Een algoritme checken op bias:

Hoe moeilijk kan het zijn?

door Willy Tadema, Rijks ICT Gilde

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

Even mijzelf introduceren:

- Willy Tadema
- Werkzaam bij het Rijks ICT Gilde
- Consultant op het gebied AI governance, ethiek en compliance
- Lid van de NEN normcommissie AI & Big Data
- Lid van de Joint Technical Committee on AI (JTC21)
- Willy.Tadema@Rijksoverheid.nl

Agenda:

- Wat is een bias assessment?
- Waarom een bias assessment?
- Wat is bias?
- Hoe test je de training- en testdata op bias?
- Hoe test je de modeluitkomsten op bias?
  - Wat is gelijkheid in een specifieke context?
  - Hoe meet en test je bias? Wanneer is het te veel?
- Hoe ga je om met bias die zich pas manifesteert nádat het AI-systeem in productie is genomen?
- Wat zijn de risico's van het verzamelen van beschermde attributen ten behoeve van bias assessments?
- Lessons learned
- Instrumenten, standaarden en best practises

Wat is een bias assessment?

- Een **systematische analyse** van biases in een AI-systeem.
- Het **identificeren** en indien mogelijk **kwantificeren** van bias.
- Het achterhalen van de **oorzaken** van bias.
- Suggesties voor **maatregelen** om bias aan te pakken.

Waarom een bias assessment?

- **Modelprestaties** verbeteren.
- Input voor het bepalen van de impact op **ethische waarden en fundamentele rechten**.
- **Juridische risico's** beheersen.

Wat is bias?

- **Technisch perspectief**:
  - Systematische fout in de modeluitkomsten
  - Negatieve impact op accuraatheid en modelprestaties

- **Ethisch perspectief**:
  - Morele ongelijkheid of vooringenomenheid waarbij bepaalde groepen of individuen systematisch worden bevoordeeld of benadeeld
  - Bedreiging voor belangrijke maatschappelijke waarden als rechtvaardigheid, eerlijkheid en gelijkheid

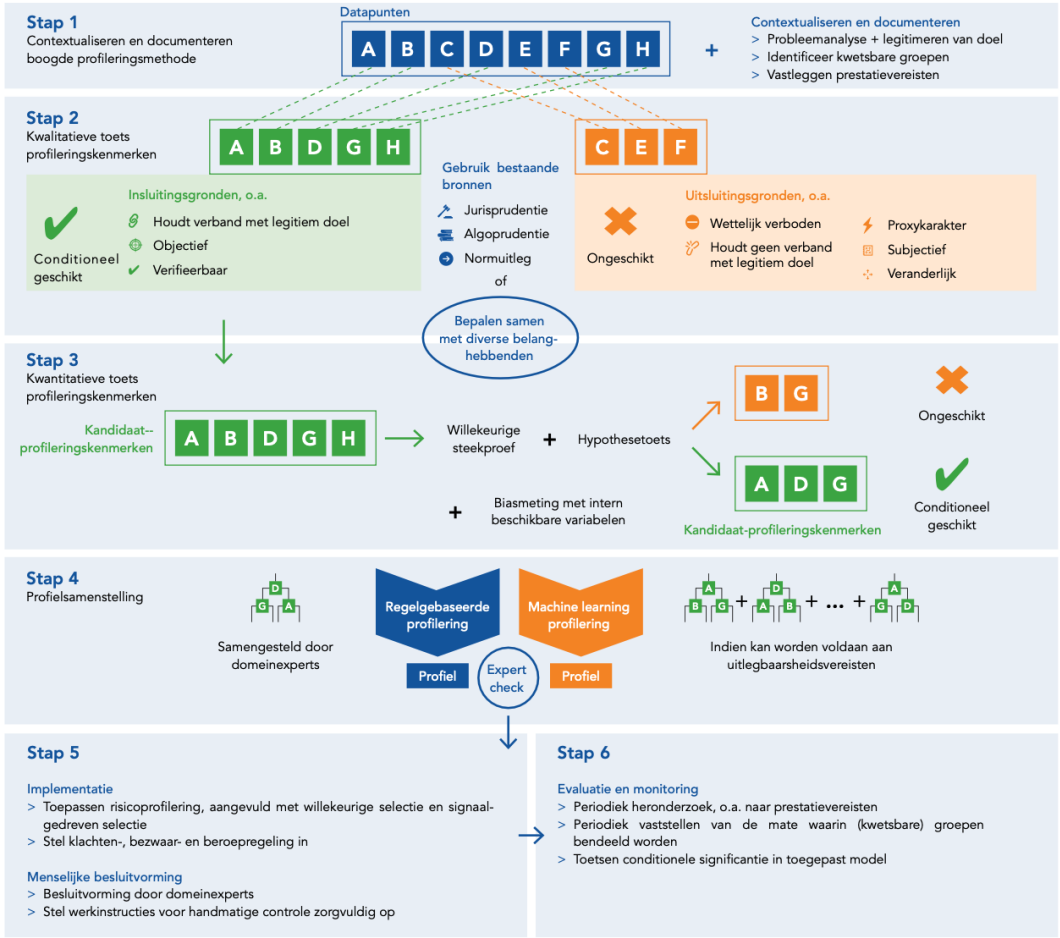
- **Juridisch perspectief**:
  - Vertekening of vooringenomenheid die leidt tot ongelijke behandeling van individuen of groepen in strijd met wet- en regelgeving
  - Non-compliance

