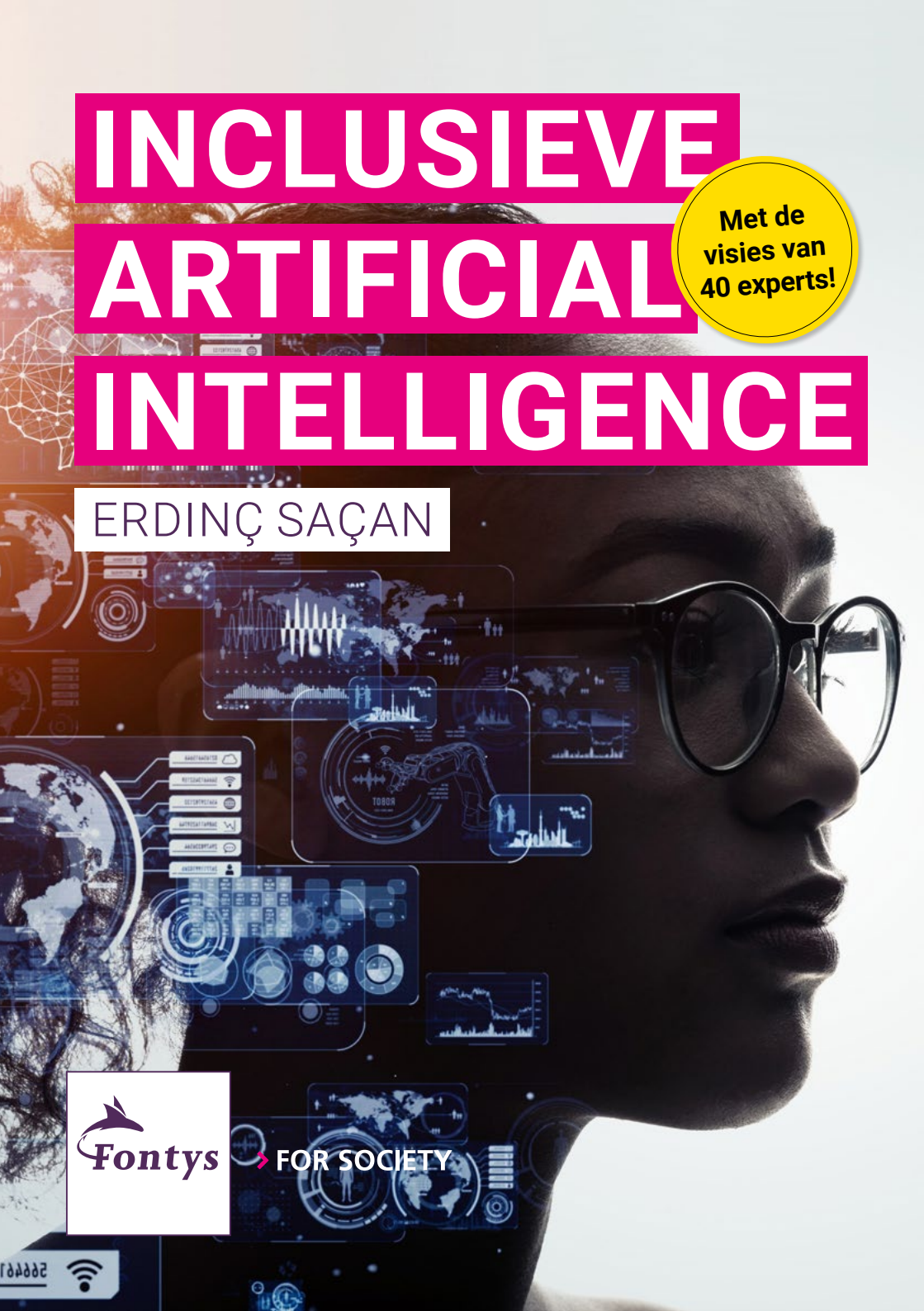
INCLUSIEVE ARTIFICIAL INTELLIGENCE

Met de visies van 40 experts!

ERDINÇ SAÇAN



Fontys
FOR SOCIETY

INCLUSIEVE ARTIFICIAL INTELLIGENCE

ERDINÇ SAÇAN

Inhoud

Voorwoord

'Onze vooroordelen zijn zo diep geworteld dat we ze nooit als vooroordelen beschouwen, maar ze gewoon gezond verstand noemen.'

George Bernard Shaw

Iers-Engels schrijver, criticus en Nobelprijswinnaar literatuur (1925) 1856-1950

Een systeem dat je adviseert over je kleding, het smakelijkste kruid bij een zalmmoot, of over boeken en films die op jouw lijf zijn geschreven. Kleine suggesties die het leven makkelijker maken, gebaseerd op informatie die is verzameld door je surf- en/of koopgedrag. Altijd handig, toch?

Laatst kocht ik een PlayStation voor mijn elfjarige zoon. Samen volgden we de installatie- en aanmeldprocedure. Mijn zoon voerde tweemaal een verkeerd wachtwoord in. Er verscheen een melding dat er onrechtmatig gebruik van de PlayStation werd gemaakt en dat het account zodoende werd opgeschort. Al googelend naar informatie ontdekte ik dat PlayStation met behulp van AI heel goed misbruik kan achterhalen. Het kan zelfs je hele device op slot gooien. Zonder dat je daar ook maar één weerwoord op kunt geven. Geen mens van vlees en bloed aan de andere kant van de lijn om aan uit te leggen dat het geen fraude, maar een foutje was. Uiteindelijk lukte het om mijn zoon aan te melden en hij speelt nu volop FIFA.

Irritant, maar relatief onschuldig voorbeeld, zeg je? Doe ik er nog eentje. Mijn moeder uit 1946 hoort bij de eerste generatie gastarbeiders. Ze kreeg een brief van de Belastingdienst met de mededeling dat ze in het datasysteem Fraude Signalering Voorziening (FSV) stond. Als huisvrouw heeft mijn moeder amper iets te maken gehad met de Belastingdienst. FSV voldeed echter niet aan de Algemene Verordening Gegevensbescherming (AVG). Te veel medewerkers

hadden toegang tot het systeem en gegevens werden te lang bewaard. Sommige van die gegevens zijn onterecht opgenomen en weer andere gegevens werden verkeerd gebruikt.

Op 27 februari 2020 werd FSV dan ook uitgezet. Goed dat de belastingdienst die stap heeft genomen, maar de schrik zat er bij mijn moeder en de ongeveer 240.000 andere mensen op die zwarte lijst goed in. Vooral mensen met een migrantenachtergrond kregen zo'n beschuldigende brief. Is dit een voorbeeld van bias in het door de Belastingdienst gebruikte algoritme? De toeslagenaffaire leidde tot een parlementaire ondervragingscommissie en bracht uiteindelijk Rutte III ten val. En wekte mijn onderzoekende interesse.

Hoe eng is het om een robot belangrijkere beslissingen te laten nemen met kunstmatige intelligentie? Waar ligt de grens? Beslissingen over je vervolgopleiding, sollicitatie, salaris, subsidie of je eerste koophuis. Of je wel of niet bevoegd bent om in een vliegtuig te stappen. En wat als de rechtsmacht, politie of defensie volledig vertrouwt op kunstmatige intelligentie? En ook: hoe pak ik zo'n uitgebreid, breed onderwerp aan? Ik heb besloten om experts uit het veld hun mening te vragen.

We kunnen niet zonder AI en er zitten zeker veel voordelen aan. De vraag is echter of we het voor alles moeten gebruiken. En als we het gebruiken hoe voorkom je dat er misbruik ontstaat, dat er geen fouten in sluipen en dat je toch altijd in uiterste gevallen terecht kunt bij een mens. Vragen, veel vragen en hopelijk krijgen we gezamenlijk antwoorden.

Bijna dagelijks verschijnt er een artikel, video, documentaire, paper en maandelijks komt een boek uit over Artificial Intelligence (AI). Er is ook steeds meer aandacht voor de ethische kant ervan. Dit boek is een poging om aan de hand van interviews met experts uit het veld te kijken hoe AI op de juiste, onbevooroordeelde (neutrale) manier kan worden ingezet. Daarom legde ik de experts twee vragen voor:

- **Kunnen we algoritmen inzetten voor het algemene belang, om zo discriminatie en ongelijkheid te bestrijden?**

- **Algoritmes nemen steeds meer beslissingen voor ons. Hoe zorgen we ervoor dat dit op een inclusieve manier gebeurt?**

Voordat we kijken naar de antwoorden, geef ik een overzicht van enkele recente ontwikkelingen die inclusieve AI bedreigen of juist bevorderen.

Recente ontwikkelingen in AI

Kansen en bedreigingen van AI

De afgelopen jaren steeg de belangstelling voor kunstmatige intelligentie explosief. Vooralsnog is er geen aanwijzing dat deze trend zal afzwakken. Aan de lange lijst van optimistische statistieken voegt Softwarebedrijf Citrix nog toe dat AI in hun Work 2035-studie de grootste aanjager is van organisatorische groei. Ook waarschuwde een door BlueDot ontwikkeld AI-algoritme de wereld voor het eerst voor Covid-19. Dat deed het maar liefst negen dagen voordat de Wereldgezondheidsorganisatie alarm sloeg. Bovendien zorgde AI ervoor dat wetenschappers de vorm van eiwitten binnen enkele minuten konden voorspellen en al drie maanden na de eerste uitbraak over werkbare vaccinaties beschikten. Ook toepassingen zoals track and trace, het prioriteren van vaccins en beademingsapparatuur, clusteranalyse en het indammen van de online verspreiding van verkeerde informatie werden dankzij de vooruitgang op het gebied van AI mogelijk.

Meer dan ooit is AI inderdaad overal en verandert het ons leven positief en fundamenteel. AI heeft echter ook het potentieel om een grote bedreiging voor de wereld te zijn. Er moet een parallel spoor naar de inzet van AI komen dat zich uitsluitend richt op het verantwoord gebruik van gegevens en kunstmatige intelligentie. Als dit niet gebeurt, kan AI zich voegen bij Covid-19 en klimaatverandering als de grootste uitdagingen voor onze wereld in 2021 en daarna. (Lang, 2021)

AI moet altijd op de mens gericht zijn. Als instrument moet AI mensen en samenleving helpen hogere doelen te bereiken. Ook moet het onder menselijk toezicht staan om oneerlijkheid en vooringenomenheid te voorkomen. Omdat AI wordt getraind op bestaande data en omgevingen, en omdat sommige van deze data inherente vooroordelen kunnen blootleggen of weerspiegelen, zijn er gevallen geweest waarin AI deze ongewenste eigenschappen heeft aangeleerd.

Zoals toen Microsoft Tay (@TayandYou) ontwikkelde. Deze Twitter chatbot AI startte als een experiment in het begrijpen van conversaties, maar in minder dan 24 uur begon het racistische berichten te genereren. Microsoft schakelde Tay aan het eind van de dag uit. Hoewel dit incident anekdotisch is, toont het aan hoe impliciete vooroordelen in gegevens zonder een verantwoord AI-kader waarschijnlijk onverwachte en ongewenste resultaten zullen opleveren. (Villanustre, 2021)

Aanpakken / neutraliseren van algoritmische bias

Technologie is in feite nooit neutraal. In elke fase - van ontwerp tot ontwikkeling, van test tot gebruik en onderhoud in de toepassingscontext - bepalen menselijke keuzes en overtuigingen het verdere verloop. In de dimensie van eerlijkheid kan de vooringenomenheid van AI-systemen menselijke vooroordelen versterken en discriminatie veroorzaken.

'Man is to Computer Programmer as Woman is to Homemaker? Debiasing van Word Embeddings'

Bolukbasi et al., 2016

Carissa Véliz,
Associate Professor aan University of Oxford zegt:

"Algoritmen zijn slechts een hulpmiddel. Helaas worden algoritmen vaker wel dan niet gebruikt om kosten te besparen en productiviteit te verhogen, zonder dat voldoende aandacht wordt besteed aan de gevolgen voor het individu en de samenleving.

Er is niet één oplossing. We moeten ervoor zorgen dat de groep mensen die algoritmen ontwerpt divers is. Als blanke, rijke mannen de meeste algoritmen ontwerpen, zou het ons niet moeten verbazen als deze tools uiteindelijk slechte gevolgen hebben voor vrouwen en minderheden. We moeten algoritmen ook voortdurend controleren om ervoor te zorgen dat ze de gelijkheid van kansen niet ondermijnen."

Is algoritmische bias eenmaal vastgesteld, hoe kunnen dan oorzaken worden geïdentificeerd en gevolgen worden beperkt? Het meest voorkomende probleem komt naar voren in de gegevens waarmee deze modellen worden getraind. Vaak blijken ze niet voldoende representatief voor de verschillende minderheden.
Een eerste eenvoudige oplossing om deze verstoringen te verminderen, is de gevoelige attributen volledig te verwijderen zodat ze niet kunnen worden gebruikt voor classificatie of door een andere fase van gegevensverzameling uit te voeren en zo een meer gebalanceerde verzameling op te bouwen.

Ook speelt de vraag naar verantwoordelijkheid: aan wie en aan welke factoren zijn deze automatische keuzes en de daaruit voortvloeiende maatschappelijke effecten toe te schrijven? Aan het algoritme, de programmeur, de Data Scientist of aan het bedrijf dat het model gebruikt?

Een van de grootste problemen bij het aanpakken van eerlijkheid van AI-modellen ligt in het ontbreken van een eenduidige definitie van deze eigenschap. Daarnaast speelt ook het ontbreken van door de wetenschappelijke gemeenschap geaccepteerde gestandaardiseerde technieken een grote rol.

Ook moet rekening worden gehouden met het feit dat het systeem dan wel eerlijk kan zijn met betrekking tot een aantal technische parameters. Maar als het vervolgens wordt gebruikt voor schadelijke doeleinden of effecten, wordt de technologie oneerlijk en gevaarlijk. Denk bijvoorbeeld aan het gebruik van gezichtsherkenningstechnologieën voor bewaking en tracking.

Wanneer datawetenschappers en advocaten wordt gevraagd ervoor te zorgen dat hun AI eerlijk is, hangen daar in de praktijk ook vervolgvragen aan. Wat betekent eerlijkheid in de context van elke specifieke user case en hoe moet dit worden gemeten? Dit kan een ongelooflijk complex proces zijn, zoals een groeiend aantal onderzoekers in de machine learning-gemeenschap de afgelopen jaren hebben opgemerkt.
https://arxiv.org/pdf/1912.06883.pdf

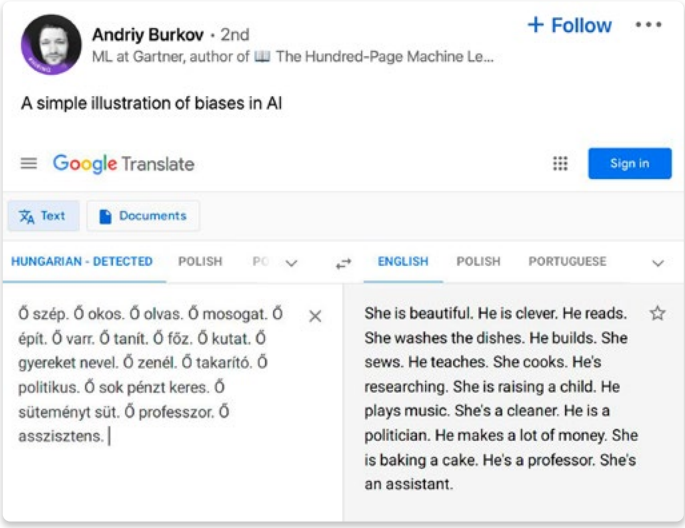
Bedrijven kunnen ook putten uit openbare richtlijnen van experts zoals Nicholas Schmidt en Bryce Stephens van BLDS. https://arxiv.org/abs/1911.05755

Reijer Passchier,
Assistant Professor in Constitutional Law at the Open University and Leiden University, stuurde mij het volgende:

"Ik vind in elk geval dat wij naast aandacht voor de gevolgen van het gebruik van algoritmes voor mensenrechten zoals privacy, veel meer aandacht zouden moet hebben voor de vraag wat algoritmegebruik betekent voor machtsverhoudingen en wat voor gevolgen eventuele verschuivende machtsverhoudingen hebben voor de effectiviteit van checks and balances."

Grip op de risico's van AI-bias

Van vooroordelen ofwel bias door AI bestaan veel voorbeelden. Andriy Burkov, de leider van een machine learning team bij Gartner, plaatste op LinkedIn onderstaande voorbeelden van Google Translate. Links staat de Hongaarse tekst en rechts de vertaling. Zij maakt schoon, zij kookt … maar hij verdient geld en hij is een professor.



Bijna 90% van de dataprofessionals zegt dat vooroordelen in data die gebruikt worden voor AI/ML-systemen 'discriminerende resultaten' opleveren. Dit leidt weer tot nalevingsrisico's, zoals blijkt uit het laatste State of Data Culture Report van Alation.

Analyse van op AI gebaseerde sollicitatiegesprekken laat de volgende opmerkelijke 'fout' zien. De software belooft persoonlijkheidskenmerken te kunnen detecteren en 'sneller, maar ook objectiever' te zijn. In de praktijk blijkt echter dat een handig geplaatste boekenplank op de achtergrond de resultaten in positieve zin verandert. https://web.br.de/interaktiv/ki-bewerbung/en/

"Toen de technologie verschoof van stoomkracht naar elektriciteit, waren de eerste pogingen om elektriciteit naar de industrie te brengen niet erg succesvol omdat men gewoon stoommachines probeerde na te maken. Ik denk dat er nu iets soortgelijks aan de hand is met AI. We moeten uitzoeken hoe we die in veel verschillende gebieden kunnen integreren: niet alleen in de gezondheidszorg, maar ook in het onderwijs, bij het ontwerpen van materialen, bij stadsplanning, enzovoort. Natuurlijk is er meer te doen aan de technologische kant, waaronder het maken van betere algoritmen, maar we brengen deze technologie in sterk gereguleerde omgevingen en we hebben nog niet echt gekeken hoe we dat moeten doen.

Op dit moment floreert AI op plaatsen waar de faalkosten zeer laag zijn. Als Google een verkeerde vertaling voor je vindt of je een verkeerde link geeft, is dat prima; je kunt gewoon naar de volgende gaan. Maar dat gaat niet werken voor een dokter. Als je patiënten de verkeerde behandeling geeft of een diagnose mist, heeft dat echt ernstige gevolgen. Veel algoritmen kunnen eigenlijk dingen beter dan mensen. Maar we vertrouwen altijd meer op onze eigen intuïtie, ons eigen verstand, dan op iets wat we niet begrijpen. We moeten artsen redenen geven om AI te vertrouwen."

