd verboden als die besluiten burgers in aanmerkelijke mate treffen.

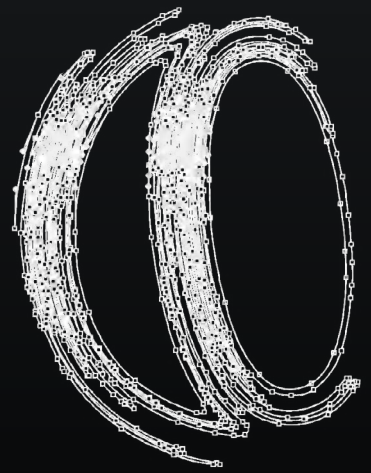
- AI-systemen die zijn gebaseerd op statistische modellen kunnen rekening houden met de principes die hiervoor leidend zijn, zoals:
  - Betrouwbaarheid
  - Onpartijdigheid
  - Objectiviteit
  - Vergelijkbaarheid
  - Consistentie
- Klik hier voor meer *informatie.*

- Controleer hoe het model presteert t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid.

Technisch

- **A**lgoritmeselectie:
  - Verklaar en documenteer keuzes voor een AI-algoritme in het licht van interpreteerbaarheid en verklaarbaarheid.
- **M**odelselectie:
  - Overweeg de vooraf gedefinieerde *fairness*-maatstaven bij het selecteren van een model.
  - Controleer of aangeleerde relaties een afspiegeling zijn van bestaande domeinexpertise en niet willekeurig zijn.
  - Gebruik het eenvoudigste model dat de prestatiespecificaties mogelijk maakt.
- **S**electeer de kandidatenpoel van AI-algoritmes in het licht van benodigde *fairness*, interpreteerbaarheid, en uitlegbaarheid. *Voorbeeld:* wanneer *fairness* criteria gesteld zijn, overweeg het gebruik van *in-processing of post-processing bias mitigation* methodes om hiervoor te optimaliseren.
- Overweeg het gebruik van *unfairness mitigation* technieken om te optimaliseren voor een *fairness metric*. Klik *hier* en *hier* voor tools. *Voorbeeld:* wanneer causaliteit een voorwaarde is, wordt gezorgd dat de kandidatenpoel bestaat uit direct interpreteerbare algoritmes, zoals *linear regression*. *Voorbeeld:* wanneer voorspellingen als doel hebben om menselijke beslissingen te ondersteunen, wordt gezorgd dat de kandidatenpoel bestaat uit methodes die uitlegbaar zijn voor mensen met een minder technische achtergrond, zoals een simpele *decision tree*.
- **V**ertaal de succescriteria naar technische maatstaven, bijvoorbeeld op het gebied van accuraatheid, foutpositieven, foutnegatieven, en *fairness*. *Voorbeeld:* om te voorkomen dat het model minder goed werkt voor minderheden, zal de *fairness metric equalized odds* worden meegenomen in de modelselectie.
- **I**ndien gebruik wordt gemaakt van bestaande modellen, neem deze dan mee in de modelselectiestrategie. *Voorbeeld:* in een NLP(Natural Language Processing)-applicatie wordt het gebruik van *pre-trained word embeddings* overwogen, de evaluatie van de *embeddings* op de succescriteria is daarom onderdeel van de modelselectiestrategie.

Technisch



- **B**ias-beperking:
- Overweeg het gebruik van bias mitigatietechnieken. Klik *hier* en *hier* voor voorbeelden.
- Pas indien nodig nabewerkingsmethoden toe om bias te beperken na training van het classificatiemodel. White-box-methoden passen het model aan; black-box-methoden passen de voorspellingen aan. *Voorbeeld:* wanneer voorspellingen als doel hebben om menselijke beslissingen te ondersteunen, wordt gezorgd dat de kandidatenpoel bestaat uit methodes die uitlegbaar zijn voor mensen met een minder technische achtergrond, zoals een simpele *decision tree.*
- Controleer hoe de modellen presteren t.a.v. categorieën die direct of indirect verwijzen naar iemands geslacht, ras, kleur, taal, godsdienst, politieke of andere mening, nationale of maatschappelijke afkomst of het behoren tot een nationale minderheid.
- Selecteer het eenvoudigste model dat de prestatiespecificaties mogelijk maakt. *Voorbeeld:* zowel een *logistic regression* model als een *random forest* model zijn voldoende accuraat, er wordt daarom gekozen voor het *logistic regression* model.