- Voor een volledige bias assessment heb je alle drie perspectieven nodig!

Hoe test je de training- en testdata op bias?

- Voorbeelden van **typen bias**:
  - Steekproefbias
  - Historische en maatschappelijke bias
  - Proxy-variabelen die een sterke relatie hebben met een kwetsbare groep

- Pas de volgende **methoden en technieken** toe:
  - Exploratory fairness analysis[1]
  - Impact assessment[2]
  - Kwalitatieve en kwantitatieve toets op voorspellende variabelen[3]



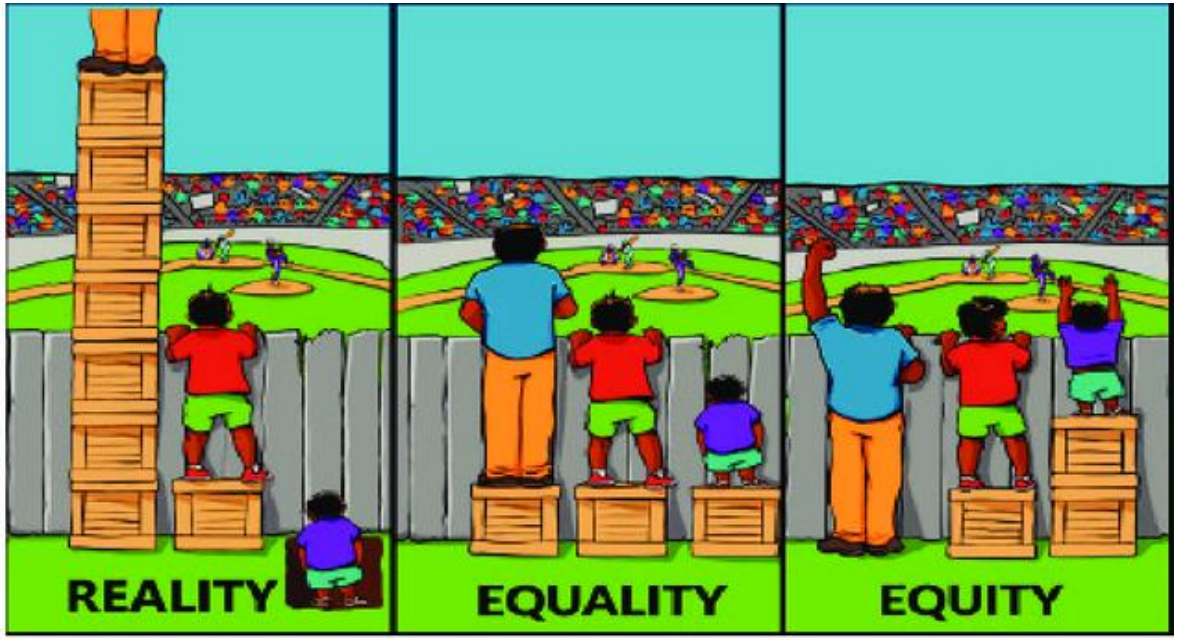
[1] Bijvoorbeeld de *data X-ray* van Tobias Baer

[2] Bijvoorbeeld de *Impact Assessment Mensenrechten en Algoritmen (IAMA)*

[3] Bijvoorbeeld conform de *publieke standaard profileringsalgoritme* van Algorithm Audit (zie afbeelding)

Hoe test je de uitkomsten van het model op bias?