Het citaat hierboven komt uit een interview met Regina Barzilay, hoogleraar aan het Computer Science and Artificial Intelligence Laboratory (CSAIL) van het MIT, de eerste winnaar van de Squirrel AI Award voor kunstmatige intelligentie ten bate van de mensheid. (Heaven, 2020)

Een goed gebruik van data biedt ons ook een kans om eerlijkheid te bevorderen. Het kan dienen als een krachtig instrument waarmee we kunnen zien waar vooroordelen zich voordoen en kunnen meten of onze inspanningen om die te bestrijden effect sorteren. Als een organisatie over harde gegevens beschikt over de verschillen in de manier waarop zij mensen behandelt, kan zij inzicht

verwerven in de oorzaken van die verschillen, en die proberen aan te pakken. (Durkee, 2021)

Een paar steden, waaronder Amsterdam, Helsinki en New York, experimenteren al met benaderingen om de transparantie te vergroten. In Nederland ontstaan meer initiatieven. Zo stelde het ministerie van Justitie en Veiligheid richtlijnen op voor het toepassen van algoritmen door overheden. Deze richtlijnen geven organisaties handvatten voor het verantwoord ontwikkelen en gebruiken van algoritmen.

Noord-Holland is van plan om alle gebruikte algoritmes openbaar te maken, valt te lezen in de datastrategie 2021-2023. De provincie maakt gebruik van kunstmatige intelligentie en geautomatiseerde besluitvorming en gaat dat ontsluiten in een algoritmeregister.

Ook een interessante actie komt uit Schotland. 'Betrouwbaar, ethisch en inclusief' is de titel van De AI-strategie van Schotland waaraan iedereen kan deelnemen.

Op haar beurt diende de EU een beleidsvoorstel in voor het reguleren van kunstmatige intelligentie. Dit beleidsvoorstel brengt gebruiksgevallen voor kunstmatige intelligentie onder in de volgende vier risicocategorieën:

Onaanvaardbaar risico
AI-systemen die als een duidelijke bedreiging voor de veiligheid, de bestaansmiddelen en de rechten van mensen worden beschouwd, zullen worden verboden. Hieronder vallen AI-systemen of -toepassingen die menselijk gedrag manipuleren om de vrije wil van gebruikers te omzeilen (bijvoorbeeld via speelgoed dat met behulp van spraakondersteuning gevaarlijk gedrag van minderjarigen aanmoedigt) en systemen die 'social scoring' door regeringen mogelijk maken.

Hoog risico
Hieronder vallen AI-systemen/technologieën die worden gebruikt in:

- Kritieke infrastructuren zoals vervoer die leven en gezondheid van burgers in gevaar kunnen brengen;
- Onderwijs- of beroepsopleidingen die bepalend kunnen zijn voor iemands toegang tot onderwijs en beroepsloopbaan. Denk aan het beoordelen van examens;
- Veiligheidscomponenten van producten zoals AI-toepassingen in robot-geassisteerde chirurgie.

De complete lijst is een stuk uitgebreider. Volg voor het totale plaatje vooral de eerdere link.

Daarnaast zijn er nog de categorieën **Beperkt risico** en **Minimaal risico**.

Internet Engineering Task Force pakt een netelig probleem aan: het schrappen van computertechnische termen die een racistisch verleden oproepen, zoals 'master' en 'slave' en 'whitelist' en 'blacklist'.

> *Mallory Knodel, hoofd technologie bij de beleidsorganisatie Center for Democracy and Technology, schreef samen met co-auteur Niels ten Oever, postdoctoraal onderzoeker aan de Universiteit van Amsterdam, een voorstel voor neutraler taalgebruik van de taskforce. 'Blocklist' zou dan verklaren wat een blacklist doet, en 'primary' zou 'master' vervangen.*

Maar terwijl de industrie afziet van verwerpelijke termen, is er geen consensus over welke nieuwe woorden moeten worden gebruikt. De programmeergemeenschap die MySQL onderhoudt, koos 'source' en 'replica' als vervangers voor 'master' en 'slave'. GitHub, de code repository eigendom van Microsoft, koos op zijn beurt voor 'main' als alternatief voor 'master.' (Conger, 2021)

Onderzoekers voerden een reeks experimenten uit waarin de invloed van AI-algoritmen in verschillende contexten werd getest. Ze rekruteerden deelnemers om te interageren met algoritmes die foto's van fictieve politieke kandidaten of online dating kandidaten presenteerden. Vervolgens vroegen ze de deelnemers om aan te geven op wie ze zouden stemmen of wie ze een bericht zouden sturen. De algoritmes promootten sommige kandidaten boven anderen. Dit gebeurde zowel expliciet (bijvoorbeeld "90% compatibiliteit") als heimelijk, door hun foto's vaker te tonen dan die van anderen. (Ujué Agudo, 2021)

In het algemeen lieten de experimenten zien dat de algoritmen een significante invloed hadden op de beslissing om te stemmen of een bericht te sturen. Expliciete manipulatie had een significante invloed op politieke beslissingen, heimelijke manipulatie bleek hier niet effectief. Het tegenovergestelde effect werd juist gezien voor beslissingen op datingvlak.

De onderzoekers voegen hier het volgende toe: "Als een fictief en simplistisch algoritme zoals het onze een dergelijk niveau van overtuigingskracht kan bereiken zonder werkelijk op maat gemaakte profielen van de deelnemers op te stellen (en in alle gevallen dezelfde foto's te gebruiken), dan moet een meer gesofisticeerd algoritme, zoals die waarmee mensen in hun dagelijks leven interageren, zeker in staat zijn een veel sterkere invloed uit te oefenen."

Gezichtsherkenningstechnologie

Schrijver van het boek Atlas of AI, Kate Crawford:

> *"AI is noch kunstmatig, noch intelligent. Het is het tegendeel van kunstmatig. Het komt voort uit de meest materiële delen van de aardkorst en uit de arbeid van menselijke lichamen, en uit alle artefacten die we elke dag produceren en zeggen en fotograferen. Het is ook niet intelligent. Ik denk dat er op dit gebied een grote erfzonde is begaan, waarbij mensen ervan uitgingen dat computers op de een of andere manier op menselijke hersenen lijken en als we ze maar trainen als kinderen, ze langzaam zullen uitgroeien tot bovennatuurlijke wezens.*
>
> *Dat is iets wat ik echt problematisch vind - dat we dit idee van intelligentie hebben gekocht, terwijl we in feite gewoon kijken naar vormen van statistische analyse op schaal die evenveel problemen hebben als de gegevens die worden gegeven."*

Emotion recognition technology (ERT) (Emotieherkenningstechnologie) heeft tot doel AI te gebruiken om emoties te detecteren aan de hand van gezichtsuitdrukkingen. De wetenschap achter emotieherkenningssystemen is echter controversieel: er zijn vooroordelen ingebouwd in de systemen. (Alexa Hagerty, 2021)

Veel bedrijven gebruiken ERT om de reacties van klanten op hun producten te testen, van ontbijtgranen tot videospelletjes. Maar ook in situaties waarin veel meer op het spel staat, zoals bij personeelswerving, luchthavenbeveiliging of grenscontroles, kan ERT worden gebruikt om gezichten te markeren als bedrieglijk of angstig.

De bekroonde film Coded Bias, momenteel op Netflix, documenteert de ontdekking dat veel gezichtsherkenningstechnologieën gezichten met een donkere

huidskleur niet accuraat detecteren. En het onderzoeksteam dat ImageNet beheert, een van de grootste en belangrijkste datasets die worden gebruikt om gezichtsherkenning te trainen, werd onlangs gedwongen om 1,5 miljoen afbeeldingen te vervagen als reactie op bezorgdheid over privacy.

Onderzoekers aan de Universiteit van Cambridge en de UCL bouwden een website met de naam Emojify. Insteek was om mensen te helpen begrijpen hoe computers kunnen worden gebruikt om gezichtsuitdrukkingen te scannen voor het detecteren van emoties en welke risico's er aan deze AI-emotieherkenning kleven. De onderzoekers zeggen dat ze hopen gesprekken op gang te brengen over de technologie en de sociale gevolgen ervan. (Russell, 2021)

Onthullingen over algoritmische vooringenomenheid en discriminerende datasets in de technologie voor gezichtsherkenning leidden ertoe dat grote technologiebedrijven, waaronder Microsoft, Amazon en IBM, de verkoop hebben stopgezet. In de EU riep een coalitie van meer dan 40 maatschappelijke organisaties op tot een volledig verbod op gezichtsherkenningstechnologie.

Bias en de kansen en risico's op (sociale) inclusie of uitsluiting door de werking van AI-algoritmen

Pedro Domingos is een Portugees computerwetenschapper en professor. Hij wordt gezien als expert in kunstmatige intelligentie. Op 19 april 2021 tweette hij het volgende:

Dit staat haaks op wat KPMG-medewerker Ylja Remmits en haar kantoorgenoot hoogleraar Sander Klous beweren in Trouw. Zij schrijven:

> *"Discriminerende factoren als etniciteit uit algoritmes halen, dat maakt vooroordelen hoogstens onzichtbaarder. Het helpt niet om afkomst uit het algoritme weg te laten. Een groot aantal andere gegevens hangt namelijk samen met iemands afkomst. Denk aan adres, opleiding of sociaaleconomische status. Deze gegevens noemen we 'proxies'. Omdat bij de ontwikkeling van het algoritme nog steeds gebruik wordt gemaakt van de door de bewoners ervaren leefbaarheid, zullen vooroordelen via proxies net zo hard in het algoritme terugkomen. Het is nu alleen een stuk ingewikkelder om ze aan te tonen, omdat we de expliciete informatie over afkomst missen. De oplossing is niet het weglaten van deze indicatoren, maar juist het inzetten op de juiste manier. We kunnen deze informatie gebruiken om relaties zoals die in het bovenstaande voorbeeld te berekenen en dus inzichtelijk te maken."*

Neuro-informaticus Sennay Ghebreab (Universiteit van Amsterdam) over de risico's, maar ook de kansen die AI biedt voor sociale inclusie: "Discriminerende AI zegt veel meer over wat er niet eerlijk is in de samenleving dan over de technologie zelf."

"AI zal nooit helemaal eerlijk zijn. Het punt is niet volledige eerlijkheid, maar de noodzaak om maatstaven en drempels voor eerlijkheid vast te stellen die het vertrouwen in AI-systemen garanderen", aldus Virginia Dignum

Gerd Leonhard

Keynote Spreker (Virtueel en RL), Auteur, Futurist & Humanist en CEO van The Futures Agency, beantwoordt mijn vraag of we algoritmen kunnen gebruiken voor het algemeen welzijn, om discriminatie en ongelijkheid te bestrijden als volgt:

"Niet echt, technologie lost geen sociale, culturele of politieke problemen op. Het maakt ze meestal erger! We kunnen technologie gebruiken als een TOOL zodra we hebben besloten de menselijke aandacht te richten op de juiste beleidsvorming. We moeten de mens te allen tijde bij de les houden (HITL), dat betekent dat mensen toezicht moeten houden op AI en hun werk moeten controleren, ook al duurt het langer of is het minder efficiënt"

Ondanks de negatieve gevolgen zou de huidige concentratie op AI-bias in zekere zin een goede zaak kunnen zijn. Vooral omdat het grote en kleine bedrijven - en andere belanghebbenden - in de huidige samenleving dwingt meer aandacht te besteden aan vooroordelen die afbreuk kunnen doen aan hun bedrijfsresultaat. Uit een recent rapport van McKinsey blijkt dat Hollywood tien miljard dollar aan jaarlijkse inkomsten kan winnen als het hardnekkige raciale ongelijkheid aanpakt.

Bedrijven die AI-oplossingen ontwikkelen en implementeren kunnen eveneens veel winnen door het actief implementeren van processen die vooroordelen in hun AI-oplossingen verminderen. Onderstaand voorbeeld heeft dan wel niet direct te maken met bias, maar het is wel een schrijnende illustratie van hoe techniek het leven van mensen soms moeilijker in plaats van makkelijker kan maken. Proctoring is software om digitaal toezicht te houden op een examen-toets die thuis wordt afgenomen. Blijkbaar is het onvoldoende getest op diverse gezichten, waaronder dus donkere.

Ingebouwde vooroordelen in AI-gerelateerde gezichtsherkenning

Joy Buolamwini, onderzoeker aan het MIT Media Lab, is een ware pionier in het onderzoek naar in kunstmatige intelligentie en gezichtsherkenning ingebouwde vooroordelen. Als afgestudeerd student aan het MIT maakte ze een spiegel die ambitieuze beelden op haar gezicht projecteerde, zoals een leeuw of tennisster Serena Williams. Maar de gezichtsherkenningssoftware die ze installeerde, werkte niet op haar zwarte gezicht, totdat ze letterlijk een wit masker opzette. (Wood, 2021)

Een recent voorbeeld is van Google. De zoekmachine onthulde een dermatologie-app die volgens eigen zeggen 288 verschillende huidaandoeningen van foto's kan herkennen. Daar is iets heel Google-achtigs aan. Het deep learning-systeem waarop de app is gebaseerd, werd oorspronkelijk getraind en getest op een dataset waarin mensen met een donkere huidskleur – net als binnen het bedrijf zelf - sterk ondervertegenwoordigd waren.

Om de taak te volbrengen, gebruikten de onderzoekers een trainingsdataset van 64.837 beelden van 12.399 patiënten uit twee staten. Maar van de duizenden afgebeelde huidaandoeningen was slechts drieënhalf procent afkomstig van patiënten met Fitzpatrick huidtypes V en VI, die respectievelijk een bruine huid en een donkerbruine of zwarte huid vertegenwoordigen. Negentig procent van de database bestond uit mensen met een blanke huid, een donkerder blanke huid, of een lichtbruine huid, aldus de studie.

Als gevolg van de bevooroordeelde steekproef, zeggen dermatologen dat de app zou kunnen leiden tot over- of onderdiagnose van mensen die niet blank zijn. (Feathers, 2021)

AI is nog lang niet feilloos. Onlineretailhandelsgigant Amazon verwijderde onlangs het N-woord uit een productbeschrijving van een zwartgekleurd actiefiguurtje. Ook gaf Amazon toe dat zijn veiligheidsmaatregelen er niet in geslaagd

zijn de racistische term uit te filteren. Het in China gevestigde bedrijf dat de goederen verkocht, had waarschijnlijk geen idee wat de Engelse beschrijving inhield, omdat een kunstmatig intelligentie (AI) taalprogramma de inhoud produceerde.

Experts op het gebied van AI zeggen dat bovenstaande deel uitmaakt van een groeiende lijst van voorbeelden waar real-world toepassingen van AI-programma's racistische en bevooroordeelde resultaten uitspuwen. (Jorge Barrera, 2021)

> *"Het opnemen van herinneringen, het lezen van gedachten en het manipuleren van wat een ander ziet via een apparaatje in hersenen lijken misschien science fiction-complotten over een verre en onrustige toekomst. Maar een team van multidisciplinaire onderzoekers zegt dat de eerste stappen naar de uitvinding van deze technologieën al zijn gezet. Door middel van een concept dat 'neuro-rechten' wordt genoemd, willen zij waarborgen invoeren voor ons kostbaarste biologische bezit: ons brein."*

Het NeuroRights Initiative, opgericht door neurowetenschapper Rafael Yuste van de Columbia-universiteit, staat vandaag aan het hoofd van deze groeiende inspanning.

Ook in Nederland zie je steeds meer initiatieven. Het Nederlands Instituut voor Beeld en Geluid, NPO en RTL Nederland hebben een intentieverklaring opgesteld voor het verantwoord gebruik van AI in media. De mediaorganisaties onderschrijven dat ze zich bij het toepassen van kunstmatige intelligentie aan ethische richtlijnen zullen houden.

Fontys ICT-docent en senior onderzoeker Petra Heck deed de afgelopen 2 jaar onderzoek naar AI engineering: hoe maak je AI-systemen die betrouwbaar genoeg zijn om in een productie-omgeving te gebruiken. Dat is echt anders dan bij traditionele software zonder AI. Bij traditionele software programmeer je alle regels zelf en weet je dus veel beter waar gek gedrag zou kunnen zitten. Bij AI-software gebruik je een algoritme waar je niet in kunt kijken en kun je alleen testen door de juiste datavoorbeelden aan dat algoritme te geven.

In dat laatste schuilt het probleem. Wat als in jouw voorbeelden alle vrouwen naar de HAVO gingen en alle mannen naar het VWO? Dan concludeert het algoritme dus dat mannen slimmer zijn dan vrouwen. Vaak zijn dit soort effecten (bias) veel subtieler en er dus ook lastig uit te halen. Gelukkig komen er steeds meer tools voor data scientists en software engineers om bias al tijdens het bouwen van het AI-systeem op het spoor te komen. In haar post over het testen van AI-systemen noemt Petra er een aantal onder de kopjes 'fairness' en 'interpretability'. Als data scientists en software engineers zorgen dat deze guidelines vertaald worden naar tools voor praktische toepassingen, is het zeker mogelijk om AI in te zetten voor het oplossen van maatschappelijke problemen zonder dat de burger daarbij benadeeld wordt.

Steeds meer bedrijven proberen te werken aan een algoritme waar vooroordelen zoveel mogelijk uit worden gehaald. Facebook heeft nu een nieuwe dataset van 45.186 video's opengesteld om eerlijkheid in computervisie en audiomodellen met betrekking tot leeftijd, geslacht, zichtbare huidskleur en omgevingsverlichting te helpen evalueren.

EXPERTS AAN HET WOORD

Maria Luciana Axente

Responsible AI & AI for Good Lead, PwC UK

Algoritmen kunnen heel duidelijk aangeven waar er sprake is van vooringenomenheid en discriminatie, maar zij zullen niet de oplossing bieden om dit te verhelpen. Algoritmen zijn het vergrootglas waarmee we de kwaadaardige tumor kunnen zien, maar zeker niet het scalpel om hem te verwijderen, noch de arts die weet hoe hij dat moet doen zonder een slagader te raken. Het is dus aan ons om discriminatie en ongelijkheid te verhelpen, niet aan algoritmen. Die moeten het middel, niet het doel, bij het aanpakken van maatschappelijke problemen zijn. En het moeten middelen zijn die met veel menselijke intelligentie worden ingezet om te begrijpen wat er waar mis is, zodat we collectief opnieuw kunnen nadenken over inclusie en gelijkheid.

> *“Algoritmen zijn het vergrootglas waarmee we de kwaadaardige tumor kunnen zien, maar zeker niet het scalpel om hem te verwijderen.”*

Eerst moeten we de context begrijpen waarin die algoritmen worden gebouwd. Door wie en op welke belangen zijn ze afgestemd? Wat is de context van het gebruik ervan? En wie wordt er op welke wijze door beïnvloed? Alleen door deze kijk op de algoritmen zullen we in staat zijn inclusiviteit te zien als een thema dat door de hele levenscyclus heen loopt. Met inbegrip van het monitoren ervan, terwijl het wordt gebruikt om driftige prestaties vast te leggen. Op basis van de impact van de uitkomst, de ernst en de waarschijnlijkheid ervan, zouden we algoritmen moeten indelen in risicoklassen, wat sterk lijkt op wat de nieuwe verordening van de Europese Commissie over AI zou betekenen. De indeling en classificatie van risico's zou ons helpen beter te begrijpen in welke gevallen we in de nabije toekomst een mens in het spel moeten houden. Ook moeten we ons zoveel mogelijk verzetten tegen automatisering, totdat we bevredigende niveaus hebben bereikt op het gebied van diversiteits- en integratie-indicatoren.

