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?
- 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?

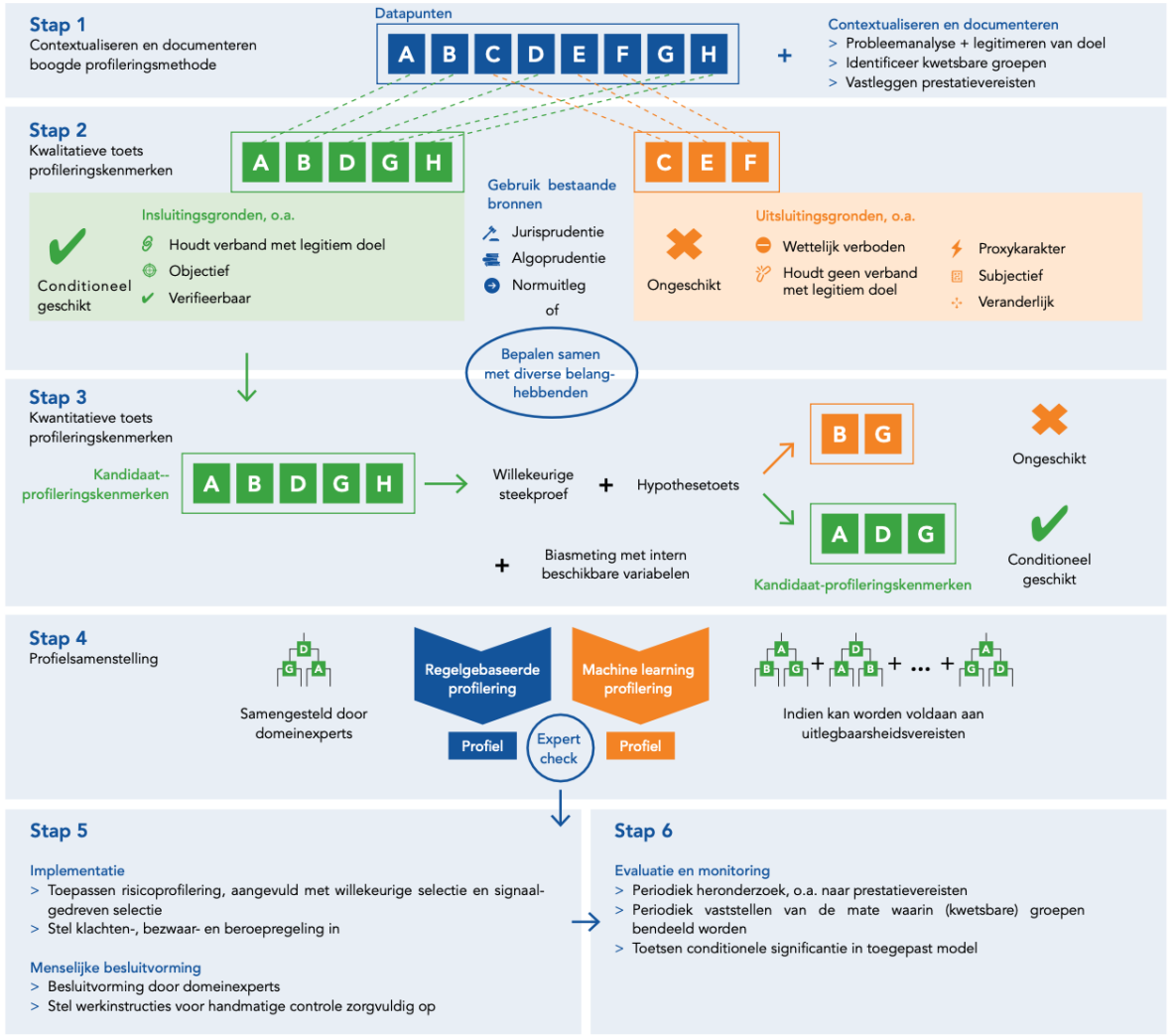
- Een systematische analyse van biases in een AI-systeem
- Kwantitatief en kwalitatief
- In alle fasen van de levenscyclus van het model
- In de training- en testdata, het model en de menselijke interactie met het model
- Het identificeren van de oorzaken van bias (root cause analysis)
- Suggesties voor het verminderen van de bias.
- Doel:
  - Modelprestaties verbeteren
  - Voorkomen dat het AI-systeem ongewenste, negatieve impact heeft op groepen of individuen
  - in de samenleving
  - Voldoen aan wet- en regelgeving

