

WAT MET FAIRNESS & AI?

Veel digitale producten, in allerlei delen van de samenleving, gebruiken **AI en machine learning algoritmes**.

Een AI-algoritme bestaat uit een reeks instructies die, als ze correct uitgevoerd worden door een computer/software, een probleem kunnen oplossen of een taak kunnen voltooien. Toch zijn deze systemen vaak een **black box** en is het niet duidelijk voor mensen hoe de AI-systemen precies tot een beslissing komen. Er zijn verschillende voorbeelden waarvan we als maatschappij vinden dat AI-systemen bestaande **vooroordelen (biases)** herbevestigen. Denk daarbij aan gezichtsherkenningsoftware die bepaalde huidskleuren beter herkent dan anderen. Om dit te verhelpen kan een algoritme gebruikt worden dat rekening houdt met deze vooroordelen door bepaalde **'fairness'**

principes' toe te passen. Er bestaan echter vele verschillende principes en iedereen heeft een eigen mening over welke nu de beste is.

In deze brainfood stellen we een aantal van deze fairness principes voor aan de hand van het voorbeeld van een geautomatiseerde sollicitatieprocedure. Welke van deze principes vind jij het meest eerlijk en rechtvaardig?

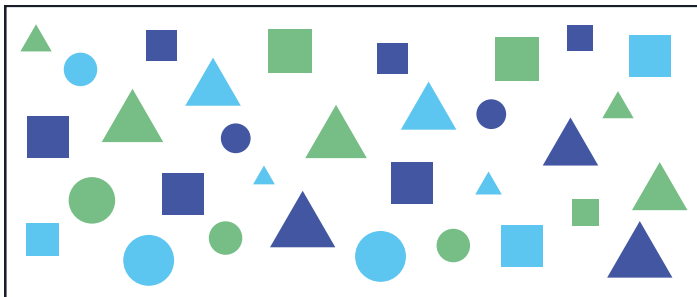
Deze brainfood kwam tot stand in samenwerking met Carmen Mazijn en An Jacobs van het IRP COMPASS project van de Vrije Universiteit Brussel.

Kenniscentrum Data & Maatschappij (2022). Wat met fairness & AI? Brussel.

Dit document is beschikbaar onder een CC BY 4.0 licentie.



ALLE SOLLICITANTEN VOOR EEN JOB

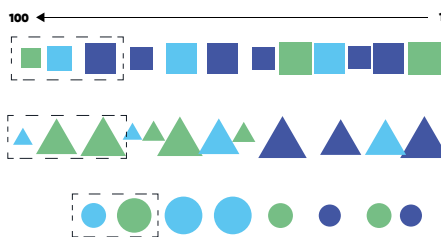


STEL JE VOOR...

Hierboven zie je alle mensen die hun CV hebben opgestuurd om te solliciteren voor een specifieke job. Elk van deze fictieve sollicitanten is een beetje anders, in vorm, grootte of kleur, maar in theorie zou elk gekwalificeerd kunnen zijn voor de job. Wie wordt er uitgenodigd voor een sollicitatiegesprek en wie niet? Het bedrijf gebruikt een AI-systeem om deze beslissing te maken.

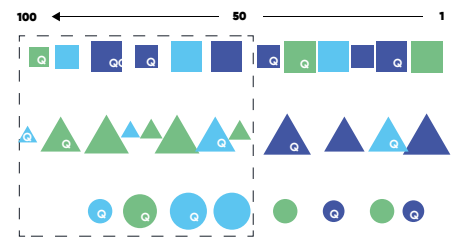
Als uitgangspunt berekent het algoritme **op basis van historische data** voor elke sollicitant een score tussen de 1 en 100. De mensen met **de hoogste score** worden uitgenodigd om op gesprek te komen. Maar dit is niet per se een eerlijke manier van selectie. Het algoritme kan bijvoorbeeld een bepaalde groep benadelen, zoals de donkerblauwe cirkels in dit voorbeeld. Om hierop in te spelen, kunnen we verschillende fairness principes toepassen.

PRINCIPE 1: STATISTICAL PARITY



Door **'statistical parity'** (statistische gelijkheid) toe te passen op je algoritme, worden de kandidaten eerst opgedeeld op basis van bepaalde (beschermde) karakteristieken, in dit geval de vorm, en daarna gesorteerd. Per groep wordt dan eenzelfde aandeel of percentage van degenen met de hoogste scores geselecteerd voor een gesprek. Op die manier wordt **elke groep evenredig meegenomen** in het selectieproces.

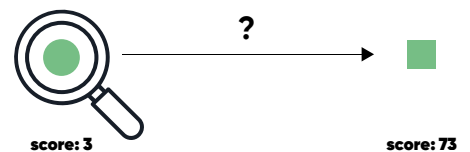
PRINCIPE 2: EQUAL OPPORTUNITY



In plaats van een vast aantal mensen uit te nodigen voor een gesprek, kan een bedrijf er ook voor kiezen om iedereen die **minimaal een bepaalde score** heeft behaald uit te nodigen. We noemen dat een **drempelwaarde**. In dit voorbeeld ligt de drempelwaarde op score 50.

De Q duidt aan dat deze persoon ook echt gekwalificeerd is voor de job. Deze waarde kan niet op voorhand met zekerheid geweten zijn en kan enkel na aanwerving getoetst worden. Met het fairness principe **'equal opportunity'** (gelijke kansen) kan je nagaan of de **foutenmarge** voor personen die effectief gekwalificeerd waren, maar niet boven de drempelwaarde uitkomen, **gelijk** is in elke groep.

PRINCIPE 3: CAUSAL DISCRIMINATION



De vorige principes keken elk naar de verschillende groepen. We kunnen ook vanuit het individu kijken, dit heet dan **individuele fairness**. Met **'causal discrimination'** (causale discriminatie) wordt bijvoorbeeld beoordeeld of iemand met, op één na, dezelfde karakteristieken een beduidend andere score zou ontvangen. Als dit het geval is bij een karakteristiek die wettelijk beschermd is, zoals gender of etniciteit, dan is het algoritme niet fair en correct.

CONCLUSIE

Er zijn nog veel andere fairness principes die ontwikkelaars kunnen toepassen op hun AI-systemen. In elke context kan een ander principe het meest geschikt zijn. Dit vormt dan ook **een uitdaging voor ontwikkelaars** om te kiezen welk principe ze zullen gebruiken. Wil jij ook graag meepraten in die discussie en **zelf ervaren** hoe het voelt als een AI-systeem jou beoordeelt? In de workshop ['Geef de computer jou de job die je verdient?'](#) kan je de fairness principes zelf aan den lijve ondervinden.

DE BEREKENING VAN HET ALGORITME ZONDER GEBRUIK VAN FAIRNESS PRINCIPES