Emma Beauxis-Aussalet

Assistant professor of ethical computing, Vrije Universiteit Amsterdam
Lab manager, Civic AI Lab

> *"Om algoritmen te gebruiken voor het algemeen belang, moeten we mensen naar het algoritme brengen en zorgen voor een eerlijk aandeel in de voordelen ervan."*

Er zijn allerlei algoritmen ontwikkeld voor het algemeen belang, bijvoorbeeld om verkeersdrukte, energieverbruik of voedselverspilling te verminderen. Nieuwe algoritmen (die 20 jaar geleden nog amper beschikbaar waren) openden nieuwe mogelijkheden om onze samenlevingen te verbeteren. Denk aan chatbots (om bijvoorbeeld mondelinge cultuur te bewaren of als ondersteuning in geestelijke gezondheid) of computervisie (bijvoorbeeld om zwerfvuil te traceren, wilde dieren in de gaten te houden en stropers of vervuiling op te sporen). Voor meer voorbeelden zie het AI for Good-initiatief van de VN of de conferenties FAccT, AIES, en workshops/tracks bij AAAI, ECML-PKKD, ICML en NeurIPS.

Het is zeker mogelijk om het algemeen welzijn aan te pakken. Maar het vereist een diep inzicht in het domein en de menselijke ecosystemen, anders dreigen potentiële nadelige effecten, Zoals het controversiële Aadhaar-systeem voor voedseldistributie in India. Dat biometrische ID-systeem werkte met vingerafdrukidentificatie, wat mensen met beschadigde vingers of wonende in gebieden met slechte connectiviteit uitsloot.

Zelfs goed ontworpen, kan een technologie voor het algemeen welzijn veranderen in een bevoorrecht goed. Elke gezondheidstechnologie zou bijvoorbeeld een gemeenschappelijk goed kunnen zijn, aangezien iedereen het bij ziekte nodig heeft. Is deze technologie echter voor sommigen ontoegankelijk, bijvoorbeeld vanwege financiën, dan is het niet langer een gemeenschappelijk goed. Dus om algoritmen te gebruiken voor het algemeen belang, moeten we mensen naar het algoritme brengen. We moeten zorgen voor een eerlijk aandeel in de voordelen van algoritmen.

In zekere zin zijn algoritmen ook een gemeenschappelijk goed. Als gemeenschappelijke goederen collectief moeten worden beheerd, moeten de algoritmen die een gemeenschappelijk goed beheren of creëren dat ook zijn. We moeten de mensen dus ook aan de ontwerptafel uitnodigen.

Het gebruik van algoritmen om discriminatie en ongelijkheid te bestrijden is het onderzoeken waard. Quota zijn bijvoorbeeld een eenvoudig algoritme om ongelijkheid te bestrijden en algoritmen kunnen ook worden gebruikt om bepaalde quota te garanderen.

> ***“ Als mensen niet bereid of in staat zijn hun organisaties of gedragingen ingrijpend te veranderen, blijven algoritmen inefficiënte - zo niet schadelijke hulpmiddelen. ”***

Algoritmen kunnen misschien wat menselijke en maatschappelijke vooroordelen wegnemen, maar we ruilen ze in voor algoritmische vooroordelen. Geen enkel algoritme is perfect en hun onvermijdelijke fouten zijn een vorm van vooringenomenheid. We kunnen bepaalde foutenniveaus tolereren en ervoor zorgen dat ze gelijk blijven voor alle bevolkingsgroepen. Maar we moeten rekening houden

met de schaal waarop algoritmen worden ingezet en met de variabiliteit van gegevens en fouten in het echte leven.

Een minieme toename van het foutenpercentage kan gevolgen hebben voor duizenden mensen en de foutenpercentages in het echte leven kunnen verschillen van die welke in testsets zijn waargenomen. Willekeurige variaties kunnen al leiden tot veel grotere foutverschillen tussen bevolkingsgroepen dan in testomstandigheden. Grotere variaties ontstaan en blijven onopgemerkt, als populaties in de loop van de tijd evolueren of verkeerd in testgegevens worden weergegeven. Tenslotte kunnen de trainings- en testgegevens menselijke vooroordelen uit het verleden weerspiegelen en algoritmen zouden deze slechts reproduceren.

Dus in plaats van de verscheidenheid aan fouten en vooringenomenheid van individuen (zowel discriminerend gedrag als eerlijke fouten), zouden we één enkele vorm van algoritmische vooringenomenheid hebben. De impact van een bevooroordeeld algoritme is veel groter en systematischer dan die van één enkel bevooroordeeld mens. Daarom moeten we zorgvuldige afwegingen maken, vraagtekens zetten bij het ontstaan van algoritmische vertekening en fouten en vertekeningen in testgegevens en het echte leven nauwkeurig onderzoeken.

Hoe dan ook, het probleem van discriminatie en ongelijkheid kan niet uitsluitend met computersystemen worden aangepakt. Discriminatie en ongelijkheid zijn diep geworteld in mensen en samenlevingen. Algoritmen zijn slechts hulpmiddelen. Als mensen niet bereid of in staat zijn hun organisaties of gedragingen ingrijpend te veranderen, blijven algoritmen inefficiënt - zo niet schadelijk omdat fundamentele kwesties over het hoofd worden gezien (om nog maar te zwijgen van ethische kwesties.

Aan de ontwerptafel kan blijken dat discriminatie en ongelijkheid ook zonder algoritmen moeten worden opgelost. Om de juiste problemen aan te pakken

en geen ergere te creëren, moeten ontwerp en evaluatie van algoritmen in veel disciplines, van technologie tot menswetenschappen, tegen ongelijkheid geborgd zijn.

De besluitvormingsprocessen moeten ook inclusief zijn, gedurende de hele levenscyclus van een algoritme, Maar vooral bij ontwerp, evaluatie, inzet en controle van algoritmen. Inclusiviteit gaat niet alleen over het beheersen van foutenpercentages en het tegengaan van bevoordeling of discriminatie van bevolkingsgroepen door lagere of hogere foutenniveaus. Inclusiviteit gaat ook over de vraag welke algoritmen worden gemaakt, voor welke doeleinden en met welke afwegingen.

Inclusiviteit is geen toegevoegde eigenschap van algoritmen en wordt niet alleen bereikt met metingen en audits. Het kan immers niet worden gewaarborgd zonder overleg met de getroffen bevolkingsgroepen. Zonder de bevolkingsgroepen waar het om draait bij de besluitvormingsprocessen te betrekken,

> ***“ Niet-technische mensen moeten hun bezorgdheid kunnen uiten en kunnen meedenken met technische mensen. ”***

worden belangrijke aspecten van het algoritmeontwerp en de praktische gevolgen ervan gemakkelijk over het hoofd gezien of niet goed afgewogen.

Er is een heel ecosysteem van mensen om rekening mee te houden: mensen die algoritmen gebruiken, mensen van wie gegevens worden verzameld, mensen wier omgeving door algoritmen wordt bewaakt ...

Om de verscheidenheid aan betrokken mensen aan de besluitvormingsprocessen te verbinden, moeten we elkaars geletterdheid ontwikkelen. Het komt erop neer dat niet-technische mensen hun bezorgdheid moeten kunnen uiten en moeten kunnen meedenken met technische mensen. En technische mensen moeten ernaar streven de perspectieven van niet-technische mensen te begrijpen. Technische mensen moeten er ook naar streven algoritmen transparant, begrijpelijk en uitlegbaar te maken voor niet technische mensen. Anders hebben technische mensen aanzienlijke macht en machtsonevenwichtigheid en tunnelvisies zijn een bedreiging voor inclusiviteit.

Siri Beerends

Tech-onderzoeker bij SETUP & promovenda, Universiteit Twente

Algoritmen worden gezien als een duizenddingendoekje voor maatschappelijke problemen. Maar veel van die problemen passen niet in een wiskundig model. Algoritmen reflecteren niet alleen menselijke vooroordelen en ongelijkheden, ze leggen deze ook nog eens vast in systemen. Hierdoor worden ongelijkheden op een nog grotere schaal verspreid. Zo hebben we gezichtsherkenningssoftware met een etnische bias, risico-indicatiesystemen die minderheden stelselmatig benadelen en algoritmen die mensen van kleur als risicogeval kwalificeren, waardoor ze moeilijker in aanmerking komen voor een huis, lening of baan.

Omdat algoritmen een aura van objectiviteit dragen, denken we met behulp van algoritmen eerlijkere beslissingen te kunnen nemen. Maar dat beeld klopt niet. De normen en waarden van de mensen die de algoritmen ontwerpen, werken door in de systemen. De datasets waarmee de systemen worden getraind bevatten vaak dezelfde vooroordelen, culturele stereotyperingen en maatschappelijke ongelijkheden als de analoge wereld.

“ Sociale ongelijkheid en racisme zijn geen ‘foutjes in de software’, maar structureel ingebed in onze technologieën. ”

Nu algoritmische bias serieus wordt genomen, zoeken we de oplossing in het verbeteren van onze algoritmen. Dat is nu eenmaal makkelijker dan het aanpassen van menselijk gedrag, is de heersende gedachtegang. Maar je kunt mens en algoritme niet los van elkaar zien; ze liggen in elkaars verlengde. En omdat de mens een lange geschiedenis kent van discriminatie en ongelijkheid, moet je eerst met dat dieper gewortelde probleem aan de slag.

Sociale ongelijkheid en racisme zijn namelijk geen 'foutjes in de software', maar structureel ingebed in onze technologieën. We proberen dat op te lossen door 'goede' waarden in het algoritme te stoppen, gevarieerde datasets te gebruiken en de 'juiste' variabelen op de 'juiste' manier te wegen. Maar de praktijk blijkt weerbarstig.

Waar halen we gevarieerde datasets vandaan als deze data niet bestaan? En wie mogen bepalen wat 'de juiste waarden' en 'de juiste variabelen' zijn? Tot nu toe zijn dat de mensen geweest die behoren tot de status quo en de meetbare meerderheid. Iedereen die niet in de meerderheidsmodelletjes past, heeft het nakijken. Het zou helpen als meer groepen hun macht kunnen claimen in het ontwerp van technologie. Maar ook dat is niet het hele verhaal.

> ***"Historische criminaliteitsdata zijn biased, onder andere doordat minderheden vaker aan politiecontroles worden onderworpen."***

Doordat we de focus leggen op het debiasen en inclusiever maken van algoritmen, slaan we automatisch de vraag over of algoritmen überhaupt wel het juiste middel zijn om in te zetten. Als het gaat om complexe maatschappelijke vraagstukken zoals welzijnskwesties en het voorspellen van menselijk gedrag, zijn algoritmen niet altijd het beste middel. Neem bijvoorbeeld predictive policing.

Met big data en algoritmen probeert de politie te voorspellen waar de meeste criminaliteit zal plaatsvinden.

Hierover schreef ik samen met Dr. Remco Spithoven een paper in het wetenschappelijke Tijdschrift voor Veiligheid. We zien dat historische criminaliteitsdata biased zijn, onder andere doordat minderheden vaker aan politiecontroles worden onderworpen. Als je een algoritme loslaat op deze data om naar patronen te zoeken, dan rolt diezelfde bias uit het voorspellende model. Je kunt vervolgens proberen het model te debiasen, maar beter is om uit te zoeken hoe het komt dat minderheden vaker aan controles worden onderworpen, welke fouten daarbij gemaakt worden en wat we kunnen doen, zodat deze patronen zich niet eeuwig blijven herhalen.

We kunnen de sociaal-economische ongelijkheden die algoritmen blootleggen dus niet oplossen met een technologische fix. Algoritmen kunnen pas ten goede veranderen als we zelf inclusiever worden. In plaats van algoritmen te bouwen met historische data die ongelijkheden herproduceren, moeten we de sociale mechanismen die tot deze ongelijkheid hebben geleid aanpakken. Bijvoorbeeld door te stoppen met het mathematiseren van dubbelzinnige kwesties die je niet kunt doorgronden en oplossen met binaire logica.

Tot slot is het belangrijk om te beseffen dat inclusie en gelijkheid ook gaan over het recht om te mogen ontsnappen aan de eigen culturele, sociale of 'biologische' identiteit. Met de inzet van algoritmen doen we precies het tegenovergestelde: we categoriseren mensen op basis van kwantificeerbare eigenschappen zoals gender, postcodegebieden, koopgedrag, culturele achtergrond, muzieksmaak, enzovoort.

Algoritmen zijn daarmee één groot festijn van sociale categorisering en culturele stereotypering. Like jij op social media bijvoorbeeld berichten over tennis, word je automatisch ingedeeld bij de categorieën 'sportief', 'politiek conservatief' en

'hoog inkomen'. Vervolgens krijg je nieuwsberichten, video's en politieke reclames aangeboden die bij jouw categorieën 'horen'. Daardoor worden we steeds minder uitgedaagd om af te wijken van de door algoritmen gebaande paden.

Uit onderzoek blijkt dat specifieke groepen aanbiedingen en informatie mislopen vanwege de stereotyperende categorieën die het algoritme voor ze heeft vastgesteld. Een voorbeeld is een vacature voor taxichauffeur die via Facebook alleen Afro-Amerikanen bereikte en een vacature voor kassamedewerker die alleen bij vrouwen belandde.

> ***"Algoritmen kunnen pas ten goede veranderen als we zelf inclusiever worden. We moeten de sociale mechanismen die tot deze ongelijkheid hebben geleid aanpakken."***

Ik hoop dat we maatschappelijke inclusie ruimer, minder categorisch en minder kwantitatief gaan benaderen. Dus niet alleen vanuit cultuur en gender, maar ook vanuit een diversiteit aan morele waarden, ideologieën, levensfilosofieën, communicatiestijlen, praktijkervaringen, vakgebieden en andere aspecten die zich juist niet goed laten kwantificeren, maar voor ons wel heel waardevol zijn.

Rudy van Belkom

Toekomstonderzoeker, Stichting Toekomstbeeld der Techniek (STT)

In de afgelopen jaren hebben uiteenlopende bedrijven, onderzoeksinstellingen en overheidsorganisaties verschillende principes en richtlijnen opgesteld voor 'ethical AI'. Ondanks de grote hoeveelheid is er slechts een beperkte spreiding van richtlijnen.

Onderzoekers van ETH Zürich analyseerden in 2019 maar liefst 84 ethische richtlijnen die de afgelopen jaren wereldwijd zijn gepubliceerd. Van de private sector tot maatschappelijke organisaties en overheden. Uit het onderzoek blijkt dat de meeste ethische richtlijnen uit de VS (21), Europa (19) en Japan (4) komen. De grootste 'richtlijn-dichtheid' is te vinden in het Verenigd Koninkrijk. Daar werden maar liefst dertien ethische richtlijnen gepubliceerd.

Foto: Stephanie Elmensdorp

> *"Tot nu toe domineren met name de rijkere landen die de wereldwijde discussie over de regulering van AI."*

Beperkte spreiding

Hoewel het overzicht een momentopname is (zo zijn er na de publicatie van het onderzoek in China twee nieuwe richtlijnen gepubliceerd), geeft het wel een duidelijke verdeling weer. Het zijn tot nu toe met name de rijkere landen die de wereldwijde discussie over de regulering van AI domineren. Hoewel sommige ontwikkelingslanden betrokken waren bij internationale organisaties die richtlijnen opstelden, hebben slechts enkele hun eigen ethische principes daadwerkelijk gepubliceerd. Dit is volgens de onderzoekers echter wel van groot

belang, aangezien verschillende culturen verschillende opvattingen hebben over AI. Er is mondiale samenwerking nodig om in de toekomst zorg te dragen voor ethische AI die bijdraagt aan het welzijn van individuen en samenlevingen.

Eerste poging

Een eerste poging om mondiale samenwerking te realiseren is gedaan door de Organization for Economic Co-operation and Development (OECD). De OECD, een coalitie van landen die zich inzet voor het bevorderen van democratie en economische ontwikkeling, heeft in 2019 een reeks van vijf principes aangekondigd voor de ontwikkeling en inzet van AI. Maar bijvoorbeeld China valt buiten de OECD en is dus niet meegenomen in de totstandkoming van de richtlijnen. De uiteengezette principes lijken in contrast te staan met de manier waarop AI daar wordt ingezet. Met name wanneer het aankomt op gezichtsherkenning en het toezicht op etnische groepen die met politieke dissidentie zijn geassocieerd. Maar juist bij tegenstrijdige opvattingen is het van belang om elkaar op te zoeken en tot een vorm van consensus te komen.

> *"Er is mondiale samenwerking nodig om in de toekomst zorg te dragen voor ethische AI die bijdraagt aan het welzijn van individuen en samenlevingen."*

Context ontbreekt

Het formuleren van ethische principes en richtlijnen vormt een belangrijke eerste stap bij de realisatie van ethisch verantwoorde AI-toepassingen. Het is echter niet eenvoudig om deze richtlijnen te vertalen naar de praktijk. We willen uiteraard allemaal dat bij de inzet van AI-systemen mensen eerlijk behandeld worden en niet worden benadeeld op basis van bijvoorbeeld geslacht of afkomst.

Fairness is dan ook een veelgebruikt principe in ethische richtlijnen en assessment tools. Toch is het niet eenvoudig om te bepalen wat precies 'fair' is. Dit vraagstuk houdt filosofen al een paar honderd jaar bezig. Er bestaat geen eenduidig beeld van hoe de samenleving eruit zou zien wanneer er geen oneerlijkheid meer zou bestaan. Is een samenleving waarin iedereen exact hetzelfde behandeld wordt überhaupt wel eerlijk? Door de komst van AI krijgt dit vraagstuk een nieuwe dimensie. Het concept van eerlijkheid moet namelijk in wiskundige termen worden uitgedrukt. Denk bijvoorbeeld aan het gebruik van AI in het rechtssysteem.

> ***"De komst van AI geeft het vraagstuk van wat fair is een nieuwe dimensie. Het concept van eerlijkheid moet in wiskundige termen worden uitgedrukt."***

Door de inzet van predictive policing kan crimineel gedrag worden voorspeld door middel van grootschalige monitoring en data-analyses. Er bestaat echter altijd het risico dat mensen zonder de gestelde criteria toch positief scoren (false-positives) en dat mensen die wel aan de gestelde criteria voldoen toch negatief scoren (false-negatives). Wat is in dit geval fair? Zet je mensen mogelijk onterecht vast, of riskeer je dat ze een misdaad begaan?