- **H**oe presteert het model op de succescriteria van effectiviteit, de gekozen *fairness*definitie en de succescriteria van foutpositieven en foutnegatieven?
- **H**oe zou het systeem functioneren met een ander model, *fairness*definitie en/of algoritme? Stel het model bij aan de hand van de uitkomsten.
- **C**ontroleer op evaluatie bias: dit kan voorkomen bij het testen van het model. Let erop dat de succescriteria die gebruikt worden om het systeem te testen overeenstemmen met de beoogde doelgroep.
- Wanneer het model gebruik maakt van persoonsgegevens, neem dan een vergelijking van de prestaties van het model en de prestaties binnen het huidige besluitvormingsproces op in de evaluatiestrategie. *Voorbeeld:* de prestaties van het model worden getoetst in een *pilot study* bij een kleine groep gebruikers.

Technisch

- Evalueer het model op:
  - Effectiviteit
  - Fairness
  - Foutpositieven en foutnegatieven
- Voer verbeteringen door.
- Beoordeel wat de context van de gekozen testcase zegt over de algemene werking van het systeem en denk na in hoeverre andere toepassingsgebieden andere contextgevoeligheden met zich brengen. Neem daarop passende maatregelen.

Organisatorisch

- De modelselectiestrategie wordt vastgelegd, gedocumenteerd en waar mogelijk openbaar gemaakt. Het ontwerp is het liefst universeel, zodat de modellen goed met elkaar kunnen worden vergeleken op uitkomst en *fairness*.
- Betrek belanghebbenden, zoals eindgebruikers en besluitvormers, bij de selectie van de kandidatenpoel om te zorgen dat de interpreteerbaarheid en uitlegbaarheid van de modellen past bij het doel van het model en de achtergrond van de gebruikers.

- Verantwoording:
  - Metadata worden bewaard en vastgelegd
  - Data zijn voor zover mogelijk toegankelijk voor derden
  - Het model moet uitlegbaar zijn aan en inzichtelijk voor belanghebbenden
  - Welk vorm van uitlegbaarheid wordt door het systeem geboden?
  - Voor wie is deze uitleg begrijpelijk?

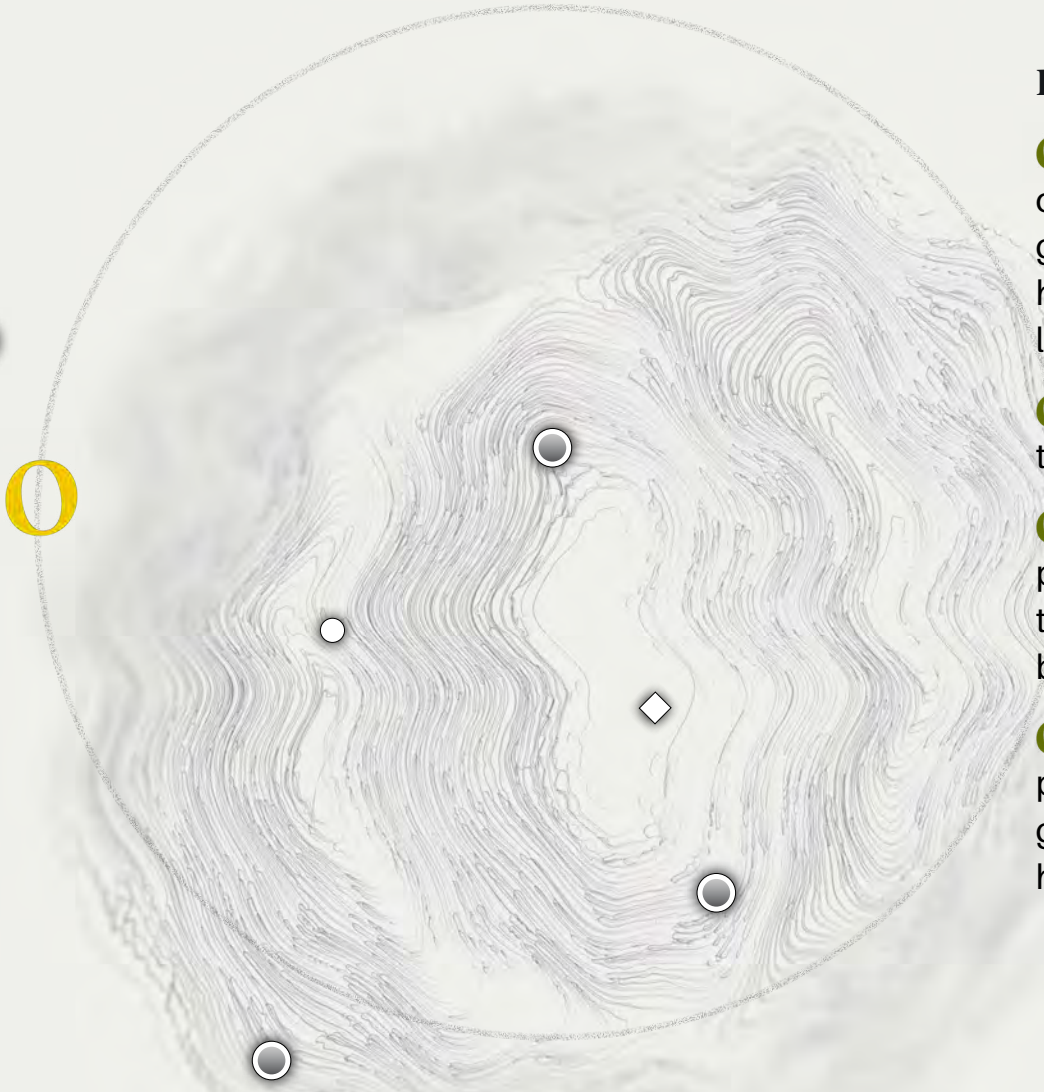
- Vraag een onafhankelijk team van experts met een diverse persoonlijke en professionele achtergrond voor een *second opinion*.
- Documenteer het gekozen model en de resultaten van de evaluatie, bijvoorbeeld in een *model card*.