Wat is bias?

- Technisch perspectief:
  - Systematische fout in de modeluitkomsten
  - Negatieve impact op accuracy 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
  - Juridisch risico van non-compliance
- Voor een volledige bias assessment heb je alle drie perspectieven nodig!

Hoe check 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
- Voer de volgende stappen uit:
  - Exploratory fairness analysis[1]
  - Impact assessment[2]
  - Kwalitatieve en kwantitatieve toets op voorspellende variabelen[3]



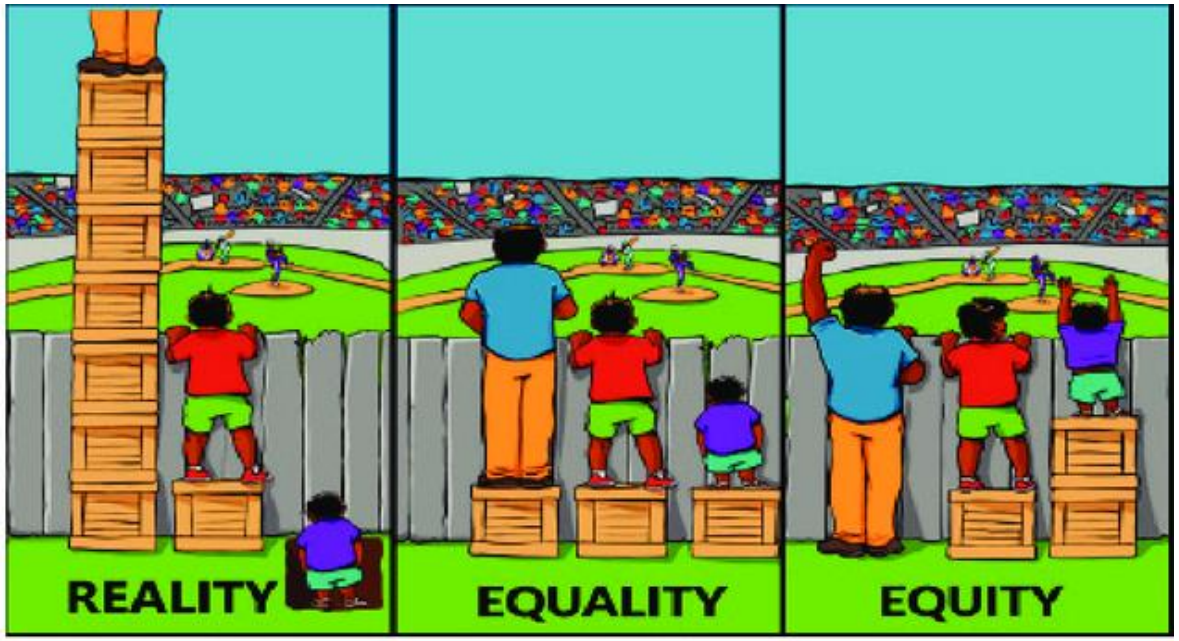
[1] Bijvoorbeeld de *data X-ray* van Tobias Baer

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

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

Hoe kwantificeer en meet je bias in de uitkomsten van het model?