Professor Yoshua Bengio

Een van ‘s werelds leidende AI-experts en pionier in deep learning


Foto: Maryse Boyce

We kunnen algoritmen op veel manieren gebruiken voor het algemeen welzijn, onder meer om discriminatie en ongelijkheid te helpen bestrijden. Allereerst moet het woord algoritme worden verduidelijkt. Alles wat op een computer draait, is gebaseerd op een algoritme. AI-algoritmen zijn speciale gevallen van algoritmen, gericht op het bereiken van vaardigheden waarvoor een vorm van intelligentie nodig is die typisch wordt toegeschreven aan mensen of andere dieren.

> *“ In het algemeen hebben computers geen inzicht in morele waarden of menselijke psychologie. ”*

Machine-learning algoritmen zijn AI-algoritmen gebaseerd op het vermogen van machines om te leren van ervaring en gegevens. Deep learning algoritmen zijn machine learning algoritmen die geïnspireerd zijn door hersenen. Er bestaan algoritmen voor machinaal leren om bijvoorbeeld vooroordelen in tekst te leren opsporen. Maar in het algemeen hebben computers geen inzicht in morele waarden of menselijke psychologie.

Wanneer wij ze trainen voor bepaalde taken, hebben wij dus mensen nodig om gegevens te labelen (bijvoorbeeld als uiting van een racistische opvatting). Hiervan gaat de machine leren. Vervolgens kunnen ze worden opgeschaald om processen te automatiseren, bijvoorbeeld het markeren van mogelijk discriminerende uitspraken in tekst. Omdat computers echter nog lang niet het menselijk begripsniveau hebben bereikt, zouden de gemarkeerde uitspraken waarschijnlijk door mensen moeten worden gecontroleerd. De computer zou de mens dus werk kunnen besparen, maar dit soort werk niet volledig kunnen automatiseren. In ieder geval niet in de nabije toekomst.

Dit is precies de reden waarom we regelgeving en bestuursmechanismen moeten invoeren, met inbegrip van vormen van transparantie en controle die het voor onafhankelijke partijen (idealiter vertegenwoordigers van de staat) mogelijk maken om deze algoritmen te valideren. Hoe zijn ze ontworpen? Welke soort gegevens zijn gebruikt? Waren ze voldoende representatief? Om zo schade te voorkomen en collectief overeengekomen waarden alom te respecteren. We doen dit voor vliegtuigen om de veiligheid van reizigers te garanderen en we zouden soortgelijke dingen ook moeten doen voor computerplatforms die grote aantallen mensen aangaan, zoals social media, banken, platforms voor e-handel of verzekeringsmaatschappijen.

Rick Bouter

Co-founder en voorzitter, Techthics

Techthics is een christelijke platform welke zich bezing op de impact van technologie op ethiek en christelijke religie. In het dagelijks leven is Rick werkzaam bij Accenture.

Foto: Myrthe Karels

> *“ AI maakt niet per definitie eerlijk. ”*

Bewustwording, bezinning, actie

Als we ons niet bewust zijn van zaken als bias en inclusie kunnen we ons er ook niet goed op bezinnen. Bezinning is nodig om een gewenste situatie te definiëren en vervolgens acties te ondernemen die naar deze ‘future state’ wijzen. Maar, voordat we met elkaar de race naar inclusiviteit beginnen, is het wel goed om de observatie te maken. AI maakt niet per definitie eerlijk. Een goed voorbeeld hiervan is het onderzoek van VU-promovenda Elmira van den Broek. Zij onderzocht de eerlijkheid van het selecteren met algoritmen binnen een grote multinational. “Er kwam allerlei casuïstiek op tafel waarbij de matchscore die AI berekent niet strookt met het beeld dat managers hebben bij een kandidaat.” (Bron: ‘Selecteren met een algoritme: eerlijk of niet?’)

Inclusieve Tech begint bij jezelf

Verder denk ik dat het goed is om niet direct naar anderen te wijzen, maar eerst naar jezelf te kijken. De laatste tijd is er veel kritiek op overheidsinstanties. Maar niet alleen overheidsinstanties zijn debet aan het ondermaats toetsen van hun algoritmen en de uitkomsten/gevolgen hiervan. Om maar te zwijgen van de mitigerende acties die zijn ingesteld om de ongewenste situatie te corrigeren of te voorkomen. Hoe gaat jouw bedrijf of organisatie met inclusieve tech om?

Zelfregulering vs. Eigen-verantwoordelijkheid
Zelfregulering door bedrijven blijkt helaas niet afdoende. Het is duidelijk dat zowel de overheid, de bedrijven als de gebruiker/burger verantwoordelijkheid moet nemen, zelf actie moet ondernemen. Eén van de manieren om burgers te mobiliseren is door de broncode in te stellen (open source), zodat er ook in dit kader proactieve participatie komt.

Ethisch kader
Een volgende observatie is dat het gebruik van algoritmen niet volledig aan een 'vrije markt' overgelaten kan worden. Onafhankelijke en juridische/ethische kaders, onder meer over de bescherming van privacy en het niet toepassen van technologie/data voor verkeerde doeleinden, zijn nodig. Toetsen kan alleen als er een consistent en eenduidig wettelijk kader voor is.

Een eerste start in dit kader van discriminatie en bias (maar ook privacy) kan een classificatie framework zijn. Dit raamwerk zou antwoord moeten geven op een vraag, niet op de situatie specifieke context van een individu. Een voorbeeld is de aan/verkoop van alcohol. Niemand hoeft je naam, geboortedatum, etniciteit te weten. Een vinkje dat voor de verkopende partij aangeeft dat de aankopende partij gerechtigd is, is afdoende.

Data-eigenaarschap
Ook zaken als datakwaliteit en minimalisatie kunnen helpen. Dit zou via data-eigenaarschap kunnen worden geregeld. Hierbij krijgt en behoudt het individu het recht om voor de gestelde vraag de juiste data te delen.

> ***"Er is een breder kader nodig om AI te duiden en op een ethisch/filosofisch niveau te bepalen welke rol het in ons leven moet spelen."***

Big Tech

Big Tech is de laatste tijd, terecht, in opspraak. Waar wijzelf continu het (data) product zijn van techgiganten, moeten we ons realiseren dat we hiermee het recht van privacy ook deels zelf hebben opgegeven. De laatste tijd zien we een kanteling waarbij mensen zich bewuster zijn van het recht van privacy en de persoonlijke data die ze delen. Een goede ontwikkeling; verandering begint bij jezelf.

> ***"Als we technologie ten volle willen benutten hebben we mensen nodig die ethisch verantwoorde keuzes kunnen maken."***

Techno-kritischeid geboden

Vandaag de dag is inclusie een sterk buzzword. Echter is het volgens mij maar een klein stukje van de responsible technology puzzel. Er is een breder kader dan inclusie nodig om technologieën als AI te duiden en om op een ethisch/filosofisch niveau te bepalen welke rol technologie in ons leven moet spelen. Als dit ethisch/filosofisch kader staat, is het ook makkelijker hier proactief beleid op te maken, voor zowel overheid als bedrijfsleven.

Daarom roep ik mensen op om zich, ongeacht levensovertuiging, rang of ras, te bezinnen op een constructief-kritische houding ten opzichte van technologie. Vooral roep ik de overheids- en onderwijssector op om dit te integreren in het onderwijs. Een terecht grote zorg is het aantal digibeten in Nederland: naar verluidt zijn zo'n vier miljoen mensen in Nederland laag digitaal vaardig. (Bron: 'Stichting lezen en schrijven'). Digitale gelijkheid start bij fysieke ongelijkheid. Vervolgens is en blijft technologie een middel. Wel een middel waar tegenwoordig niemand meer omheen kan. Als we technologie ten volle willen benutten (de positieve zaken omarmen en de negatieve kant voorkomen) hebben we mensen nodig die ethisch verantwoorde keuzes kunnen maken. Zo helpen we allemaal mee aan responsible tech. Ook jij.

Dr. Marijke Brants

Onderzoeker digitaal en duurzaam ondernemen

Een aantal pijlers is essentieel voor een inclusieve samenleving: een arbeidsmarkt, woningbeleid en onderwijssysteem die niet discrimineren. Discriminatie is echter nog steeds een wezenlijk probleem in onze samenleving en binnen deze respectievelijke domeinen.

Artificiële Intelligentie (AI) kan hierin ondersteuning voor een oplossing bieden. Door patronen te analyseren en objectieve(re) criteria vast te leggen, of door het blootleggen van bias in historische data (bijvoorbeeld in het aanwervingsbeleid). Deze ondersteuning is echter enkel een meerwaarde wanneer we AI op een ethische manier inzetten en de mens centraal blijven stellen.

Onbedoelde bias kan een AI-algoritme binnensluipen, waardoor het op een systematische en oneerlijke wijze bepaalde personen discrimineert. Er bestaat niet zoiets als een neutraal algoritme. Het creëren van iets nieuws brengt onvermijdelijk keuzes met zich mee die de eigenschappen van het eindproduct mede bepalen. Deze keuzes worden (te) vaak alleen gemaakt door de technische ontwikkelaars, die niet (altijd) de meeste kennis hebben over het onderwerp/thema waarvoor het algoritme een oplossing biedt.

> ***“ De ondersteuning door AI is echter enkel een meerwaarde wanneer we het op een ethische manier inzetten en de mens centraal blijven stellen. ”***

Ondanks dit uitgangspunt worden door een algoritme gemaakte aanbevelingen en selecties gewoonlijk gepresenteerd alsof ze inherent vrij zijn van (menselijke) vooroordelen, enkel en alleen omdat de beslissingen 'gebaseerd zijn op gegevens'. Dit is echter een misvatting. Er is grote noodzaak om bij aanvang van de ontwikkeling van een AI-algoritme voldoende ondersteuning en aandacht te schenken aan potentiële bias en onbedoelde neveneffecten, niet enkel vanuit een technisch oogpunt.

> ***"Er is grote noodzaak om bij aanvang van de ontwikkeling van een AI-algoritme voldoende ondersteuning en aandacht te schenken aan potentiële bias en onbedoelde neveneffecten."***

Binnen de onderzoeksgroep Creative and Innovative Business loopt momenteel een project in opdracht van de cel Gelijke Kansen van de federale overheid: RaiS. Binnen dit project ontwikkelen en dissemineren we een instrument dat bedrijven en organisaties laat nadenken over een ethische en non-discriminerende manier om AI in te zetten in het selectie- en rekruteringsproces.

We werken binnen RaiS met een sterk co-creatietraject waarbij we verschillende partijen (tech, HRM-professionals, KMO's, rekruteringbureaus, et cetera.) samenbrengen. Hierbij is aandacht voor verschillende vormen van bias: bestaande bias (in historische data), technologische bias (inherent aan het gebruikte algoritme) en 'emergent' bias (problemen die ontstaan door al dan niet plotselinge veranderingen in de samenleving). We onderzoeken ook de rol die 'explainable' AI kan spelen.

Gabriela Arriagada Bruneau

Postgraduate Researcher AI & Data Ethics Inter-disciplinary Ethics Applied Centre University of Leeds

We kunnen en we moeten algoritmen inzetten voor het algemene belang. Een groot deel van het debat heeft zich gericht op het kijken naar vooroordelen en discriminatie, genderongelijkheid en racisme, maar zij zien het als een onvermijdelijk probleem dat zich in de gegevens bevindt en dat wordt versterkt door algoritmen. Dit perspectief heeft het veld geholpen door de bewustwording van de mogelijke slechte werking en kwaadaardige gevolgen van datagestuurde technologieën, maar het is tijd om aan dit probleem een laag van diepte toe te voegen.

Om het algemeen welzijn te bevorderen en kwesties van discriminatie en ongelijkheid te bestrijden, moeten technische oplossingen een minimumvereiste zijn. Het gebruik van algoritmen om ethische waarden als eerlijkheid en betrouwbaarheid te bevorderen vereist dat ze worden gezien als maatschappelijke instrumenten. Dit betekent dat ze niet simpelweg kunnen worden gezien als geavanceerde manieren om de nauwkeurigheid of efficiëntie te verhogen, maar dat algoritmen moeten worden geconstrueerd met een waarde-ontwerp en een mensgerichte benadering in gedachten.

Als zij worden gezien als socio-technische instrumenten die de samenleving vorm geven en door ons worden gevormd, kan de feedback-lus die wij met op data gebaseerde technologieën hebben, een troef worden. Dit betekent dat wij onze verwachtingen ten aanzien van deze technologieën gegrond moeten houden en ze niet meer macht moeten toekennen dan wij ze kunnen geven.

Door deze technologieën zorgvuldig af te stemmen op weloverwogen ethische kaders hebben we een grote kans om algoritmen te gebruiken om heersende maatschappelijke ongelijkheden te corrigeren. Hiervoor moet echter verder interdisciplinair werk worden verricht om een dialoog tot stand te brengen die

meer omvat dan alleen richtsnoeren en checklists voor ontwikkelaars. Een dialoog die zich richt op verfijning van de integratie van ethische beginselen in de praktijk.

Ten eerste zou ik niet willen beweren dat AI beslissingen voor ons neemt, bovendien zijn algoritmen niet geschikt om voorspellingen op individuele basis te doen! Een veel voorkomende misvatting is dat zeer autonome systemen oneerlijke beslissingen nemen. Dat is niet het geval. Wat algoritmen kunnen doen is ons een output (voorspellende modellen) of een effect (causale modellen) geven dat een context nodig heeft om te worden ingezet, en vertekend kan zijn. Om de inclusiviteit te waarborgen is dus een nauwgezet controleproces nodig dat een bias bewuste aanpak omvat.

Maar het allerbelangrijkste is dat we ervoor zorgen dat we begrijpen wat de algoritmen doen wanneer ze dat specifieke resultaat opleveren. Dit vereist een robuust begrip van uitlegbaarheid, dat aansluit bij het transparantiebeginsel. Dat

> ***“ Het gebruik van algoritmen om ethische waarden te bevorderen vereist dat ze worden gezien als maatschappelijke instrumenten, geconstrueerd met een waarde-ontwerp en een mensgerichte benadering in gedachten. ”***

beginsel is een van de zeven vereisten voor 'betrouwbare AI' door de deskundigengroep op hoog niveau voor AI (AI HLEG) van de EU-Commissie).

Transparantie staat centraal in het huidige debat over AI, waarbij diverse onderzoekers methoden en technieken ontwikkelen om het 'black-box'-probleem op te lossen.

> ***"Diversiteit is niet iets dat alleen met cijfers kan worden gemeten; hier tellen ook ervaringen."***

Toch houden de meeste van deze opvattingen geen rekening met belangrijke inzichten uit de filosofie en de sociale wetenschappen. Verklaringen vereisen meer dan alleen het kunnen definiëren of identificeren van de elementen die een resultaat produceren. De 'kenmerken' die in het algoritme werkzaam zijn, maken meestal deel uit van een contextueel geheel van variabelen die het algoritme beïnvloeden. En deze variabelen zijn niet alleen in het model te vinden. Er zijn fundamentele elementen die in het debat over verklaarbare AI moeten worden opgenomen en verband houden met de manier waarop mensen begrijpen en beslissingen nemen, maar vaak over het hoofd worden gezien.

Een op waarden gebaseerde benadering zou meer in overweging moeten nemen dan enkel statistisch relevante informatie en de oorzaken buiten de interne mechanismen moeten beschouwen die tot die beslissing hebben geleid. Hoe beïnvloeden sociale factoren dit? Wat is de context van die gegeven verklaring? Wat verwachten we van die verklaring? Zo kunnen we te weten komen waarom algoritmen systematisch discrimineren en hoe we dat kunnen voorkomen.

Een andere factor die veel wordt genoemd, is het diverser maken van de AI-industrie (AI Now Report, 2019). Het is zinloos om meer gediversifieerde

gegevens te verzamelen en de contextualisering te verbeteren als de mensen die het doen geen diverse achtergrond hebben. Diversiteit is niet iets dat alleen met cijfers kan worden gemeten; hier tellen ook ervaringen. Een echt inclusieve oplossing voor algoritmische systemen vereist goede gegevenspraktijken die zijn ingebed in een ontwerp- en toepassingsomgeving met een constructief standpunt. Betere algoritmen zijn mogelijk door minderheden - die vaak worden genegeerd of getroffen door slecht ontwikkelde algoritmen - in het ontwerp-proces te integreren.

Stefan Buijsman

Onderzoeker in de filosofie, Institute for Futures Studies, Stockholm

Algoritmen zijn niet automatisch discriminerend en kunnen ook meer gelijke beslissingen bevorderen. Een algoritme dat kaarten genereert op basis van satellietbeelden kan bijvoorbeeld helpen om ongelijkheid te bestrijden in landen waar anders eerst geld en tijd gestoken moet worden in het vinden van begaanbare routes (al helemaal na natuurrampen).

Het is ook prima mogelijk dat algoritmen die persoonsgegevens verwerken dat op een eerlijke en gewenste manier doen. De grote uitdaging echter is om dat te bereiken bij zelflerende algoritmen, die hun beslissingen baseren op de patronen in grote datasets. In die datasets zit vaak een zekere mate van ongelijkheid (bijvoorbeeld dat verhoudingsgewijs meer mannen worden aangenomen voor programmeursfuncties) en het is ontzettend lastig om ervoor te zorgen dat algoritmen die ongelijkheden niet oppikken en mogelijk uitvergroten.

> *“Als er geen ongelijkheden in de data zitten, voegt het algoritme die niet zomaar toe.”*

Er is dus geen principiële reden waarom algoritmen discriminerend zouden zijn. Als er geen ongelijkheden in de data zitten, voegt het algoritme die niet zomaar toe. Bovendien kunnen algoritmen de problemen meetbaarder maken, doordat we kunnen uitrekenen op welke manier en hoezeer een algoritme ongelijk handelt.

Foto: Merlijn Doomernik

Tegelijkertijd is er een grote 'maar': in de praktijk is het lang niet altijd mogelijk om discriminerend gedrag van algoritmen te voorkomen. Algoritmen die nu automatisch teksten schrijven, associëren moslims vaak met geweld, vanwege de internetteksten waar het algoritme van geleerd heeft. Dat is niet zomaar op te lossen met een aanpassing in de code, ondanks dat het een bekend probleem is en er hard aan gewerkt wordt. Je kunt er daarnaast alleen iets aan doen als je er actief op let. Geavanceerde algoritmen zijn complex genoeg dat niet zichtbaar is op welke manieren ze discrimineren, tenzij je zelf tests uitvoert. Het kan dus, maar het is tegelijkertijd een lastig probleem.

In Europa is het momenteel niet toegestaan om de zelflerende algoritmen die de meeste problemen opleveren te gebruiken voor beslissingen over persoonsgegevens. AI is het alsnog mogelijk om andere discriminerende algoritmen, die immers niet meer dan een set instructies zijn voor het maken van een beslissing, te gebruiken.