fase
5 - Implementatie

Praktijk-test

Aanpassing model

Toepassing



Belangrijke punten in deze fase:

- Selecteer een afgebakende toepassing om het systeem te testen; zorg dat deze afgebakende toepassing representatief is voor het gehele domein waarop het AI-systeem later wordt ingezet.
- Pas het model aan op basis van de resultaten van de testcase.
- Pas de verwachtingen aan voor de toepasbaarheid van het systeem op basis van de testcase; welke mogelijke toepassingen vallen bijvoorbeeld af?
- Documenteer de mogelijkheden en beperkingen van het systeem en informeer gebruikers over de condities waaronder het systeem gebruikt kan worden.

Praktijktest

1. Wat is de toepassingsstrategie?

2. Welke beperkte en afgebakende testcase is representatief en kan goed worden gemonitord?

3. Hoe werkt het model binnen de gekozen testcase en is dat volgens verwachting?

Aanpassing model

4. Welke aanpassingen zijn er nodig om de werkzaamheid te verhogen?

5. Welke aanpassingen zijn er nodig om de *fairness* van het model te verhogen?

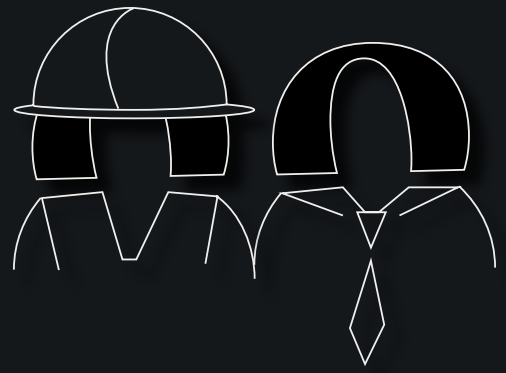
6. Welke aanpassingen zijn er nodig om de foutmarges te verkleinen?

Toepassing

7. Welke beperkingen volgen uit de vorige stappen voor de toepassingsmogelijkheden en het implementatietraject voor het breed uitrollen van het systeem?

8. Welke aandachtspunten zijn er voor de toepassing en hoe kan er bij de implementatie voor worden gezorgd dat deze goed kunnen worden gemonitord?

9. Hoe worden belanghebbenden en anderen op de hoogte gesteld van en betrokken bij de implementatie van het systeem?



Als eerste zal de sollicitatiebrief en het CV kwantificeerbaar worden gemaakt worden voor het model. Daarna worden relevante variabelen geïdentificeerd. Op basis van deze variabelen worden de sollicitanten vergeleken, waarbij alleen de variabelen worden gebruikt waarin geen bias naar voren komt. Als representatieve testcase wordt gekeken naar huidige werknemers die via het systeem zijn aangenomen, waarbij wordt gelet of de werkgever tevreden is over de *true positives.*

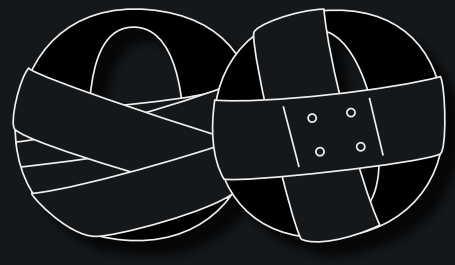
Het model wordt aangepast naarmate er meer gegevens uit sollicitatieprocedures kunnen worden gehaald. Door de toename van informatie uit deze procedures kan het model beter presenteren. Naarmate er meer gegevens uit de procedure worden gehaald, kunnen deze ook leiden tot het verhogen van de *fairness* in het model. De variabelen die leiden tot bias worden eruit gefilterd zodat deze geen deel uitmaken van het verdere proces.