- Bepaal de **definitie van gelijkheid** of 'fairness',
  bijvoorbeeld **kansengelijkheid** ('equality of opportunity') versus **uitkomstengelijkheid** ('equality of outcome')
- Er zijn veel verschillende definities van gelijkheid. Welke definitie de juiste is in een specifieke context,
  is niet alleen een ethische, maar ook een **politieke vraag**.[1]
- Kies een **fairness metriek** die past bij de definitie van gelijkheid. [2]
- Bepaal voor welke (combinaties van) **kwetsbare groepen** je wilt testen.
- Bepaal **drempelwaarden**.
- De juiste fairness metriek en drempelwaarden,
  worden **niet voorgeschreven door wet- en regelgeving**.[3]



[1] Zie *The political philosophy of AI* van Mark Coeckelbergh

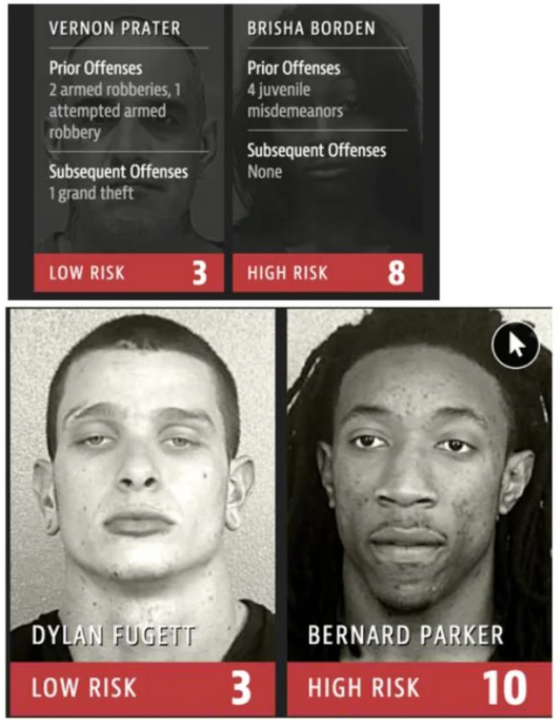
[2] Zie bijvoorbeeld de publicaties van Arvind Narayanan over fairness

[3] Zie *Why Fairness Cannot Be Automated* van Sandra Wachter e.a.

Voorbeeld: de **COMPAS**-casus

- een model uit de VS om het **risico op recidive** te voorspellen
- ProPublica analyseerde de **foutpercentages** en concludeerde dat het model discrimineerde:

  Zwarte verdachten werden vaker onterecht als 'hoog risico' bestempeld, en witte verdachten vaker onterecht als laag risico. [1]
- Northpointe – de ontwikkelaar van COMPAS – hanteerde een andere definitie van fairness en gebruikte andere fairness metrics om te beargumenteren dat het model juist *niet* discrimineerde.



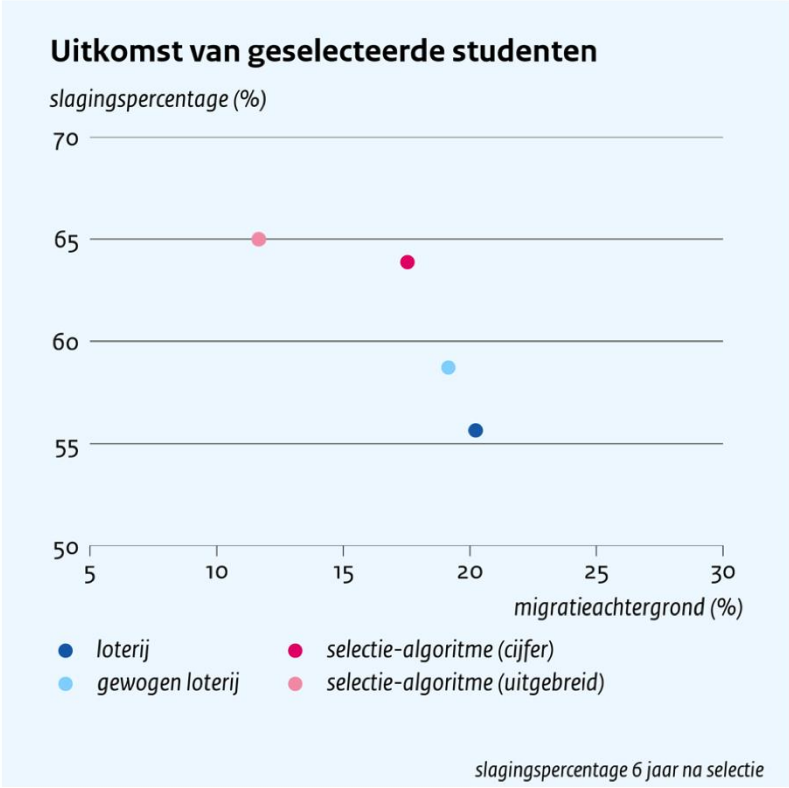
[1] Zie https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Voorbeeld: **selectie van geneeskundestudenten**