> ***"Algoritmen die nu automatisch teksten schrijven, associëren moslims vaak met geweld, vanwege de internetteksten waar het algoritme van geleerd heeft. Dat is niet zomaar op te lossen met een aanpassing in de code."***

Het voornaamste wat we kunnen doen, is algoritmen te blijven zien als instructies die door mensen zijn goedgekeurd. Er is niets mysterieus aan een algoritme, en als het resultaat is dat er discriminatie plaatsvindt, dan moet er dus een ander (en beter) algoritme gebruikt worden. Om dat bij alle algoritmen mogelijk te maken, is het belangrijk zo goed mogelijk te begrijpen waarop algoritmen hun beslissingen baseren, aan welke standaarden die beslissingen moeten voldoen en welke data gebruikt worden om zelflerende algoritmen te trainen (bijvoorbeeld niet alleen foto's van witte mannen). Het is een flinke uitdaging, maar wel een die heel belangrijk is voor een goed gebruik van algoritmen.

Tessa Cramwinckel

Onderzoeker fair and explainable AI

Als een algoritme berekent wat de ratio van mannen en vrouwen is binnen een bedrijf en het bedrijf doet daar vervolgens wat mee, is er technisch gezien een algoritme gebruikt voor het algemene belang. In dit geval is het algoritme waarschijnlijk niet veel meer dan een functie van Excel. Dus het eerste wat ik zou afbakenen is de definitie van het woord algoritme.

Ik zie dat het veel gebruikt wordt, met name wanneer men niet precies weet wat er gebeurt. Anders zou je het beestje wel bij zijn naam noemen. Ik denk dat je in deze context eerder complexere modellen binnen het sociale domein bedoelt, zoals neurale netwerken om criminaliteit te voorspellen.

Wanneer een model ongewenst oneerlijk is, heeft dat meestal te maken met de data die oneerlijk zijn en dat heeft meestal weer te maken met de 'oneerlijke' samenleving. Ik denk dat het belangrijk is om dat ten eerste te beseffen.

> ***"Wanneer een model ongewenst oneerlijk is, heeft dat meestal te maken met de data die oneerlijk zijn en dat heeft meestal weer te maken met de 'oneerlijke' samenleving."***

Een algoritme, in de breedste zin van het woord, kan gebruikt worden om dit recht te trekken. Al ben ik soms wat sceptisch, omdat je dan aan symptoombestrijding doet. Dat is niet per se slecht, maar wel als vervolgens te weinig aandacht besteed wordt aan het onderliggende probleem, namelijk een oneerlijke samenleving.

Hier zijn allerlei methodes voor die je ongetwijfeld al hebt opgezocht. En er zijn ook weer verschillende metrieken die meten hoe eerlijk/inclusief een dergelijk algoritme is. Mijn masterthesis pleit ervoor om die metriek, of eigenlijk de definitie van wat eerlijk is, niet aan de programmeur over te laten. De publieke notie van eerlijkheid moet juist worden geïncorporeerd in het model. Hiervoor heb ik een methode ontwikkeld.

Een tweede puzzelstuk in deze kwestie is het zo transparant mogelijk maken van een algoritme. Misschien maakt een algoritme wel een beslissing die overgenomen wordt door beleid, maar het is de kunst om ook die 'waaromvraag' te kunnen beantwoorden. Waarom maakt het algoritme deze beslissing? Dit is eerlijk naar de eindgebruiker toe, maar het is ook een heel mooi controlemechanisme om te kunnen zien of een algoritme wel op een wenselijke manier werkt. Dit is overigens best een lastig onderzoeksterrein, maar wel erg leuk!

Walter Diele

Artificial Intelligence Consultant

Wel eens een whatsappje gemaakt door steeds willekeurig een van de drie gesuggereerde woorden te kiezen? Dat levert misschien een grammaticaal correcte zin op, maar is natuurlijk inhoudelijk onzin, en al helemaal niet wat je had willen zeggen.

Blind vertrouwen op een algoritme voor een goede suggestie is onverstandig. Maar omdat we iets kunstmatige intelligentie noemen (en 'algoritme' een moeilijk woord is), wordt de uitkomst van een model vaak zonder veel vragen te stellen gebruikt. Voorbeelden waarbij dat mis gaat, zijn legio, maar veel ervan die je in de media ziet, zijn duidelijk en bijna grappig.

> ***"Omdat we iets kunstmatige intelligentie noemen, wordt de uitkomst van een model vaak zonder veel vragen te stellen gebruikt."***

Problematischer is het als het niet direct zo duidelijk is, maar het wel delen van de bevolking discrimineert. Mensen discrimineren; onze hersenen kunnen functioneren door allerlei cognitieve biases te gebruiken. De besten onder ons zijn zich daar erg van bewust en proberen het effect te minimaliseren. Maar hoe dan ook levert ons gezamenlijke gedrag data op waar deze discriminatie ingebakken zit. Modellen die deze data voeren, leren de modellen op dezelfde manier te discrimineren.

Valt daar dan niets aan te doen? Wel een beetje; er zijn technieken om data deels te corrigeren voordat het model erop getraind wordt. Gelukkig is daar toenemende aandacht voor, zeker onder data scientists. Maar ook de top van organisaties heeft inmiddels pijnlijk ervaren dat zij verantwoordelijk zijn voor gebruikte modellen. Hopelijk resulteert dat in meer betrokkenheid van management en model governance binnen organisaties, zodat de deel-verantwoordelijkheden voor goed datagebruik op de juiste niveaus in de organisatie neergelegd worden.

Maar kan je het ook omdraaien? Kun je algoritmen gebruiken om menselijke discriminatie te detecteren? Eerlijk gezegd denk ik dat je daar niet heel ingewikkelde modellen voor nodig hebt. Een simpele kruistabel is vaak voldoende om misstanden aan te tonen.

> ***“ Ook de top van organisaties heeft inmiddels pijnlijk ervaren dat zij verantwoordelijk zijn voor gebruikte modellen. ”***

Als hoogopgeleide blanke Nederlandse man van in de vijftig heb ik schokkend weinig ervaring met onheuse behandeling. Ik ben nog nooit staande gehouden door de politie, mijn koffers zijn nog nooit open gemaakt op een vliegveld, mijn kinderen worden op school ingeschat als slim (wat een selffulfilling prophecy is), nooit een probleem gehad om werk te vinden, enz. Natuurlijk allemaal terecht, maar geen toeval.

We kennen ook allemaal de voorbeelden van het effect van westerse versus niet-westerse namen op cv’s voor stages of banen, het veelvuldig aanhouden van donkere profvoetballers of rappers in dure auto’s, verschil in schooladvies per culturele achtergrond, kans op een huurwoning voor allochtonen versus autochtonen … Ga zo maar door.

> *“Iedereen weet dat het onzin is dat een hele raad van bestuur van een organisatie toevallig blank, oud en man is en alleen bij HR de beste kandidaat toevallig een vrouw is.”*

Helaas zijn er geen moeilijke modellen nodig om een klein verschil in behandeling aan te tonen; het verschil is zo groot dat tellen voldoende is. Ook hier ligt de oplossing niet (vooral) bij de data scientist, maar bij de governance. Als bestuurders van organisaties verantwoordelijk gesteld zouden worden voor discriminatie als uit de data (tellingen) blijkt dat de verschillen tussen groepen echt geen toeval zijn, zouden ze er misschien beter op sturen en waarborgen inbouwen. Nu wordt per persoon bekeken of er sprake is van discriminatie, wat vaak lastig te bewijzen is. Maar iedereen weet dat het onzin is dat een hele raad van bestuur van een organisatie toevallig blank, oud en man is (alleen bij HR was toevallig de beste kandidaat een vrouw). En datzelfde geldt voor alle andere vormen van discriminatie.

Als de verschillen zo duidelijk zijn, zou je een organisatie verantwoordelijk moeten kunnen houden voor klaarblijkelijke discriminatie. Ik ben bang dat in dezen de stok het beter doet dan de wortel, omdat bias ingebakken zit in onze hersenen en het dus niet automatisch goedkomt. Ik kijk uit naar de tijd dat we AI modellen nodig hebben om te zien of er wordt gediscrimineerd.

Dr. Steven Dorrestijn

Lector Ethiek & Technologie, Saxxion, University of Applied Sciences

Volgens mij is een belangrijke les van de techniekfilosofie dat techniek ambivalent is. Dat wil zeggen dat techniek ethisch geladen is, dus niet neutraal, en dat de waarde van techniek niet eenduidig is, maar dat er altijd positieve én negatieve kanten aan zitten. Dus ja, AI kan bijdragen aan de bestrijding van discriminatie en ongelijkheid, door bijvoorbeeld slim door data in nieuws, social media, beleidsstukken te zoeken naar voorbeelden van discriminatie. Maar ja, een typisch probleem met AI dat nu overal naar voren komt, is dat het zorgt voor bubbels en daarmee discriminatie en ongelijkheid ook kan versterken.

Bij deze formulering gaat het erom of de automatisering van beslissingen beter en inclusiever zou kunnen worden gemaakt. Dat kan vast en zeker. Partijdigheid, discriminatie in de opgestelde algoritmen of ontstaan of vergroot in zelflerende algoritmen, kunnen bij ontdekking worden gecorrigeerd. Dat is een manier om met de ambivalentie van techniek, van AI, om te gaan. De ontwikkeling moet reflexief worden gemaakt: evaluatie en bijsturing moeten worden ingebouwd. Dit heeft te maken met inclusiviteit in de zin van de vraag wie te maken hebben met de positieve en negatieve effecten van AI.

> *"Techniek is ethisch geladen, en de waarde ervan is niet eenduidig. Er zitten altijd positieve én negatieve kanten aan."*

Ten tweede sluit het transparanter maken van algoritmen aan bij het principe van inclusiviteit: iedereen kan dan meekijken hoe de rekensommen en de beslissingen worden gemaakt. Ten derde wordt inclusiviteit bevorderd door het organiseren van inbreng van meer mensen bij het ontwerp en de ontwikkeling van algoritmen (co-creatie, co-design).

Volgens mij is het voor het goed omgaan met AI ook heel hard nodig om AI op z'n plek te zetten. Het besef van de ambivalente uitwerking ervan zou bij iedereen goed moeten indalen. Dat voorkomt hopelijk een onterecht vertrouwen in de kracht van AI. Het lijkt me goed om als standaardvisie te hebben dat AI ons kan helpen bij ons werk, ons denken en handelen, maar dat het nooit wenselijk is om beslissingen helemaal uit te besteden aan geautomatiseerde systemen.

> ***"Voor het goed omgaan met AI is het heel hard nodig om AI op zijn plek te zetten."***

Een belangrijke vraag die we moeten stellen, is wat we doen met de antwoorden en voorstellen die geautomatiseerde systemen ons geven. Volgen we ze blind op of gebruiken we ze als input voor onze eigen beslissingen? Dus, heel concreet:

- Hoeveel laten we ons voorzeggen door data?
- Hoe houden we 'the human in the loop'?

Die laatste wezenlijke vragen passen bij het praktijkgericht onderzoek op een hogeschool. Ik moet, ter illustratie, altijd denken aan een kunstwerk dat jaren geleden in een stad in de buurt werd geplaatst. De kunstinstallatie verandert van kleur op basis van een peiling van de stemming in de stad, zo begrijp ik na het uitpluizen van berichten op social media.

De stemming van een stad kun je normaal niet zo zien. Nu dus wel. Dat is echt een mogelijkheid van deze tijd. Denk ook aan gezondheidshorloges. Stel je nu eens voor dat zo'n horloge ook je stemming peilt en dat die stemming wordt weergegeven met een van kleur veranderend licht op je hoofd. Op een dag kleurt het rood: boos. Maar je voelt je helemaal niet boos. Wat doe je dan? Gooi je het gezondheidshorloge weg omdat het niet goed werkt? Of denk je: 'Ik voelde me niet boos, maar de data zeggen dat ik diep van binnen wel boos ben dus ja, misschien ben ik toch wel boos. Ik geloof dat ik al iets begin te voelen.'

> ***"Ik voel me niet boos, maar de data zeggen dat ik wel boos ben, dus misschien ben ik toch wel boos."***

Ik geloof dat het problematisch is dat wat de data zeggen te veel prestige heeft. Het valt niet mee om er tegenin te gaan. Een uitdaging voor onze tijd.

Rob Elsinga

National Technology Officer, Microsoft Nederland

De vraag of algoritmen discriminatie en ongelijkheid kunnen helpen bestrijden is lastig, maar ik denk het wel. Bij het ontwerpen, bouwen en implementeren van algoritmen is de kans groot dat discriminatie of uitsluiting ongemerkt optreedt. Technologie en AI kunnen nu al helpen machine-learning-modellen te begrijpen, eerlijkheid in AI te beoordelen en verbeteren en de levenscyclus van AI (ontwerpen, ontwikkelen en toepassen) te managen. Microsoft maakt zulke tools beschikbaar voor AI ontwikkelaars in onze Azure Machine Learning Cloud services.

> *"Verstandig gebruikt, kan de taal van de mensenrechten discussies over verantwoorde AI verrijken."*

De grote beloften van AI en algoritmen zijn vaak gericht op het algemeen belang, zoals verbeteringen in gezondheidszorg, landbouw en klimaat. Bij deze beloften van AI moet wat mij betreft de mens centraal staan. Ik geloof in AI die zich aan ons aanpast en onze vindingrijkheid versterkt. Zinvolle innovaties die ons helpen een positieve impact te maken op de zaken die voor ieder van ons belangrijk zijn.

Gelukkig zijn we het steeds meer eens over het ethische kader en de principes waar verantwoorde AI aan moet voldoen. Het bestrijden van discriminatie en ongelijkheid begint bij de waarden van de organisatie. Deze waarden zijn soms impliciet of expliciet uitgesproken en leiden tot principes die de basis zijn voor elk belangrijk besluit in de organisatie.

Een ander belangrijk aspect hierbij is diversiteit. Het is bij het ontwerpen, bouwen en implementeren van algoritmen belangrijk om diversiteit in teams na te streven. Hoewel we enthousiast zijn over de kansen die AI biedt, erkennen we dat AI betrouwbaar moet zijn om het geaccepteerd te krijgen. Net als in andere discussies over onze waarden, moeten we hier pleiten voor een inclusief en mondiaal perspectief. Verstandig gebruikt, kan de taal van de mensenrechten discussies over verantwoorde AI verrijken.

In de wereld van de clouddiensten vertalen waarden als vrijheid van meningsuiting en persoonlijke privacy zich onder andere in het internationale mensenrechtenrecht en de Universele Verklaring van de Rechten van de Mens. In discussies over kansen en uitdagingen waarmee we worden geconfronteerd om te zorgen dat AI-technologieën (vooral cognitieve diensten) mensgericht blijven, verdienen volgens mij vier bepalingen speciale aandacht. Deze bepalingen vloeien allemaal voort uit het respecteren van de inherente waardigheid van elk persoon:

1. Non-discriminatie: 'Iedereen heeft recht op alle rechten en vrijheden die in deze Verklaring worden uiteengezet, zonder onderscheid van welke aard dan ook, zoals ras, kleur, geslacht, taal, religie, politieke of andere opgave, nationale of sociale afkomst, eigendom, geboorte of andere status.' (artikel 2);
2. Bescherming tegen willekeurige overheidsbemoeienis (artikel 12);
3. Vrijheid van meningsuiting: het recht om allerlei soorten informatie en ideeën te verstrekken en te ontvangen, ongeacht de grenzen (artikel 19);
4. Vrijheid van vereniging: Het recht op vreedzame vergadering en vereniging (artikel 20)

Het vermogen om weloverwogen beslissingen te nemen bij het uiteenzetten van de wil van het volk bij het vaststellen van regels voor de samenleving en het vestigen van het gezag van overheden (artikel 21), dus instellingen zoals onafhankelijke journalistiek en academische vrijheid die de verschillende rechten in de artikelen 18-21 betekenisvol maken.

Om AI verantwoord toe te passen, is het de uitdaging om fundamentele rechten, beginselen, organisatiewaarden en principes om te zetten naar standaarden en naar de praktijk. Het gaat om de vraag welke beslissingen AI zou móeten maken, niet om welke beslissingen het kán maken. Dit vraagt om dialoog tussen verschillende disciplines en een benadering vanuit de wetenschap om te komen tot systemen en tools voor verantwoorde AI.

> ***“ Het gaat om de vraag welke beslissingen AI zou móeten maken, niet om welke beslissingen het kán maken. ”***

Foto: Rene Opstals Photography

Dr. Katleen Gabriels

Moraalfilosofe, gespecialiseerd in computerethiek, Universitair docent (Vakgroep Wijsbegeerte), Universiteit van Maastricht

Gaat het om AI, dan moeten we ons altijd afvragen waarvoor we AI willen inzetten. We moeten heel goed nadenken over waar we AI wel en niet voor gebruiken. Wanneer is de inzet wenselijk en wanneer problematisch?


Foto: Hugo Thomassen Photography

> *"De toeslagenaffaire maakte de valkuilen erg duidelijk. Hier overschreed de overheid niet enkel ethische, maar ook juridische grenzen."*

AI helpt al bij het bepalen van een diagnose in een medische context (conform medische oordeelsvorming), zoals bij het diagnosticeren van borst- of huidkanker. Als AI het systematisch beter doet dan artsen, dan zou het in de toekomst zelfs als onethisch kunnen worden beschouwd als we AI hier niet voor zouden inzetten.

Bij medische oordeelsvorming worden algoritmen getraind met foto's, bijvoorbeeld van melanomen (huidkanker). Die zijn complex, maar nog lang niet zo complex als sociale data. Een goed uitgebalanceerde dataset is essentieel en bovendien mogen er geen (sociale) vooroordelen in sluipen.

De toeslagenaffaire maakte de valkuilen erg duidelijk: de algoritmen selecteerden onder meer op dubbele nationaliteiten en 'exotische' namen. Hier overschreed de overheid niet enkel ethische, maar ook juridische grenzen. We hebben onafhankelijke toezichthouders nodig om AI op de juiste manier te gebruiken en om dat gebruik vervolgens te controleren.

Het is belangrijk om op verschillende niveaus de besluit- en oordeelsvorming door algoritmen te analyseren. Technisch gezien is het mogelijk en haalbaar. Maar in hoeverre is het ook ethisch wenselijk? Beleidsmakers focussen graag op het economische niveau: vaak is het al voldoende als het tijd- en kostenbesparend is. Dit vanuit een focus op efficiëntie en optimalisatie. Maar ook hier kunnen ze voorbijgaan aan de ethische wenselijkheid.

Uiteraard is het juridische aspect ook essentieel; bij de toeslagenaffaire werd de AVG overtreden. En dan is er ook nog de filosofische vraagstelling. Bijvoorbeeld welk mens- en wereldbeeld zit achter die drang tot optimalisering? En moet alles geoptimaliseerd worden? Welke nieuwe problemen kunnen zo ontstaan?