Een CV bevat beperkte informatie, daarnaast is een sollicitatiebrief lastig en op meerdere manieren te kwantificeren. Hierdoor is het aantal variabelen beperkt en is de weging daarvan subjectief. Er zal een tweede test worden ingericht waarbij deze data wordt verrijkt met informatie over sollicitanten uit openbare bronnen. Daarna zal worden beoordeeld of dit model beter presteert. Belanghebbenden worden voorafgaand aan het proces op de hoogte gesteld van hun rechten en het feit dat er gebruik wordt gemaakt van een AI-systeem.

Daar het model een voorspelling maakt van waar en wanneer de kans het grootst is dat een misdrijf zal plaatsvinden, zullen daar preventief patrouilles heen worden gestuurd. Daarbij wordt niet volledig vertrouwt op het algoritme in de zin dat er waakzaam wordt omgegaan met de uitkomst. Ook andere buurten zullen dus in het oog gehouden worden om geen misdrijven te missen. Na verloop van tijd zal er worden gekeken in welke mate het model een goede of slechte indicatie is voor het zich voordoen van misdrijven. Als testcase wordt het aantal inbraken en het aantal moorden voor een bepaalde periode in een afgebakend territorium gemonitord.

Meer data zijn nodig om de voorspellende kracht te verhogen, bij voorkeur niet-gevoelige data. Een audit van het systeem moet duidelijk maken welke data geen voorspellende waarde hebben, die moeten er vervolgens worden uitgehaald.

Uit de test bleek dat de voorspellende werking van het model goed werkte op veelvoorkomende criminaliteit, zoals diefstal, maar niet goed op zwaardere criminaliteit, zoals levensdelicten. Daarom wordt er voor gekozen om het systeem, in ieder geval in aanvang, slechts toe te passen voor het eerste soort criminaliteit. Over drie jaar wordt geëvalueerd of het systeem inmiddels beter is in het voorspellen van de zwaardere criminaliteit.



Op basis van historische gegevens worden de risicogroepen in kaart gebracht. Als representatieve testcase wordt gekeken naar de huidige cases die binnenkomen bij een ziekenhuis in Amsterdam, dat een gemêleerde patiëntenpopulatie heeft.

Er hoeven niet veel aanpassingen worden gemaakt in het model, omdat er veel historische gegevens beschikbaar zijn uit eerdere cases die binnen zijn gekomen. Door middel van specifiekere categorisering kunnen de data verrijkt worden, wat kan zorgen voor een betere voorspelling. Dit geldt ook voor de foutmarges.

De historische data reiken ver, waardoor er weinig beperkingen optreden. Echter, dit verhoogt ook de kans op *overfitting,* te veel regels die worden meegegeven aan het model waardoor het model te veel regels meegeeft, wat leidt tot verkeerde constateringen. Als aandachtspunt kan *overfitting* worden tegengegaan door het gebruik van *pruning.* Belanghebbenden worden voorafgaand aan het proces de hoogte gesteld van hun rechten en het feit dat er gebruik wordt gemaakt van een AI-systeem.

Juridisch

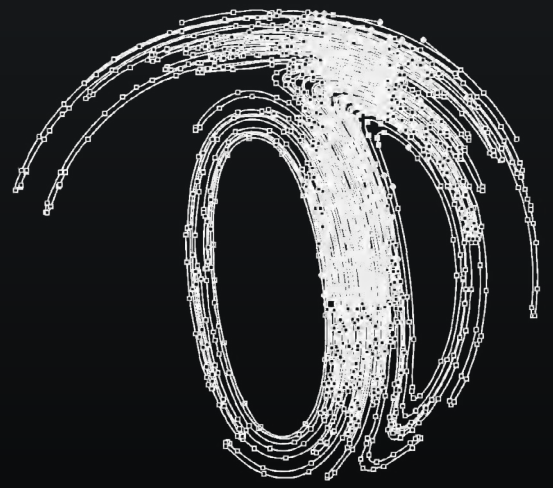
- Stel de belanghebbenden op de hoogte van hun rechten:
  - Recht op informatie, inclusief informatie over het algoritme
  - Recht om het besluit aan te vechten
  - Recht om aanvullende informatie aan te dragen
  - Recht om niet onderworpen te worden aan automatische besluitvorming
- Klik hier voor meer *informatie.*

- Controleer hoe het model presteert t.a.v. categorieën die direct of indirect verwijzen de beschermde gronden zoals burgerlijke staat, geslacht, godsdienst/levensovertuiging, hetero- of homoseksuele gerichtheid, nationaliteit, politieke gezindheid, ras/etniciteit.
- Als het model significant onderscheid maakt, direct of indirect, op basis van een van deze groepen, is daar een rechtvaardiging voor en zo ja welke? Leg dit voor aan een jurist.