- Afhankelijk van wat je in de specifieke context onder gelijkheid of 'fairness' verstaat, 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]
- Er zijn nog veel meer fairness metrieken voor het meten van bias. [2]
- Hoe je fairness meet en wanneer het te veel is, wordt niet duidelijk gedefinieerd in wet- en regelgeving.[3]



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

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

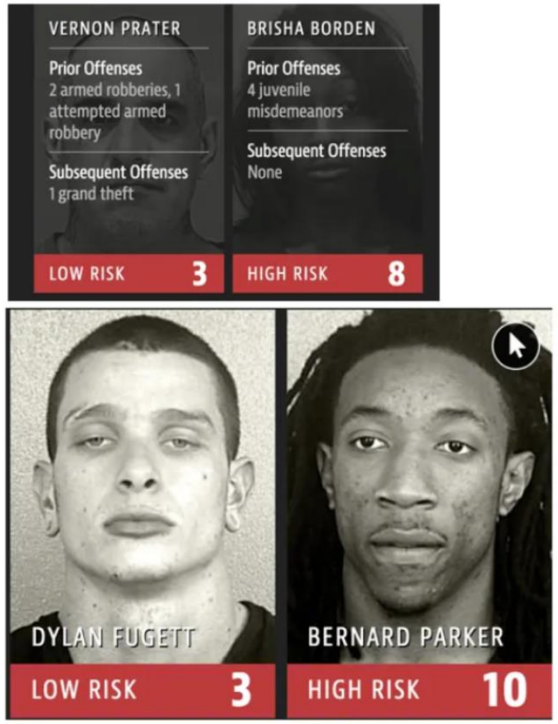
[3] Zie *Why Fairness Cannot Be Automated* van Sandra Wachter e.a.

Ter illustratie de COMPAS-casus uit de Verenigde Staten, een model om het risico op recidive te voorspellen.

ProPublica verzamelde gegevens via FOIA requests en onderzocht of de percentages fout-positieven en fout-negatieven gelijk verdeeld waren.[1] Dat bleek niet zo te zijn.

COMPAS maakte ongelijkmatig fouten tussen verschillende demografische groepen. Zwarte verdachten werden vaker onterecht als 'hoog risico' bestempeld, en witte verdachten vaker onterecht als laag risico.

Northpointe – de ontwikkelaar van COMPAS – hanteerde een andere definitie van fairness en gebruikte andere fairness metrics om te beargumenteren dat het model *niet* discrimineerde.