> ***"Om de winstcijfers de hoogte in te jagen, pleegde Volkswagen op massale schaal bedrog."***

AI die slimme, geconnecteerde apparaten scheppen mogelijkheden, maar ook valkuilen. 'Dieselgate' maakte duidelijk hoe je slimme technologieën zo kunt programmeren dat je ze voor onethische doeleinden kunt inzetten. De sjoemelsoftware van Volkswagen bepaalde zelfstandig wanneer het een test betrof. Zo kon men de emissiewaarden manipuleren, waardoor de auto zuiniger leek dan hij was. Om de winstcijfers de hoogte in te jagen, pleegde Volkswagen op massale schaal bedrog. Met technologieën die steeds slimmer worden, moeten we nog meer op onze hoede zijn dat er niet mee wordt geknoeid.

Pieter van Geel

Director Data, Greenhouse (GroupM / WPP)

Algoritmen zijn bij uitstek geschikt om discriminatie en ongelijkheid te detecteren of te herkennen. Ik zie hier momenteel alleen nog niet veel toepassingen van. Eerder andersom.

> *“ De metriek waarop in een algoritme wordt gestuurd zou wel iets ‘maatschappelijker’ mogen in plaats van business. ”*

Doordat de data waarop een algoritme getraind wordt vaak biased zijn, zien we vaak discriminatie/ongelijkheid ontstaan door een algoritme. Daarnaast zou ook de metriek waarop in een algoritme wordt gestuurd, wel iets ‘maatschappelijker’ mogen (dus inderdaad bestrijding van discriminatie) in plaats van business (meer winst/meer clicks/langere exposure time).

Wat mij betreft kan dat dus alleen als de data unbiased zijn en de sturingsmetriek een maatschappelijk doel dient (bijvoorbeeld inclusief handelen) en niet winst van bedrijven nastreeft. Unbiased data zijn uiteraard niet altijd beschikbaar en daar moet dan dus op gecorrigeerd worden. Hier moet dan wel toezicht en controle op worden gehouden. Benieuwd welke overheid of instantie deze taak op zich neemt.

Oumaima Hajri

MSt AI Ethics & Society, MSc Data Science & Society

Discriminatie en profilering liggen op de loer bij het gebruik van algoritmen en al helemaal als het zelflerende algoritmen zijn. Er zijn tot op heden nog steeds geen duidelijke kaders over wat landelijke overheidsinstanties (of andere partijen) doen om discriminatie en profilering bij beslissingen te voorkomen.

Hier komt nog eens bij dat er geen controle is op het gebruik van algoritmen (dit zou gereguleerd moeten worden) en burgers niet voldoende inzicht hebben in het gebruik hiervan. Dit zijn toch wel minimale eisen die gesteld moeten worden, alvorens er überhaupt wordt gesproken over het inzetten van algoritmen voor het algemene belang.

Het is belangrijk om eerst twee concepten te benoemen die de kern van deze vraag beantwoorden. Ten eerste is het belangrijk dat algoritmen uitlegbaar zijn. Dat betekent dat de technische structuur en werking van een algoritme niet alleen door een programmeur zelf worden begrepen, maar dat de programmeurs deze ook aan mensen kunnen uitleggen. Wanneer dit niet het geval is, resulteert het vaak in het feit dat algoritmen als ‘black boxes’ worden gezien.

Ten tweede is het ook ontzettend belangrijk dat een algoritme interpreteerbaar is. Dat wil zeggen: een verklaring kunnen hebben, maar ook kunnen begrijpen hoe bepaalde inputvariabelen bijdragen aan een algoritme-output die uitgelegd moet worden.

> *“Net als dat wij geen willekeurige beslissingen accepteren van mensen of entiteiten die wij niet begrijpen, zouden we dit ook niet moeten accepteren van algoritmen.”*

Als we het hebben over beslissingen die op een inclusieve manier worden genomen, dan denken we vaak aan de mensen achter de knoppen: welke belangen hebben zij en zijn zij wel een weerspiegeling zijn van onze maatschappij? Echter, het is belangrijk om in deze discussie eerst een stap naar achter te nemen en te focussen op bovengenoemde concepten. Want in hoeverre hebben wij iets aan ‘inclusieve beslissingen’ omdat de juiste mensen achter de knoppen zitten, als we uiteindelijk de redenering achter de beslissingen niet kunnen begrijpen?

Het hebben van een inclusief team achter de knoppen tackelt alleen één deel van het probleem. Het tweede deel zou namelijk moeten zijn dat algoritmen zowel uitlegbaar als interpreteerbaar zijn. Want net als dat wij geen willekeurige beslissingen accepteren van mensen of entiteiten die wij niet begrijpen, zouden we dit ook niet moeten accepteren van algoritmen.

De afgelopen jaren is Explainable AI (XAI) een belangrijk veld geworden waarin onderzoekers pleiten voor uitlegbaarheid en interpreteerbaarheid, zodat het altijd te verklaren én te begrijpen is hoe algoritmen tot een bepaalde uitkomst komen. Echter blijkt het nog steeds lastig voor algoritmen om context te geven waarom ze tot een bepaald besluit zijn gekomen. En dit kan, zeker bij serieuzere voorbeelden zoals de toeslagenaffaire, absoluut van belang zijn.

Tot de tijd dat algoritmen niet voldoen aan bovenstaande eisen, zouden ze niet ingezet moeten worden voor grote beslissingen die significante impact kunnen hebben op burgers. Zo simpel is dat.

Er zijn nog steeds geen duidelijke kaders over wat landelijke overheidsinstanties doen om discriminatie en profilering bij beslissingen te voorkomen.

Dr. Marcel Heerink

Researcher, trainer en auteur in ethiek, sociale robots, AI

Stel, je hebt als leidinggevende een applicatie tot je beschikking die bepaalt wie een loonsverhoging moet krijgen en wie beter ontslagen kan worden. Je wilt dat baseren op prestaties en potentieel. Goede mensen wil je niet alleen belonen, maar ook vasthouden.

> ***“ Bij loonsverhoging rollen er werknemers uit die weinig beperkingen hebben, bij ontslag rollen er juist mensen uit die wél beperkingen hebben. ”***

Daarom is de applicatie zo geprogrammeerd dat ze cijfers heeft over ieders prestaties en indicaties van potentiële groei (we gaan er even van uit dat beide goed gemeten kunnen worden). Ze krijgt geen informatie over iemands beperkingen, geslacht, etniciteit of leeftijd. En stel, wat loonsverhoging betreft rollen er werknemers uit die weinig beperkingen hebben, maar gaat het om ontslag, dan rollen er juist mensen uit die wél beperkingen hebben.

Dan zou je het volgende kunnen doen:

- De uitkomsten volgen en op grond daarvan mensen ontslaan of loonsverhoging geven;
- De uitkomsten negeren en gewoon je eigen regeltjes volgen;
- De applicatie zo aan laten passen dat beperkingen meegenomen worden en leiden tot een hogere score;
- Kijken of je een betere applicatie kunt krijgen die (bijvoorbeeld met simulaties) uitzoekt op welke plek of met welke aanpak de mensen met lage scores hoger zouden kunnen scoren.

Dat laatste lijkt me het mooist. Maar zolang we daar nog geen goede applicaties voor hebben, moeten we dat dan maar handmatig doen.

Ir. Reinoud Kaasschieter

Eur.Ing. Artificial Intelligence en ECM Consultant,
Insights & Data Capgemini Nederland

Het gebrek aan inclusiviteit is een maatschappelijk probleem, algoritmen kunnen slechts een hulpmiddel zijn om dit de wereld uit te helpen. De grootste uitdaging zit hem in de data waarmee we de modellen voeren. Het maakt niet uit of deze modellen wel of niet intelligent zijn. De data die we verzamelen over mensen, bijvoorbeeld hun digitale 'footprint', kunnen worden gebruikt om discriminerende systemen te maken.

Het is niet zo dat modellen vanzelf gaan discrimineren: fout ontworpen modellen gaan dat doen. En ontwerpen is een menselijke activiteit. Dat niet alle systemen bewust ontworpen zijn om te discrimineren is natuurlijk waar. Maar het is de verantwoordelijkheid van de ontwerper om na te denken over alle mogelijke consequenties van zijn systeem. Het is al een hele opgave systemen en modellen te maken die niet discrimineren, laat staan algoritmen te maken die helpen discriminatie te bestrijden.

> ***"Het is al een hele opgave systemen en modellen te maken die niet discrimineren, laat staan algoritmen te maken die helpen discriminatie te bestrijden."***

Foto: Marnix Klooster

Daar kan tegenin worden gebracht dat mensen ook discrimineren, al dan niet bewust. En dat algoritmen daar misschien geen last van hebben en objectiever kunnen beslissen. Maar telkens weer blijkt het lastig dit soort systemen te bouwen. De meest voorkomende oorzaak is dat deze systemen leren op basis van historische (bedrijfs)data. In deze data kan ongelijkheid ingebakken zitten. Het blijkt heel moeilijk data van deze bias op te schonen.

Bij recruitment- en sollicitatiesystemen streeft men ernaar de impliciete vooroordelen van menselijke personeelsfunctionarissen te vermijden. Deze systemen beloven bijvoorbeeld geen onderscheid te maken naar gender. We zouden dan gender uit de gegevens waarmee we het keuze-algoritme voeden kunnen halen. Maar dat blijkt niet voldoende.

Ben je lid geweest van een vrouwenteam of op zwangerschapsverlof geweest? Het algoritme weet je feilloos als vrouw te typeren. En al lukt het bijvoorbeeld om genderneutraal te worden, gaan deze systemen op andere criteria discrimineren. Zo kan bij sollicitaties gezichtsherkenningssoftware worden gebruikt om te kijken of een sollicitant oprechte antwoorden geeft. Maar deze software werkt dan weer niet bij mensen met autisme of gezichtsverlamming. Mensen kunnen zich bewust worden van hun discriminerende gedrag en zelf hun gedrag daarop aanpassen. Algoritmen doen dat niet vanuit zichzelf.

> ***“ Ben je lid geweest van een vrouwenteam? Het algoritme weet je feilloos als vrouw te typeren. ”***

Algoritmen worden ingezet om processen efficiënter te laten verlopen. Of om veel meer productie te maken. Zonder algoritmen kunnen we geen grote hoeveelheden data meer verwerken. Het probleem ontstaat wanneer we algoritmen beslissingen laten nemen.

Het gebruik van algoritmen wordt pas echt efficiënt wanneer we de mens uit het proces halen. Mensen zijn langzaam en duur. Het idee om iedere computerbeslissing door een mens te laten controleren heeft daarom zijn beperkingen. Het proces wordt te langzaam, te duur en eigenlijk onuitvoerbaar. Mensen de uitkomsten van computerbeslissingen laten controleren op ongelijkheid en discriminatie is volgens mij een menselijk onmogelijke taak. Vanuit bijvoorbeeld ergonomie en kennismanagement. Of het is in ieder geval economisch gezien problematisch. Vergelijk het met het handmatig opsporen van verkeerde content op sociale media.

> ***“ Zolang we niet in staat zijn ongelijkheid en discriminatie in het normale maatschappelijke verkeer weg te krijgen, blijven algoritmen altijd lapmiddelen. ”***

Misschien moeten we stoppen met proberen allerlei maatschappelijke problemen met technologie op te lossen. En daar dan de ultieme oplossing in te zien. Zolang we niet in staat zijn ongelijkheid en discriminatie in het normale maatschappelijke verkeer weg te krijgen, blijven algoritmen altijd lapmiddelen.

Het enige wat we kunnen doen, is zelf proberen inclusief te handelen en te denken. En deze waarden niet alleen in ons dagelijks gedrag uit te dragen, maar ook in de systemen die we maken. Daar zijn allerlei technieken en organisatievormen voor ontwikkeld. Daarbij moeten de waarden ‘gelijkheid’ en

'non-discriminatie' wel bovenaan staan. En moeten we geen systemen gebruiken die deze waarden niet honoreren. We moeten veel meer empathie hebben voor de slachtoffers van de onjuiste beslissingen van onze algoritmen. Neem de toeslagenaffaire maar als voorbeeld.

Ik constateer dat 'foute' AI, als in onethische, discriminerende systemen, aandacht krijgt in academische kringen. Ook aan universiteiten wordt onethische AI gemaakt en onderzocht. Maar de discussie wordt daar openlijk en scherp gevoerd. Daarbij gaat het vooral om de vraag of AI bijdraagt aan het gemeenschappelijke goed ('common good').

Binnen het bedrijfsleven wordt natuurlijk ook met AI gewerkt en geëxperimenteerd. Maar daarover wordt niet gepubliceerd en foute en discriminerende AI wordt hier eigenlijk nog 'per ongeluk' ontdekt. Vaak door buitenstaanders. Dus eigenlijk hebben we als buitenstaanders geen goed inzicht in wat er allemaal gebeurt bij bedrijven. Je moet goed begrijpen dat AI in het bedrijfsleven wordt ingezet om organisaties winstgevender te maken. Dat is logisch. Waar de universiteiten zich concentreren op niet-discriminerende, inclusieve AI, zal het bedrijfsleven zich dus vooral concentreren op winstgevende AI. Deze twee uitgangspunten hoeven elkaar natuurlijk niet te bijten, maar er kunnen wel waardeconflicten ontstaan.

Jo-An Kamp

Docentonderzoeker bij het lectoraat Moral Design Strategy, Fontys Hogeschool

We moeten ons eerst afvragen wat algoritmen nu eigenlijk zijn. De meest eenvoudige definitie luidt 'een eindige reeks instructies (uitgevoerd door computers) die vanuit een gegeven begintoestand naar een beoogd doel leidt'. Het feit dat deze reeks berekeningen door computers wordt uitgevoerd, suggereert dat computers ook de actoren zijn die de beslissingen nemen. Maar dat is niet altijd het geval en ook niet het hele verhaal.

> *"Als we slechte data in het systeem gooien, dan komen er ook slechte data uit."*

Ten eerste zijn algoritmen meestal door mensen geschreven en ten tweede beslissen mensen welke data in het systeem worden ingevoerd en dus ook uit de berekeningen naar voren zullen komen. Als we algoritmen inzetten voor het algemene belang, bijvoorbeeld voor een overheidsinstantie, dan is het dus belangrijk dat we goed nadenken over zowel de rekenregels als over de data die we in het systeem invoeren. Als we er slechte data ingooien, dan komen er namelijk ook slechte data uit.

Algoritmen zijn niet per definitie eerlijker of beter dan mensen. Wel kunnen ze, door de rekenregels die er achter liggen, patronen blootleggen die we in specifieke, losgeknipte gevallen wellicht over het hoofd zouden zien. Zo is het bijvoorbeeld een bekend verschijnsel dat blanke mannen in een sollicitatieprocedure met een panel van blanke mannen meer kans hebben om aangenomen te worden dan mannen met een andere afkomst of dan vrouwen. We kiezen namelijk snel voor mensen die op ons lijken. Of hebben een onbewuste voorkeur voor kandidaten die lijken op mensen die in het verleden ook goed gepresteerd hebben binnen een bepaalde functie. Daardoor zien we potentieel talent dat niet op ons lijkt vaker over het hoofd.

Algoritmen kunnen ons ervan bewust maken dat deze onbewuste vooroordelen er zijn, waardoor we de regels en de procedures (zowel in de computer als in het echt!) daarop kunnen aanpassen. Op deze manier zouden we discriminatie en ongelijkheid dus actief kunnen bestrijden. Maar dan moeten we het wél op de juiste manier doen!

Ik denk dat we ons ten eerste af moeten vragen hoe inclusief het team is dat de algoritmen maakt. Hoe eenzijdiger dit team, hoe lastiger het is om je te verplaatsen in de wellicht veel bredere groep waarvoor het algoritme is bedoeld.

Maak je team dus inclusiever en/of test je product in die bredere doelgroep, met mensen van verschillende leeftijdscategorieën, huidskleuren, politieke voorkeuren, enzovoort. En stel jezelf de volgende vragen: heeft iedereen toegang tot je product (is iedereen bijvoorbeeld in het bezit van een computer of smartphone en kan iedereen hier even goed mee omgaan)? Sluit je (al dan niet bewust) mensen uit? Heeft jouw technologie een ingebouwde bias? Kun je de beslissingen van je algoritmen transparant maken? En is de uitslag na de 780e stap nog wel uit te leggen?

In de Technology Impact Cycle Tool staan nog een heleboel van dit soort vragen die je kunnen helpen om je technologie beter aan te laten sluiten op de menselijke maat en de wereld waarin je zelf zou willen wonen. Probeer het gerust een keer uit op www.tict.io. Het is gratis en voor iedereen.

> ***"Hoe eenzijdiger het team dat algoritmen maakt, hoe lastiger het zich kan verplaatsen in de veel bredere groep waarvoor het algoritme is bedoeld."***

Yori Kamphuis

AI-expert en spreker

In de vraag of algoritmen kunnen worden ingezet voor het algemeen belang, zitten bepaalde waarden omsloten. Een AI systeem streeft een doel na, iets dat kan worden gemaximaliseerd of geoptimaliseerd. Hoe wordt dat algemene belang gedefinieerd en gemeten?

> *“ Het is belangrijk dat voorspellingen geen selffulfilling prophecies worden. ”*

AI is niet automatisch neutraal, want het streeft een bepaald doel na. Daarbij is het bovendien afhankelijk van data. Als die data een bias hebben, dat wil zeggen gekleurd zijn of vooroordelen bevatten, zal de AI diezelfde bias waarschijnlijk in stand houden. Ook is het belangrijk dat voorspellingen geen selffulfilling prophecies worden.

Stel dat slechts naar één groep wordt gekeken of een zekere verkeersovertreding wordt gemaakt. Dan is het een foute conclusie om te stellen dat alleen mensen van deze groep dit soort overtredingen maakt. Want er is niet gekeken of iemand uit een andere groep ook dat soort overtreding maakt. Maar als AI is getraind en door de foute conclusie alleen die ene groep controleert, kan het discriminatie of ongelijkheid in de hand werken.

> *“ Als het AI-systeem een andere aanbeveling doet dan jij had verwacht, wordt het interessant. ”*

Belangrijk om te onthouden is dat onderscheid mág worden gemaakt maken op basis van bepaalde kenmerken. Zolang deze kenmerken er toe doen is dit geen discriminatie. Een voorbeeld van het College voor de Rechten van de Mens stelt dat iemand die zonder rijbewijs solliciteert als chauffeur geweigerd mag worden. Zo kan specifiek een “gekleurde/zwarte acteur [worden gezocht om de rol van] Martin Luther King [te] spelen”.
Zie https://mensenrechten.nl/nl/gelijkebehandelingswetgeving en https://mensenrechten.nl/nl/discriminatie-uitgelegd.

Je zou mogelijk een dataset kunnen trainen waarin je telkens een (potentieel discriminerend) element weglaat, en dat model opnieuw trainen, om zo te kijken

of de voorspellende waarde van het model verbetert. Verbetert het, dan had dit element geen toegevoegde waarde en zou je het wellicht niet meer moeten verzamelen. Maar ook dit gaat mogelijke discriminatie niet uitbannen.