Technisch

- Beantwoord de volgende vragen:
  - Wat is de toepassingsstrategie?
  - Welke beperkte en afgebakende toepassing is representatief en kan goed gemonitord worden?
  - Hoe werkt het model en is dat volgens verwachting?
  - Wat is de exitstrategie?
- Overweeg gebruikerstesten om gebruiksgemak en toegankelijkheid voor mensen met een beperking te toetsen.
- Wanneer het model regelmatig wordt geüpdatet op basis van nieuwe data, zorg dat ook deze nieuwe modellen grondig worden geëvalueerd.
- Plan voor het monitoren van veranderingen in de datadistributie, zoals concept drift en een verschuiving in de demografie van data subjecten.

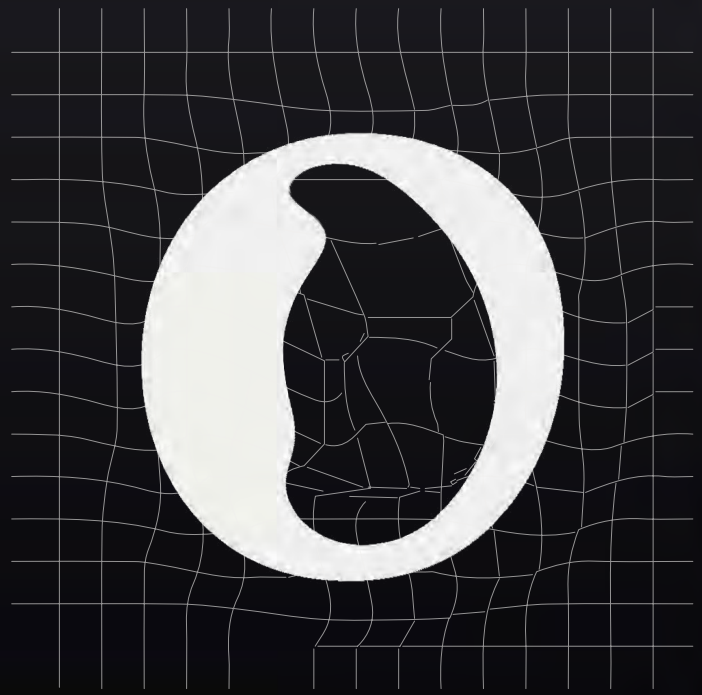
Technisch



- Evalueer hoe het model in de testcase presteert op:
  - Effectiviteit
  - *Fairness*
  - Foutpositieven en foutnegatieven
- Voer verbeteringen door.
- Beoordeel wat de context van de gekozen testcase zegt over de algemene werking van het systeem en denk na in hoeverre andere toepassingsgebieden andere contextgevoeligheden met zich brengen. Neem daarop passende maatregelen.

- Monitoring en controle:
  - Plan evaluatiemomenten.
  - Documenteer de gebruiksresultaten.
  - Controleer op exitcriteria.
  - Maak API's beschikbaar voor externe auditors.
  - Maak databladen en modelkaarten voor zover mogelijk openbaar.

Organisatorisch



- Stel betrokkenen en belanghebbende op de hoogte van het feit dat het AI-systeem in een testsetting wordt geïmplementeerd. Doe dat waar mogelijk voordat zij in aanraking komen met of de gevolgen ondervinden van het systeem.

- Overleg met zowel belanghebbenden als externe experts over de functionaliteit, *fairness* en accuratesse van het model.
- Implementeer voor zo ver mogelijk hun adviezen en suggesties.
- Leg veranderingen in het systeem vast.
- Documenteer het gekozen model en de resultaten van de evaluatie.