- Bron: de publicatie Rechtvaardige Algoritmes van het Centraal Planbureau
- Vergelijken van verschillende selectiemethoden langs twee assen: efficiëntie en representativiteit.
- Metrieken:

  efficiëntie → slagingspercentage 6 jaar na selectie

  representativiteit → percentage met migratieachtergrond



[1] Zie https://www.cpb.nl/sites/default/files/omnidownload/CPB-Publicatie-Rechtvaardige-Algoritmes.pdf

Sommige bias ontstaat pas nadat het AI-systeem in gebruik is genomen door, bijvoorbeeld door:

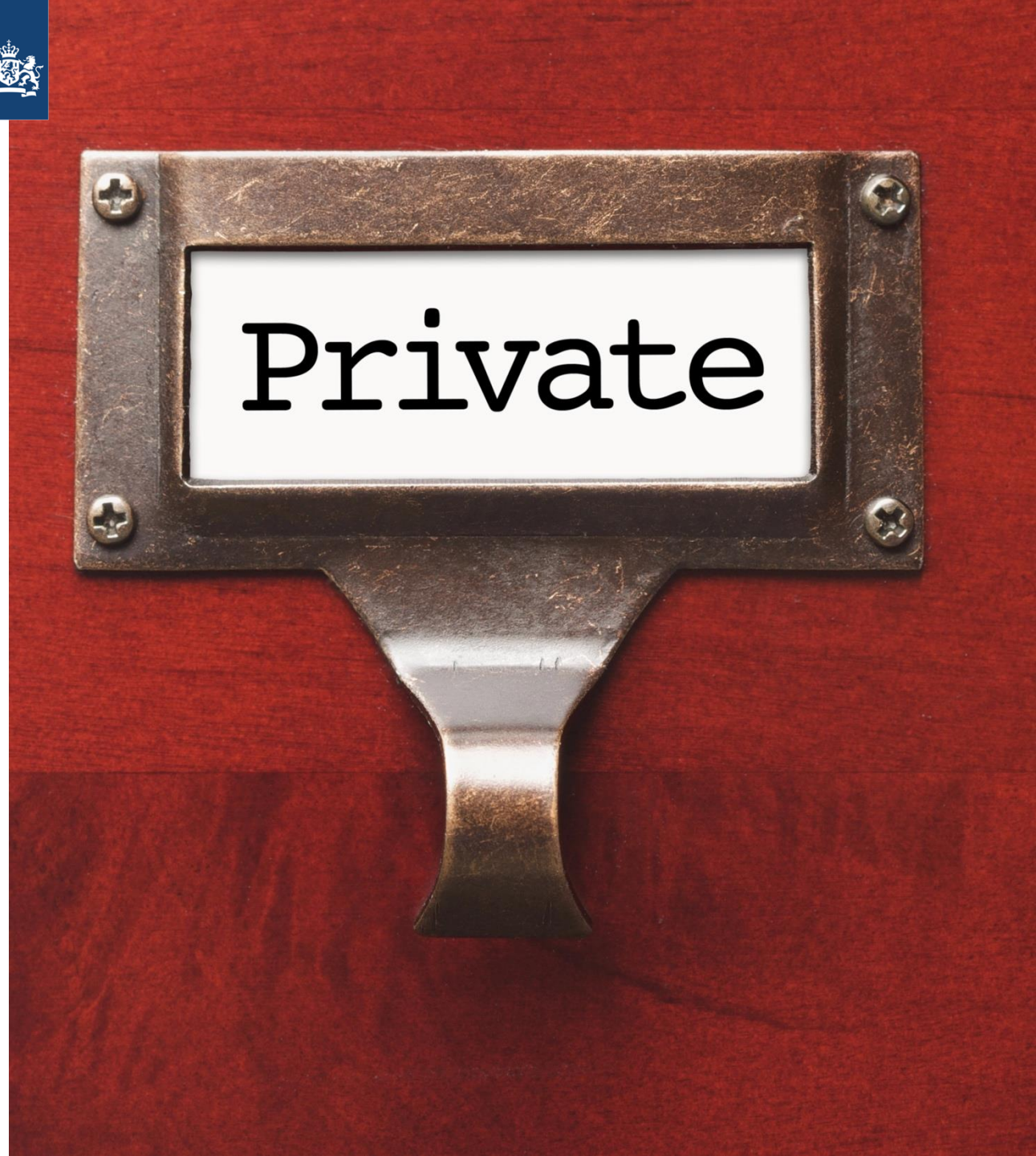
- de interactie tussen gebruikers en het AI-systeem, bijvoorbeeld **automation bias**,
- **reinforcement bias** doordat het systeem opnieuw getraind wordt op basis van feedback uit zijn omgeving,
- wijzigingen in de context waarin het AI-systeem wordt toegepast (**data drift** of **concept drift**).