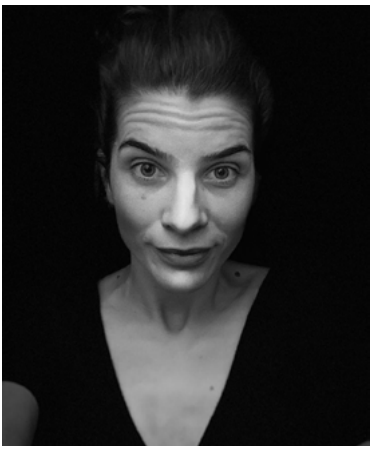
Door AI als ondersteuning voor menselijke beslissers in te zetten, denk ik dat we mensgerichte AI mogelijk kunnen maken. Daarbij moet er aandacht zijn voor monitoring - doet het AI systeem wat het moet doen? Als het AI-systeem een andere aanbeveling doet dan jij had verwacht, wordt het interessant.

Explainable AI moet inzicht geven in waarom het een bepaalde aanbeveling doet. Hieraan zit een zekere mate van (interne) transparantie gekoppeld. Ofwel: kun je snappen waarom het systeem met een bepaalde aanbeveling komt? Stel dat je nog steeds van mening bent dat het AI-systeem tot een andere beslissing had moeten komen, hoe komt het dat dit zo is? Met andere woorden: waarin schiet het AI-systeem tekort? Zo kan het systeem verder worden getraind, om de menselijke kant in de gaten te blijven houden.

Traditioneel werden in de medische wereld de meeste behandelingen en medicijnen op mannen getest, niet op vrouwen. Op basis hiervan kregen ook vrouwen echter een aanbeveling. Elke niet-representatieve groep als modelgroep gebruiken zorgt daarmee voor minder inclusiviteit. Ik zou dus met name richten op de verbetering van (het verzamelen van) relevante data op basis waarvan een AI wordt getraind.

Julia Keseru

Technology and justice activist, ED of the Engine Room



> *"Veel van wat vandaag de dag als AI wordt verkocht, is eigenlijk gewoon slangenolie."*

Hoe meer onderzoek ik zie, hoe meer ik ervan overtuigd raak dat algoritmen, hoewel ze in sommige gevallen nuttig kunnen zijn om de efficiëntie te verhogen, echt verschrikkelijk zijn om besluitvormingsprocessen rechtvaardig, eerlijk en genuanceerd te ondersteunen.

Het grootste probleem is niet dat geautomatiseerde besluitvorming nog niet verfijnd of geavanceerd genoeg is. Hoewel het belangrijk is om op te merken dat veel van wat vandaag de dag als AI wordt verkocht, eigenlijk gewoon slangenolie is. Het belangrijkste probleem is dat de gegevens die algoritmen gebruiken, vaak al zeer problematisch zijn.

Momenteel is er in elke sector een trend naar evidence-based besluitvorming waarbij kwantificeerbare markers voorrang krijgen op kwalitatieve en contextuele analyse, een lovenswaardige trend. Vertrouwen op kwantitatieve gegevens zonder context kan echter meer problemen veroorzaken dan oplossen - systematisch racisme wordt bijvoorbeeld voortdurend gereproduceerd door het feit dat minderheden oververtegenwoordigd zijn in de gegevens die worden gebruikt voor voorspellend politiewerk of in kinderwelzijnssystemen.

Toch is het niet onmogelijk om algoritmische besluitvorming voor goede doelen te gebruiken. We hebben gezien dat geautomatiseerde besluitvorming helpt om de gevolgen van klimaatverandering in de landbouw te verzachten, corruptie bij overheidsopdrachten aan het licht te brengen, de resultaten van patiënten te verbeteren, schendingen van de mensenrechten te monitoren en noodsituaties te voorspellen.

> ***"We hebben strengere regels en stimulansen nodig voor de technologie-industrie om te voldoen aan de normen voor mensenrechten."***

Omdat het om een systemische kwestie gaat, moet een groot aantal dingen gebeuren om de huidige koers te veranderen van de manier waarop AI wordt ontworpen, geïmplementeerd, gerepliceerd, enzovoorts om ervoor te zorgen dat AI inclusievere beslissingen voor ons maakt. We hebben strengere regels en stimulansen nodig voor de technologie-industrie om te voldoen aan de normen voor mensenrechten.

We hebben meer diverse teams nodig achter het ontwerp van technische hulpmiddelen die algoritmen gebruiken. We hebben veel meer transparantie en verantwoording nodig rond geautomatiseerde besluitvormingsprocessen in zowel de publieke als de private sector (zie deze primer van Data and Society: https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf). We hebben meer data equity-trainingen nodig zoals die van We All Count, en nog veel meer. Om een goed idee te krijgen van wat er nog meer nodig is, raad ik dit rapport van de Mozilla Foundation aan: https://assets.mofoprod.net/network/documents/Mozilla-Trustworthy_AI.pdf

Gary Marcus

Oprichter en CEO, Robust.AI Professor Emeritus, New York University.
Nieuw boek: REBOOTING AI

Toekomstige, nog uit te vinden algoritmen kunnen misschien ooit helpen bij rechtvaardigheid en ongelijkheid, maar de huidige algoritmen zijn daar veel te primitief voor. Ze recapituleren slechts het verleden, en hebben geen begrip van de toekomst die we proberen te bereiken. Ze hebben geen waarden en normen en geen begrip van de complexe dynamiek van de mensheid. Zolang zij niet veel meer zijn dan blinde getallenkrakers, zijn zij ons vertrouwen niet waard.

> *"De huidige algoritmen recapituleren slechts het verleden, en hebben geen begrip van de toekomst die we proberen te bereiken."*

Foto: Athena Vouloumanos

Iris Muis

MA Project coördinator, DEDA-lead Utrecht Data School, Universiteit Utrecht

Algoritmen, waaronder AI, bieden de maatschappij enorm veel kansen. We hebben de mogelijkheid om onze processen effectiever, efficiënter en soms zelfs op een totaal andere manier in te richten. Tegelijkertijd bieden deze technologieën geen oplossing voor alles en kunnen algoritmen, net als mensen, vooroordelen bevatten en bepaalde waarden boven andere plaatsen.

AI kan waarden transporteren, versterken of zelfs aantasten. Het is en blijft onze taak om erop toe te zien dat AI de waarden transporteert die wij willen overdragen. Ik ben voorstander van een verantwoorde ontwikkeling en inzet van algoritmen, waarbij bewust wordt gereflecteerd op de impliciete aannames die besloten liggen in de code en de manier waarop het algoritme geïmplementeerd wordt. Dit geeft ons de kans om de algoritmen bij te laten dragen aan een maatschappij waar wij graag in zouden willen leven.

Inclusiviteit is een belangrijke waarde die je mee zou moeten nemen in het design van een algoritme. Al bij de ontwikkelingsfase van een algoritme worden bepaalde keuzes gemaakt, door de data scientist (of misschien diens manager), die verankerd worden in het design en die consequenties hebben wanneer het algoritme daadwerkelijk wordt ingezet. Wanneer je aan de voorkant nadenkt over waarden, bijvoorbeeld inclusiviteit, geeft dat je de kans om deze te verankeren in het design van het algoritme.

> *“Een verantwoorde ontwikkeling en inzet van algoritmen geeft ons de kans om ze bij te laten dragen aan een maatschappij waar wij graag in zouden willen leven.”*

Thijs Pepping

Humanisticus & Trendanalist ViNT, Sogeti

> *"Algoritmen kwantificeren onze onbewuste vooroordelen en wrijven deze in ons gezicht."*

In de eerste plaats moeten we ons realiseren dat bias in algoritmen in veel gevallen een afspiegeling is van de bias in mensen. Neem bijvoorbeeld de AI-tool die Amazon ontwikkelde om te helpen in het wervingsproces. Hartstikke handig volgens sommigen: je geeft het programma honderd cv's en de beste vijf worden geselecteerd. Amazon moest de stekker uit het project trekken omdat de tool een voorkeur voor cv's van mannen had.

Dat is natuurlijk kwalijk. Maar de situatie is net zo kwalijk als daarna alles bij het oude blijft en er geen reflectie is op waar die vooroordelen in het algoritme vandaan kwamen. De tool is namelijk getraind op tien jaar wervingsdata waarbij mensen beslissingen maakten over wie werd aangenomen en waarbij het grootste deel van de sollicitanten man was. Hoe komt dat? Wat kunnen we daaraan veranderen?

In die zin zie ik vooroordelen in algoritmen als een zegen: ze tonen het menselijk tekort. Algoritmen kwantificeren onze onbewuste vooroordelen en wrijven deze in ons gezicht. We moeten uiteraard de bias in algoritmen bevechten, zeker

omdat met de schaalvergroting van technologie ongelijkheid in een algoritme al snel grote gevolgen voor grote groepen mensen heeft. Maar we moeten ook vooral de bias in onszelf aankaarten. De spiegel die algoritmen ons voorhouden kan hier een goed vertrekpunt voor zijn.

> ***“ Het werken aan inclusieve algoritmen is geen objectieve wetenschap, maar waarde- en politiek geladen. ”***

Het wordt nu echt tijd voor een Minister van Innovatie & Technologie om dit punt goed op de agenda te zetten en te reguleren. Hoewel het al iets beter is dan in 2017, mis ik in 2021 nog steeds uitgedachte visies en stellingnames in partijprogramma's als het gaat om AI en technologie. De geavanceerde en allesdoordringende technologie om ons heen bevraagt ons over onze diepste normen en waarden. Wat vind je bijvoorbeeld van een liefdesrelatie tussen een mens en een chatbot? Of van het virtueel tot leven brengen van overleden dierbaren om daar nog een gesprek mee te voeren? Hoe ver mogen algoritmen ingrijpen in onze maatschappij? En wie maakt en beoordeelt die algoritmen?

Uiteindelijk kom je bij het beantwoorden van die vragen uit bij je eigen wereldbeeld, politieke voorkeur en levensbeschouwing. Plat gezegd: geloof je bijvoorbeeld in een ziel, dan keur je een relatie met een chatbot wellicht af, omdat die chatbot geen ziel heeft. Ben je atheïst dan oordeel je weer anders. Het werken aan inclusieve algoritmen is dus geen objectieve wetenschap. Het is waarde- en politiek geladen. Werkvormen voor Ethical AI, Explainable AI, Transparant AI, Inclusive AI et cetera zijn belangrijk en nuttig, maar laten we vooral ook de enorme impact van technologie op landelijk niveau serieuzer nemen.

Gerard Schouten

Lector AI & Big Data Fontys University of Applied Sciences

Ik ben ervan overtuigd dat we - in the long run - slimme algoritmen zeker kunnen inzetten voor het algemeen belang. Er zijn echter wel zaken die we nog niet goed ‘onder de knie’ hebben en nog moeten leren. Maar daar wordt wel heel hard aan gewerkt.

Foto: Ektor Tsolodimos

> *“Het werken aan inclusieve algoritmen is geen objectieve Bij een AI-model dat die doet met sociaal maatschappelijke aspecten moet letterlijk ‘the human in the loop’ toegepast worden.”*

1. Hoe om te gaan met bias in data waarmee we deze algoritmen trainen? Er wordt zelfs gekeken om bijvoorbeeld demografische bias (leeftijd, gender) automatisch te meten in datasets en deze met AI (adverserial neural networks) te corrigeren. In 2020 verscheen een researchpaper getiteld 'Risk of Training Diagnostic Algorithms on Data with Demographic bias'. Zie: https://www.researchgate.net/publication/344657158_Risk_of_Training_Diagnostic_Algorithms_on_Data_with_Demographic_Bias.

2. Naar mijn (sterke) mening moeten we bij AI die iets zegt of doet met sociaal maatschappelijke aspecten (toegang tot leningen of uitkeringen, toegang tot onderwijs, et cetera) altijd zorgen dat een AI-model met letterlijk 'the human in the loop' gebouwd/toegepast wordt. Dit heeft de toeslagenaffaire ons wel geleerd.

3. Het privacyprobleem is ook nog niet goed opgelost. Met technieken als edge computing (bijvoorbeeld op het IoT device zelf) kun je echter al wel grote stappen maken om geen persoonsgegevens of hieraan gerelateerde data in 'the Cloud' te zetten.

Fredo De Smet

Curator en digitaal humanist

We moeten ons ervan bewust zijn dat AI met een verlies aan controle gepaard gaat. In mijn boek 'Artificial Stupidity' beschrijf ik uitvoerig hoe we in een controlecrisis leven. Dat is niet nieuw. We hebben dat al eerder meegemaakt in de geschiedenis.

> ***"We hebben meer kracht om informatie te creëren dan om deze te controleren."***

In de negentiende eeuw hadden we meer kracht om materialen en energie te produceren dan om deze te controleren. We stortten ons in een industriële revolutie die ons de moderne wereld in hielp. Spierkracht werd vervangen door industriële kracht. Onze armen en benen werden als het ware verlengd door machines van ijzer en stoom. We overwonnen de natuur. De mens liet duizenden jaren agrarisch leven achter zich en trok de stad in. Gepaard hiermee kwamen situaties van onrecht, slechte werksituaties in fabrieken, kinderen in arbeid, dodelijke ongevallen. Drama. Het politieke en sociale weefsel werd door elkaar geschud.

Vandaag leven we in een gelijkaardige controlecrisis. In dit geval gaat het echter niet over materie maar over informatie. We hebben meer kracht om informatie te creëren dan om deze te controleren. Er zijn gewoon te veel data om door een mens te laten lezen. Enter: AI. Een geautomatiseerd proces om deze overvloed aan data te beheersen. De ironie van de zaak is natuurlijk dat AI van nature niet te controleren is. Het is een zelflerend, zelfsturend algoritme. We zullen, kortom, altijd te laat zijn.

Dat is natuurlijk geen antwoord op je vraag. Maar ik geef het mee omdat we realistisch moeten zijn. Een verlies aan controle is inherent aan AI.

Dat merk je ook aan de organisaties en lichamen die voor ethische AI pleiten. Tal van rapporten pleiten voor een ethische manier om met AI om te gaan. The Future of Life introduceerde in 2017 een set van regels (Asilomar principles). De EU was de voorbije jaren druk in de weer. De wensen die worden uitgedrukt blijven steeds vaag. Het komt in essentie neer op transparantie, uitlegbaarheid en accountability.

Je voelt al dat het moeilijk is om een organisatie accountable te maken als de technologie in essentie niet transparant en uitlegbaar is. Kortom, we zijn nog niet thuis. Toch geven deze drie aspecten een idee waar het naartoe moet.

Zo zouden we bijvoorbeeld een Ethical AI-label kunnen ontwikkelen, zoals dat ook voor duurzaam voedsel bestaat. Dit label krijgt een organisatie alleen als het tenminste een poging doet uit te leggen met welke data het werkt en wat het met die data doet (uitlegbaarheid).

Daarnaast zou het goed zijn als onderzoeksinstellingen de toegang kunnen opvragen naar de AI. Een grote frustratie van Stanford is dat ze de ingenieurs van Silicon Valley opleiden, maar geen inzicht krijgen in de algoritmen waarmee deze ingenieurs enkele kilometers verderop werken. Het is wederom geen quick

fix, maar ik vind dat de slimste koppen eveneens toegang moeten krijgen tot de meest complexe algoritmen, niet enkel de zelflerende (transparantie).

> ***“ De scheermesjes die ooit elektrisch werden, zijn binnenkort ‘powered by AI’. ”***

Ik geloof ook dat de race voor AI een opportuniteit is om andere vormen van ongelijkheid aan te pakken. Slimme producten worden binnenkort een differentiator voor organisaties. De meest gekke producten zullen gecognificieerd worden. Denk aan de scheermesjes die ooit elektrisch werden (dankzij Philips). Ook die zijn binnenkort ‘powered by AI’.

Als deze bedrijven het Ethisch-AI label willen, zouden ze ook aan enkele andere voorwaarden moeten voldoen. Het is een opportuniteit om meer inclusie te organiseren, door bijvoorbeeld diversiteitsquota op te leggen in de tech-teams (vier van de vijf AI-experts zijn man). De mens centraal zetten betekent ook de hegemonie van de man doorbreken.

Om het echt humanistisch te maken, zouden we een waardensysteem aan het label kunnen koppelen. Productontwikkeling zou purpose-driven moeten zijn. Zo kunnen we meer inzetten op designprocessen die niet enkel optimaliseren voor de eindconsument, maar tevens voor de gemeenschap.

Jim Stolze

Tech-ondernemer, oprichter Aigency

Strikt genomen is een algoritme binnen de computerwetenschap niets anders dan een stapsgewijze instructie om een taak uit te voeren. En zo zijn er binnen de overheid genoeg IT-systemen die trachten op een eerlijke en rechtvaardige wijze beslissingen van ambtenaren te ondersteunen.

Waarschijnlijk doel je op datagedreven AI-modellen. Mijn bedrijf Aigency heeft recentelijk voor het ministerie VWS dergelijke modellen ontworpen én gevalideerd op basis van de principes achter FACT (Fairness Accuracy Confidentiality & Transparency). Daarmee konden wij statistisch aantonen of bepaalde bevolkingsgroepen waren oververtegenwoordigd in de trainingsdata en/of zij daarvan in de echte wereld voor- of nadeel zouden ondervinden. Dit zou eigenlijk bij elk dataproject verplicht moeten worden. Er moéten dergelijke checks gedaan worden op de data voordat je überhaupt het woord AI durft uit te spreken.

> ***“We hebben mondige, kritische consumenten nodig die zelf opkomen voor hun rechten in het digitale domein.”***

Hierbij maak ik onderscheid tussen top-down en bottom-up. Top-down is alles wat al is vastgelegd in de wet (onder meer GDPR) of waarop wordt gehandhaafd door toezichthouders als de Autoriteit Persoonsgegevens. Maar let op: dit is slechts het topje van de ijsberg. De meeste datagedreven modellen worden niet gecontroleerd of nemen de wet letterlijk, terwijl daar ethisch wel het een en ander op valt af te dingen.

Daarom ben ik meer geïnteresseerd in de bottom-up aanpak. We hebben mondige burgers nodig, kritische consumenten die zelf opkomen voor hun rechten in het digitale domein. Dat is onder meer waarom ik destijds de Nationale AI-Cursus ben gestart. We kunnen wel allemaal naar de overheid wijzen en roepen om regulering, maar dat ontslaat ons niet van de plicht om vooral zélf invulling te geven aan diversiteit, inclusiviteit en rechtvaardigheid. If not us, who else?