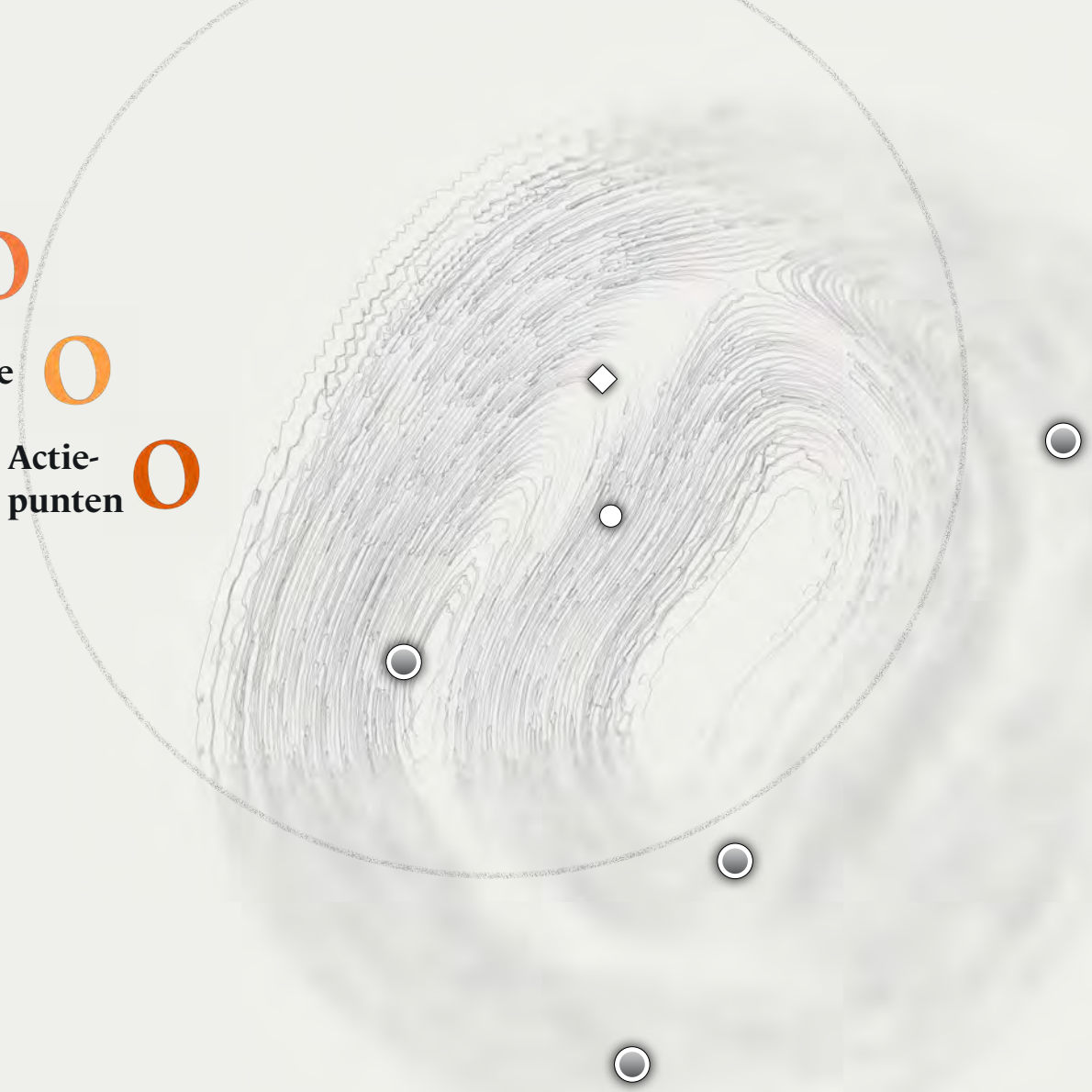
- Stel iedereen op de hoogte als het AI systeem wordt geïmplementeerd buiten de aanvankelijke testsetting en voer een klachtenprocedure in.
- Vraag een onafhankelijk team van experts met een diverse persoonlijke en professionele achtergrond voor een *second opinion.*

fase
6 - Evaluatie

Evaluatie-voorbereiding

Evaluatie

Actie-punten



Belangrijke punten in deze fase:

- Kies een implementatiestrategie en formuleer een evaluatiestrategie. Betrek bij voorkeur externen bij de evaluatie.
- Beoordeel hoe het systeem zou functioneren met een ander model, *fairness*definitie en/of algoritme.
- Beoordeel na evaluatie of het systeem moet worden stopgezet, aangepast of doorgang kan vinden.

Evaluatievoorbereiding

1. Wordt er gekozen voor een permanente evaluatie, specifieke evaluatiemomenten of beide en waarom?

2. Wordt er gekozen voor een interne evaluatie, een evaluatie door externen of beide en waarom?

3. Hoe wordt de evaluatie getest en met welke meetpunten?

Evaluatie

4. Hoe functioneert het systeem ten aanzien van de succescriteria?

5. Welke aanpassingen zijn er nodig ten aanzien van de beschermde categorieën?

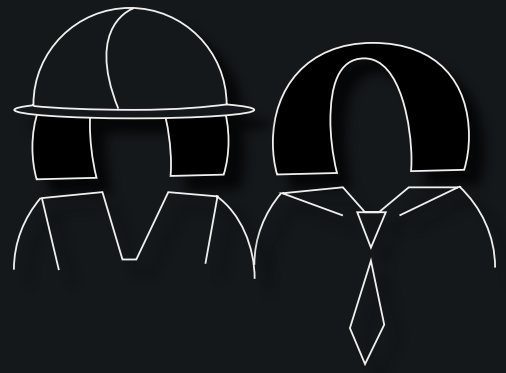
6. Hoe zou het systeem functioneren met een ander model, *fairness*definitie en/of algoritme?