Welke maatregelen?

- Continu **monitoren**
- Periodiek **(her)evalueren**
- Random **controlegroep**
- **Incidentenregister**

Mogen en willen we wel gevoelige gegevens verzamelen voor het uitvoeren van een bias assessment?

- De meeste organisaties beschikken niet over dit soort gegevens.
- Variabelen als ras, gender en etniciteit zijn **sociale constructen.**
- Risico op **datalekken** en **function creep.**
- **Spanning** tussen het non-discriminatierecht en het gegevensbeschermingsrecht[1]:
  Hoe interpreteren we de AVG en AI Act juist?



[1] Zie het blog van Marvin van Bekkum en Frederik Zuiderveen Borgesius in iBestuur

- Kies voor een **interdisciplinaire aanpak**.
- Zorg voor een **duidelijk doel** en **gemeenschappelijke taal**.
- Laat de keuze voor de definitie van gelijkheid, fairness metrieken en drempelwaarden niet over aan het technische team.
- Volg een **risicogebaseerde aanpak**.
- Blijf **monitoren** op bias na ingebruikname van het AI-systeem.
- Implementeer waar mogelijk **standaarden** en **best practises**.
- **Documenteer** aannames, keuzes, beperkingen, enzovoorts, zodat je het later kunt uitleggen en verantwoording af kunt afleggen.

Links naar meer verdiepende informatie:

- Handreiking non-discriminatie by design en de e-learning Non-discriminatie in algoritmes en data
- Impact Assessment Mensenrechten en Algoritmes van de Utrecht Data School in opdracht van het ministerie van Binnenlandse Zaken
- Algoprudentie database, bias detectie tool en publieke standaard profileringsalgoritme van Algorithm Audit
- Understand, Manage and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists van Tobias Baer
- Fairness and machine learning van Solon Barocas, Moritz Hardt en Arvind Narayanan
- 21 fairness definitions and their politics van Arvind Narayanan
- The political philosophy of AI van Mark Coeckelbergh
- Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI van Sandra Wachter e.a.
- Machine Bias. There's software used across the country to predict future criminals. And it's biased agains blacks. van ProPublica.
- Rechtvaardige algoritmes van het Centraal Planbureau
- Gevoelige gegevens verwerken om discriminatie door AI tegen te gaan? van Marvin van Bekkum en Frederik Zuiderveen Borgesius
- Algoritmekader
- ISO/IEC TS 12791, Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks, binnenkort voor Rijksoverheden waarschijnlijk gratis te downloaden via NEN Connect.
  De komende jaar worden er meer Europese en ISO/IEC standaarden over bias gepubliceerd.

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

Bedankt voor je aandacht!

- Meer weten of eens sparren? Kom zo langs bij onze stand in het **Advanced Technology Lab**.
- Je kunt ook contact opnemen via **Willy.Tadema@Rijksoverheid.nl** of LinkedIn **https://www.linkedin.com/in/willytadema**
- Meer informatie over het Rijks ICT Gilde vind je op onze website: **www.rijksorganisatieodi.nl/rijks-ict-gilde**
- De slides van mijn presentatie zijn te downloaden met behulp van de QR code.