Non-discriminatie by design

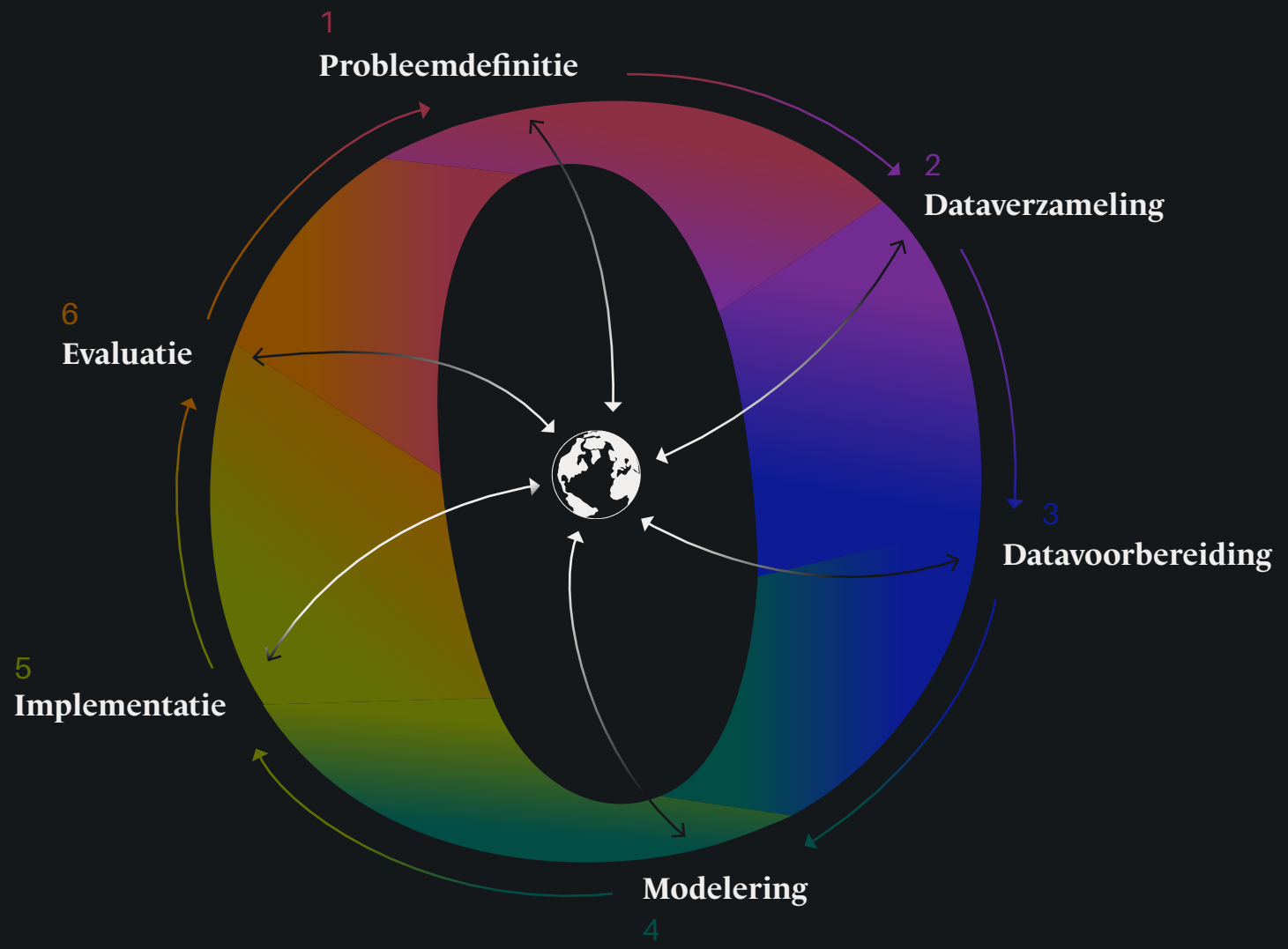
Tekst

- **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.



Samenvatting

Dit is de samenvatting van de Handreiking Non-Discriminatie by Design. De handreiking 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?

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. 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.

In dit document zijn zes stappen onderscheiden die worden doorlopen bij het ontwikkelen van een AI-systeem. Per stap zijn drie clusters vragen geformuleerd die je kunnen helpen om de juiste vragen te doordenken, maatregelen te treffen en acties te ondernemen. Het non-discriminatierecht geeft uitgangspunten, maar geen absolute ge- of verboden. 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'. Het gaat er dus vooral om dat je je bewust bent van het gevaar voor discriminatie en nagaat of het onderscheid tussen groepen noodzakelijk en rechtvaardig is.

Binnen het non-discriminatierecht is niet alleen ‘directe discriminatie’ verboden, maar ook ‘indirecte discriminatie’. Van directe discriminatie kan sprake zijn indien een persoon op een andere wijze wordt behandeld dan een ander in een vergelijkbare situatie. Een voorbeeld is een werkgever die alleen mannen aanneemt of een verhuurder die geen huurders met een migratieachtergrond wil. 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 voorbeeld is een vacaturetekst waarin wordt gezocht naar mensen die langer zijn dan 1 meter 80 (veel meer mannen zijn langer dan 1 meter 80 dan vrouwen).

Direct onderscheid maken tussen groepen kan natuurlijk legitiem zijn, zelfs als dat gebeurt op de in de grondwet genoemde gronden. 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.

Indirecte discriminatie kan bovendien legitiem zijn als het onderscheid ‘objectief gerechtvaardigd’ kan worden. Dat is zo als er sprake is van een legitieme reden om het onderscheid te maken, als het maken van onderscheid proportioneel is en als 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 op etniciteit. Maar een taaleis kan toch gerechtvaardigd zijn, bijvoorbeeld als het gaat om een baan met veel contact met klanten die Nederlands spreken.

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? 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.

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:

<u>Sollicitatiealgoritme</u>

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.

Goede reden

↓ Wettelijke uitzonderingen

↓ Objectieve rechtvaardiging: legitiem doel

↓ Is het middel dat ik gebruik om mijn doel te bereiken:

→ **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 & 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?

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?

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?

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?

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?

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?