Actiepunten

7. Moet het systeem al dan niet tijdelijk worden stopgezet?

8. Kunnen gevonden problemen en obstakels worden verholpen?

9. Wat vinden belanghebbenden en externe experts van de evaluatieresultaten?



Er wordt gekozen voor specifieke evaluatie-momenten, steeds na het proces waarin een vacature is vervuld. Op een later moment zal een onafhankelijk team de gegevens van de kandidaten met de hoogste scores bekijken en fouten eruit filteren.

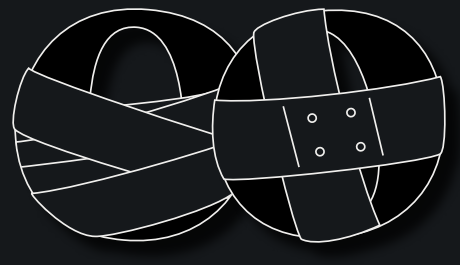
De oude en de nieuwe situatie worden met elkaar vergeleken; hierbij wordt gekeken naar de tevredenheid over de aangenomen kandidaten en wat het verschil is tussen deze twee. De beschermde categorieën worden in het model opgegeven als restrictie, dus deze zullen buiten beschouwing gelaten worden. Wel zullen deze data worden bewaard om te kunnen testen of het systeem niet indirect discrimineert.

Het systeem zal (tijdelijk) stopgezet worden als blijkt dat in het systeem een ongewenste bias zit.

Er is gekozen om een extern team van specialisten permanent het systeem, de uitkomsten en eventuele klachten te laten beoordelen. In dit team zitten twee juristen, twee data-analisten en twee voormalige politieagenten.

Er zijn paralleltesten uitgevoerd met zowel andere modellen, als andere algoritmen, als andere *fairness*definities. Die testen laten zien dat er in het algemeen de minste bias zit in de gekozen definities en modellen; toch preseren de andere definities en modellen op deelaspecten beter. Daarom wordt het huidige systeem op punten aangepast.

Evalueren van de ideeën van belanghebbenden is in dit geval lastig. Wel zullen er twee belangengroepen worden gevraagd kritisch mee te kijken, namelijk Amnesty International en Bits of Freedom. Uit onze interne evaluatie kwam naar voren dat het systeem in bestaande vorm kan worden gecontinueerd, maar dat een permanente vinger aan de pols noodzakelijk is.



Door de redelijk constante stroom aan gegevens over patiënten wordt er gekozen voor een permanente evaluatie, zodat deze constante stroom regelmatig wordt geëvalueerd. Naderhand wordt er gekozen voor een interne evaluatie; deze artsen en statistici zullen kijken naar de resultaten en controleren of er geen ongerechtvaardigde bias in het systeem zit.

Er wordt gekeken of door de signalering van de risicogroepen meer mensen geholpen kunnen worden. Op basis van de resultaten wordt gekeken naar de mogelijkheid om de voorspellingen van het systeem te vertalen naar preventiebeleid.

Het systeem zal (tijdelijk) stop worden gezet als blijkt dat door het systeem slechtere keuzes worden gemaakt, daarna zal worden achterhaald welke variabelen verkeerde voorspellingen veroorzaken. Gevonden problemen en obstakels kunnen worden verholpen door het toevoegen van meer restricties. Naast de variabelen die al worden meegenomen, worden meer variabelen toegevoegd om het probleem uit te sluiten.

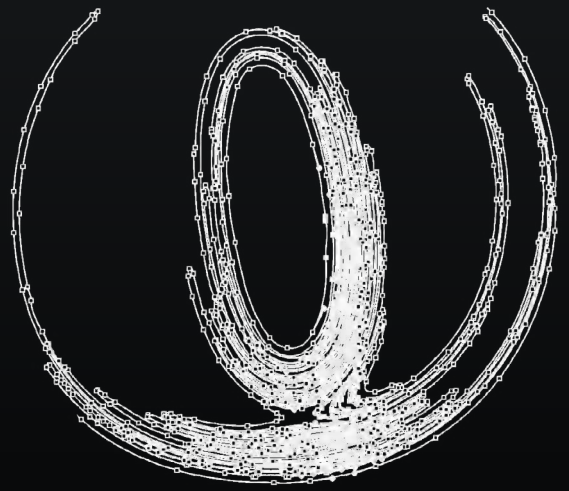
Juridisch

- Als er persoonsgegevens worden verwerkt moet het systeem duidelijk effectiever zijn dan de status quo zonder het AI-systeem om te voldoen aan de noodzakelijkheids-, proportionaliteits- en subsidiariteitstest. Verantwoord dit.