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

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

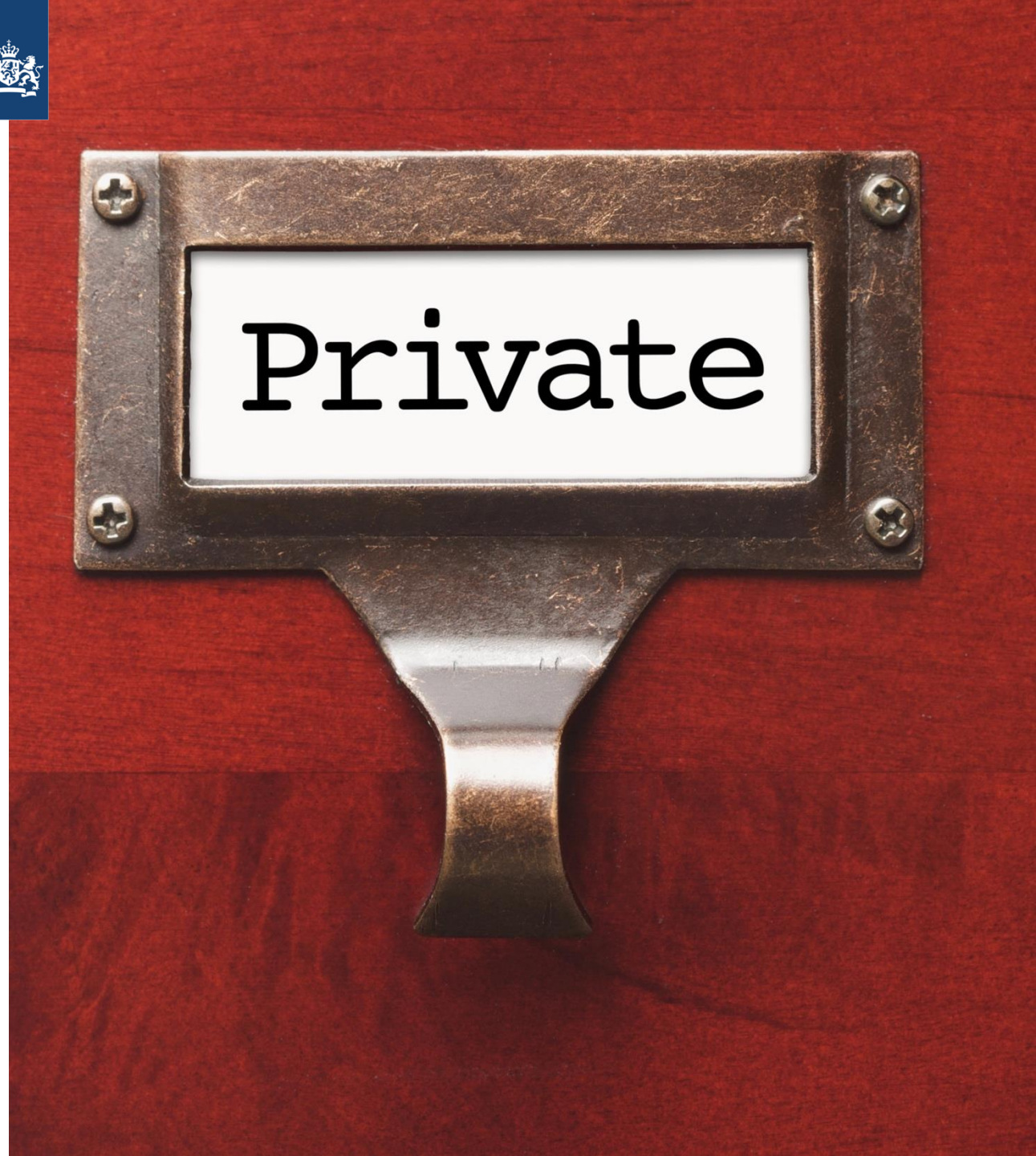
- de wijze waarop mensen met het systeem interacteren, bijvoorbeeld *automation bias*,
- *reinforcement bias* doordat het systeem opnieuw getraind wordt op basis van feedback uit zijn omgeving,
- wijzigingen in de context waarin het AI-systeem wordt toegepast (*data drift* of *concept drift*).

Voorbeelden van maatregelen:

- Continu monitoren op bias en periodiek (her)evalueren
- Controlegroep
- Incidentenregister

Het verzamelen en gebruiken van gevoelige gegevens voor het uitvoeren van een bias assessment is een technisch, ethisch en juridisch vraagstuk:

- 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 heldere taal
- Voer een exploratory fairness analyse uit
- :Laat de keuze voor de definitie van gelijkheid en het kwantificeren van bias niet over aan het technische team
- Blijf monitoren op bias na ingebruikname van het AI-systeem
- Volg een risicogebaseerde aanpak
- Implementeer waar mogelijk standaarden
- Leg alle keuzes, overwegingen, aannames, tekortkomingen van de data , gebruikte fairness metrieken, enzovoorts vast, zodat je het later kunt uitleggen en er verantwoording over kunt afleggen

Links naar meer verdiepende informatie:

- Algoritmekader
- Handreiking en leermodule non-discriminatie by design
- 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.
- 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.
- Gevoelige gegevens verwerken om discriminatie door AI tegen te gaan? van Marvin van Bekkum en Frederik Zuiderveen Borgesius

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.