- Controleer hoe het systeem presteert t.a.v. categorieën die direct of indirect verwijzen naar de beschermde gronden zoals burgerlijke staat, geslacht, godsdienst/levensovertuiging, hetero- of homoseksuele gerichtheid, nationaliteit, politieke gezindheid, ras/etniciteit.
- Als het model significant onderscheid maakt op basis van een van deze groepen, is daar een rechtvaardiging voor, en zo ja, welke? Ga na of het systeem al dan niet tijdelijk moet worden stopgezet of dat naar een van de vorige fasen in het proces moet worden teruggegaan om aanpassingen te doen.

- Vraag een externe jurist om een *second opinion* te doen en advies te geven ten aanzien van de juridische randvoorwaarden voor het systeem.

Technisch



- Leg de volgende punten nauwgezet vast:
  - Hoe wordt de evaluatie ingericht en waarom zo?
  - Waar wordt de verantwoordelijkheid voor de evaluatie belegd en waarom?
  - Welke meetpunten worden gekozen en waarom?

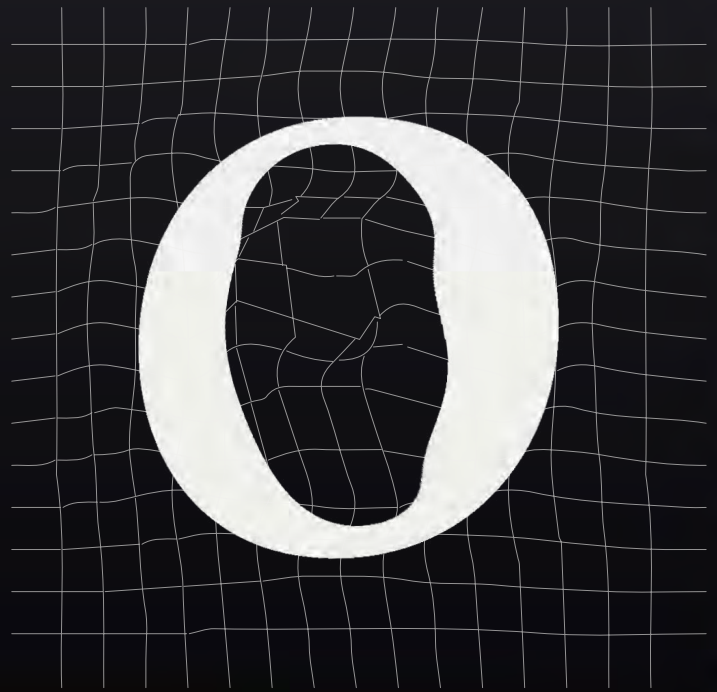
- Evalueer het model op:
  - Effectiviteit
  - Fairness
  - Foutpositieven en foutnegatieven

  Voldoen de prestaties aan de voorgestelde succescriteria? Zo nee, stop het project tijdelijk of permanent. Zo ja, zijn er mogelijkheden om het systeem nog effectiever, rechtvaardiger of accurater te maken?

- Hoe functioneert het systeem:
  - Op andere data?
  - Met een ander algoritme?
  - Met een andere *fairness*definitie?
  - Met een ander model?

- Vraag een externe data-analist om een *second opinion* te doen en advies te geven ten aanzien van het technisch systeem.

Organisatorisch



- **B**etrek belanghebbenden bij de evaluatie. *Voorbeeld:* Doe een survey of focusgroep met belanghebbenden om hun ervaringen in kaart te brengen.
- **G**a na of het systeem, het model, de data en/of de evaluatie openbaar kunnen worden gemaakt, al dan niet in geanonimiseerde vorm.
- **D**ocumenteer de evaluatiemethodes, motivaties voor keuzes, en verantwoordelijkheid voor de evaluatie.
- **V**raag een bedrijfskundige om een *second opinion* te doen en advies te geven ten aanzien van de procesinrichting van het systeem.

Colofon

Tekst

- **Bart van der Sloot** (Tilburg University)
- **Esther Keymolen** (Tilburg University)
- **Merel Noorman** (Tilburg University)
- **Het College voor de Rechten van de Mens**
- **Hilde Weerts** (Eindhoven University of Technology)
- **Yvette Wagensveld** (Tilburg University)
- **Bram Visser** (Vrije Universiteit Brussel)

Ontwerp

- **Julia Janssen** (kunstenaar)
- **Suzan Slinger** (productie manager bij Studio Julia Janssen)

Deze handreiking is opgesteld in opdracht
van het Ministerie van Binnenlandse Zaken.