Janienke Sturm

Lector Mens en Technologie Fontys HRM en Psychologie

Ik ben ervan overtuigd dat AI-algoritmen ingezet kunnen worden om discriminatie en ongelijkheid te verminderen, zie bijvoorbeeld: https://www.volkskrant.nl/nieuws-achtergrond/bij-unilever-is-de-helft-van-alle-managers-vrouw-dankzij-kunstmatige-intelligentie~b8535708/

Echter zijn er twee kanttekeningen:

- 1. Algoritmen lijken neutraal, maar dat zijn ze eigenlijk niet. Er zal altijd bewust of onbewust bias zitten in de set met data waarop de algoritmen getraind worden. Vooroordelen worden dan in de dataset gerepliceerd, en in het slechtste geval leidt deze dataset juist tot meer discriminatie en ongelijkheid, bijvoorbeeld op basis van gender, etniciteit, seksuele voorkeur of postcode. Zie voor goede voorbeelden: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2020/09/14/onvoorziene-effecten-van-zelflerende-algoritmen/Onvoorziene+effecten+van+zelflerende+algoritmen.pdf

- 2. Algoritmen kunnen positief ingezet worden door mensen met goede intenties. Maar dezelfde algoritmen kunnen door mensen met slechte intenties of tegenstrijdige belangen, ook ingezet worden voor doeleinden die minder positief, minder ethisch et cetera zijn. Zie het voorbeeld in dit artikel: https://www.computable.nl/artikel/nieuws/security/7060485/250449/noldus-schendt-via-tech-mensenrechten-in-china.html

Hierin wordt een Nederlands bedrijf dat software maakt voor gedragsonderzoek (onder andere gezichts- en emotieherkenning) ervan beticht mensenrechten te schenden in China, omdat de Chinese overheid de software inzet voor surveillancedoeleinden.

We moeten er ook oog voor hebben dat de inzet van AI in het algemeen juist kan leiden tot meer ongelijkheid. Een behoorlijk grote groep mensen heeft niet de juiste apparatuur om toegang te hebben tot AI-toepassingen, of niet de juiste vaardigheden om AI-toepassingen te gebruiken.

Een nog veel grotere groep heeft onvoldoende inzicht in de werking van algoritmen om deze op waarde te kunnen schatten en de risico's die het gebruik van AI met zich meebrengt in te kunnen schatten. Digitale ongeletterdheid zorgt er op deze manier voor dat niet iedereen in gelijke mate kan profiteren van de kansen die AI biedt. Daarmee leidt de inzet van AI, en van technologie in het algemeen, juist tot een toenemende kansenongelijkheid in de samenleving, bijvoorbeeld op het gebied van participatie en gezondheid.
https://www.capgemini.com/research/the-great-digital-divide/

"Door digitale ongeletterdheid kan niet iedereen in gelijke mate profiteren van de kansen die AI biedt."

Farid Tabarki

Oprichter Studio Zeitgeist

Veel mensen gebruiken social media voor hun nieuwsgaring; de oplage van kranten daalt intussen gestaag. Grote platforms voor kennis, nieuws en meningen zijn niet automatisch onafhankelijk of transparant: hun algoritmen en beleidskeuzes zijn medebepalend voor onze sociale realiteit. De bedrijven moeten daarnaar handelen.

> ***“ Wat hedendaagse samenlevingen doet werken hangt in toenemende mate af van bits in plaats van atomen. ”***

De grote vraag is hoe je de ethische verantwoordelijkheden verdeelt in de huidige onoverzichtelijke informatiesamenleving. Een samenleving waarin mondiaal opererende bedrijven en individuele bloggers rollen spelen die twintig jaar geleden nog niet bestonden en waarin natiestaten steeds minder in de melk te brokkelen hebben.

Luciano Floridi, professor filosofie en ethiek op de universiteit van Oxford, heeft hier een antwoord op. Volgens hem is het een teken van onze tijd dat, wanneer huidige politici over infrastructuur spreken, ze vaak informatie- en communicatietechnologieën (ICT’s) in gedachten hebben. En dat klopt. Van succes in het bedrijfsleven tot cyberconflicten, wat hedendaagse samenlevingen doet werken hangt in toenemende mate af van bits in plaats van atomen. Afhankelijk van hun digitale infrastructuur kunnen samenlevingen groeien en bloeien.

En de ICT’s van samenlevingen vertegenwoordigen vaak één van hun zwakste kanten op het gebied van cyberveiligheid. We weten dit allemaal. Minder voor de hand liggend en filosofisch interessanter is dat ICT’s ook een nieuw soort vergelijking lijken te hebben onthuld.

Sta eens stil bij de ongekende nadruk die ICT's leggen op cruciale verschijnselen zoals verantwoordingsplicht, intellectuele eigendomsrechten, neutraliteit, openheid, privacy, transparantie en vertrouwen. Deze worden waarschijnlijk beter begrepen in termen van een platform of infrastructuur van sociale normen, verwachtingen en regels die er zijn om het morele of immorele gedrag van de betrokken agenten te vergemakkelijken of te belemmeren.

Door onze informatieve interacties zo belangrijk in de kern van ons leven te plaatsen, legden ICT's iets bloot wat er natuurlijk altijd al was, zij het in het verleden minder zichtbaar: het feit dat moreel gedrag ook een kwestie is van 'ethische infrastructuur', of zoals ik het eenvoudigweg noem: infra-ethiek. Voor de grote informatie- en communicatieplatforms is er dus werk aan de winkel.

"WICT's legden het feit bloot dat moreel gedrag ook een kwestie is van infra-ethiek."

Prof.dr.ir. Bedir Tekinerdogan

Chair Information Technology, Wageningen University & Research

Een algoritme is een eindige reeks instructies met het doel een probleem op te lossen. Vaak zijn voor eenzelfde probleem (bijvoorbeeld sorteren of zoeken in een lijst) meerdere algoritmen mogelijk. Er wordt dan gekeken naar het algoritme dat het snelste is of het minste geheugen vereist. Algoritmen worden ook onderscheiden op basis van hun complexiteit die veelal bepaald wordt door de hoeveelheid tijd en/of geheugenruimte die een algoritme nodig heeft.

Veel problemen kunnen we oplossen door er efficiënte algoritmen voor te ontwikkelen. Echter, sommige problemen kunnen moeilijk opgelost worden en ook bestaan er geen duidelijke algoritmen voor. De vraag of we algoritmen kunnen inzetten om ongelijkheid te bestrijden, is eigenlijk ook niet heel triviaal. We dienen dit verder te herleiden en er een juiste context aan te geven. Het is goed om hierbij een onderscheid te maken tussen de reële wereld en de cyberwereld. De cyberwereld bestaat uit alle computersystemen die ten dienste behoren te staan van de besluitvorming voor de reële wereld.

“ Binnen de context van reële wereld en cyberwereld moeten technologische ontwikkelingen en hun invloed ook een aandachtspunt van de politiek en de regel- en wetgeving worden. ”

Algoritmen kunnen nu ingezet worden om het doen en handelen van de reële wereld te onderzoeken en, daar waar nodig, ongelijkheid te bestrijden. Dit kan technisch geïmplementeerd worden door 'computervision' en kunstmatige intelligentie te gebruiken.

Met behulp van slimme sensoren en bijbehorende softwaresystemen kan het gedrag van mensen geobserveerd worden en, indien nodig, actie ondernomen worden. Algoritmen kunnen ook gebruikt worden voor het analyseren van (social) media op racisme, discriminatie en andere overtredingen van de huidige wetgeving. De vraag is echter waar de grens van het bewaken ligt en in hoeverre de privacyrechten worden geschonden.

Technisch is heel veel mogelijk, maar het blijft een kwestie die een bredere en meer diepgaande discussie vereist. Wat we wel zien, is dat de technologie zich heel snel ontwikkelt en de bestaande politiek en wetgeving eigenlijk hierbij achterblijven. De cyberwereld is een belangrijk onderdeel van de maatschappij. Dit is de virtuele wereld met alle computersystemen die van invloed zijn op de besluitvorming in de reële wereld. Binnen deze nieuwe veranderde context is het daarom van groot belang dat technologische ontwikkelingen en hun invloed ook een aandachtspunt van de politiek en hiermee de regel- en wetgeving worden.

Het lijkt hier vooral te gaan om algoritmen die betrekking hebben op soft-waresystemen van de overheid of particuliere bedrijfsleven. De vraag is dan of deze algoritmen al dan niet eerlijk zijn. Binnen een democratische rechtsstaat zijn alle burgers gelijk en dienen ze in gelijke gevallen gelijk behandeld te worden. Dat is ook vastgelegd in artikel 1 van de Nederlandse Grondwet over gelijke behandeling en discriminatieverbod. Er wordt normaliter veel belang gehecht aan de handhaving van deze wet. Met recht. Met de digitalisering heeft de handhaving van de wet echter een andere dimensie gekregen.

> ***“ Burgers werden in gelijke gevallen ongelijk behandeld door algoritmen die zijn geïmplementeerd door mensen die het gelijkheidsbeginsel zouden moeten respecteren. ”***

Algoritmen zijn doorgaans geïmplementeerd in computerprogramma's die gebruikt worden voor een besluitvorming binnen de overheid of particuliere bedrijven. De in deze computerprogramma's geïmplementeerde besluitvorming is echter veelal minder zichtbaar. Er kan dan helaas te gemakkelijk sprake zijn van een overtreding van het gelijkheidsbeginsel en dus de beginselen van een democratische rechtsstaat.

Dit hebben we ook gezien in de toeslagenaffaire, waarbij een overheidsorgaan bij bepaalde besluiten etnisch profileerde. Burgers werden dus in gelijke gevallen ongelijk behandeld. In de reële wereld zou zo'n handelswijze gemakkelijk opgemerkt worden, maar nu gebeurde het door algoritmen, in de virtuele wereld. Algoritmen die overigens zijn geïmplementeerd door mensen die het gelijkheidsbeginsel zouden moeten respecteren.

Om deze en andere wettelijke overtredingen te voorkomen, is het van belang om er bij het implementeren van de algoritmen in computersystemen zeker van te zijn dat de algoritmen in overeenstemming zijn met de overeengekomen rechten en wetten. In een vroeg stadium, vóór de implementatie, dienen de bestaande wettelijke en ethische eisen voor een algoritme of computersysteem vastgelegd te worden.

Naast de functionele eisen dient de implementatie van deze systemen dan overeenkomstig deze kwaliteitsaspecten te gebeuren. Wellicht zijn er nu al meerdere systemen geïmplementeerd die niet eerlijk en inclusief zijn. Deze dienen geanalyseerd te worden en, daar waar nodig, dienen ze weer bijgewerkt te worden om toch te voldoen aan de wettelijke en ethische normen. Het gelijkheidsbeginsel met dus ook toegepast worden op de cyberwereld van computersystemen.

Eric van Tol

Commisaris at The PMO Company

Ironisch genoeg zullen we discriminerende algoritmen met dezelfde algoritmen moeten bestrijden. Het is de data die het algoritme voedt. Wij maken die data en ja wij zijn discriminerend. De oplossing is ons algoritme-ontwerp én is onze data set keuze. Wellicht kunnen algoritmen kleinere ongelijkheitjes bestrijden. En we kunnen we ongelijkheid beter in kaart brengen of de oorzaken vaker begrijpen met zelflerenede algoritmen, maar op grote schaal ongelijkheid bestrijden is een morele revolutie met een veelheid van acties die voorlopig niet door zelflerende algoritmen gedaan gaat worden.

Zorg dat algoritmen uitlegbaar zijn voor iedereen.
Dit is om twee redenen vaak niet zo.

Ten eerste wordt door vele -vaak ten onrechte- ICT complex gevonden. Gemakkelijke toegang tot training is het enige antwoord.

Ten tweede veel algoritmen zijn een black-box en dus niet uitlegbaar. Het antwoord daarop is onderzoek. Onderzoek naar andere volgbare zelflerende algoritmen en onderzoek naar ondersteunden verklaringen/mechanisme voor zo'n black-box algoritme.

Zorg dat de algoritme-consument ontwerpmacht heeft. In het service design proces van bedrijven en overheden moet de consument meer aan het woord komen. Burgers of consumenten kunnen hun diensten niet aan passen en nauwelijks terugkoppelen of al helemaal niet weigeren. Zo is de privacy van een consument vaak niet het probleem maar het gebrek aan autonomie van de consument veel meer. Het mooiste is als de consument een dienst kan sturen.

Anders moeten we inclusiviteit voorschrijven.

Dr Eva Vanmassenhove

Assistant Professor Department of Cognitive Science & Artificial Intelligence, Tilburg University

Er zijn verschillende tools/algoritmen die kunnen helpen bij het detecteren van bias/discriminatie:

- Testing with concept activation (TCAV), een systeem dat biases (ras, gender, locatie) kan detecteren;
- AI Fairness 360 van IBM (zeventig fairness metrics die kunnen helpen bij het detecteren van bias, en tien bias mitigation algoritmen);
- FairML (een open-source toolkit).

Bovendien proberen verschillende initiatieven AI in te zetten voor 'social good'. Deze initiatieven focussen niet noodzakelijk op inclusiviteit of bias, maar op het detecteren van cyberpesten, online intimidatie, armoededetectie, enz. Het ethische aspect van AI-technologie krijgt ook steeds meer aandacht. Zo zijn er steeds meer richtlijnen voor betrouwbare, transparante en innovatieve AI, zoals die van de Europese Commissie.

De eerste stap is waarschijnlijk (h)erkennen dat, ondanks het feit dat algoritmen vaak als objectief, neutraal en dus bijgevolg als unbiased worden omschreven, ze juist vaak inclusief noch objectief zijn en stap twee is de oorzaken hiervan proberen te begrijpen.

> ***"De eerste stap is (h)erkennen dat algoritmen juist vaak inclusief noch objectief zijn. Vervolgens moet je de oorzaken hiervan proberen te begrijpen."***

Momenteel kan, in grote lijnen, onderscheid gemaakt worden tussen drie specifieke technieken om ervoor te zorgen dat onze algoritmen minder biased zijn: (1) preprocessing van data, (2) in-processing, en (3) post-processing. Dit zijn allemaal voorbeelden van debiasing techniques of bias mitigation techniques.

> ***"Een divers team dat aan applicaties werkt is zich waarschijnlijk meer bewust van de verschillende manieren waarop data biased kunnen zijn."***

1. **Pre-processing:** Besteed meer aandacht aan de trainingsdata ('garbage in, garbage out'). Trainingsdata kunnen voor een bepaalde taak bijvoorbeeld incompleet, niet divers, biased en niet representatief of slecht gedefinieerd zijn. Meer aandacht besteden aan de kwaliteit van de trainingsdata is dus zeker belangrijk als je inclusievere modellen wilt trainen. Wanneer je algoritme alleen getest en getraind werd op 'blanke mannen van 40 jaar' zal het hoogstwaarschijnlijk ook alleen goed werken voor deze specifieke doelgroep. Prima als je dat voor ogen had, maar niet goed als je iets probeert te ontwikkelen wat in principe voor iedereen zou moeten werken. 'Reweighing technieken' worden gebruikt om bias te voorkomen. De pre-processing approach is echter vaak moeilijk, prijzig (veel manuele interventie) en kan ook privacyproblemen met zich meebrengen, aangezien je bepaalde sensitieve attributen moet opslaan/weten om ervoor te zorgen dat er een zekere balans is.

2. **In-processing:** Je kan ook het model/algoritme zelf aanpakken en ervoor zorgen dat een classifier bijvoorbeeld tegelijkertijd focust op accurate voorspelling en het reduceren van bias tegen een bepaald, mogelijk sensitief attribuut. Deze techniek heet 'adversarial debiasing'. Je krijgt dan eigenlijk een combinatie van twee modellen waarbij het eerste probeert een accurate voorspelling te maken, en het tweede model (de 'adversary') probeert dat sensitieve attribuut op basis van de voorspelling van het eerste model te voorspellen. Het doel is om het eerste model zo accuraat mogelijk te laten voorspellen, terwijl je de voorspellingscapaciteit van het tweede model zo laag mogelijk houdt. Dit zou ervoor moeten zorgen dat voorspellingen minder biased zijn, aangezien het op basis van de voorspellingen niet mogelijk zou mogen zijn om sensitieve attributen te voorspellen (dankzij de 'adversary', zie het tweede model).

3. **Post-processing:** Zodra de voorspelling gemaakt is, kun je ook nog correcties proberen door te voeren die de voorspellingen eerlijker of meer inclusief maken. Hiervoor bestaan ook verschillende technieken.

   Een makkelijk voorbeeld binnen het vakgebied van automatische vertalingen is:
   - EN: 'I am happy' (genderneutraal)
   - FR: 'Je suis heureux' (masculien gender)

   In een post-processing step zou je deze zinnen kunnen identificeren en vervolgens ook een vrouwelijk/neutraal alternatief aanbieden:
   - FR 'Je suis heureuse' (feminien gender)

De meeste van deze technieken zijn taakafhankelijk. Bias mitigation technieken voor classificatie-taken zien er hoogstwaarschijnlijk anders uit dan bias mitigation technieken specifiek voor bijvoorbeeld vertalingen.

Andere belangrijke elementen om bias te bestrijden en inclusievere AI modellen te creëren zijn:

A. De kwaliteit/diversiteit van de testdata waarmee onze modellen getest worden;

B. Humans in the loop (manuele evaluatie waar nodig om eventuele problemen bloot te leggen);

C. Diversiteit in teams die technologie ontwikkelen. Aangezien wij ook heel erg biased zijn, is het goed om een divers team samen te stellen dat werkt aan applicaties. Een meer diverse groep is zich waarschijnlijk meer bewust van de verschillende manieren waarop data biased kunnen zijn;

D. Tijdige updates van de richtlijnen en het legale framework/de regelgevende en juridische systemen rond AI die kunnen helpen discriminatie te voorkomen.

Rens van der Vorst

Technofilosoof

De algoritmen van LinkedIn tonen mij elke dag wel een artikel van de één of andere kunstmatige intelligentie die iets fantastisch kan. In China werd een AI gelanceerd die emoties kan herkennen. In Amerika ontwikkelde men een AI die gewone mensentaal om kan zetten in programmacode. In Engeland schrijft een AI artikelen voor The Guardian. Je zou zeggen dat als AI zich zo snel ontwikkelt, dat kansen biedt om AI in te zetten om discriminatie en ongelijkheid te bestrijden. Of niet? Ik denk van niet.

> ***"Mensen discrimineren en zijn bevooroordeeld. Maar ik betwijfel of een AI betere beslissingen of eerlijkere beslissingen neemt."***

Neem nu bijvoorbeeld AI die emoties kan herkennen. Kan het dat echt? Of heeft AI een wel erg beperkte opvatting van wat emoties zijn? Worden emoties teruggebracht tot bepaalde gezichtsuitdrukkingen? En is dat niet het allergrootste probleem? Niet dat AI steeds slimmer wordt, maar dat AI beperkt is? Maar dat we er ons toch aan conformeren en de hele dag rondlopen met een stupide grijns? Of met een overdreven serieus gezicht?

Daarom ben ik huiverig om AI in te zetten om discriminatie en ongelijkheid te bestrijden. Natuurlijk, mensen discrimineren. Natuurlijk, mensen zijn bevooroordeeld. Maar neemt een AI betere beslissingen? Eerlijkere beslissingen? Ik betwijfel het. Kun je er wel voor zorgen dat de input van AI niet bevoordeeld is?

Het klinkt makkelijker dan het is. In heel veel vooroordelen zit namelijk een kern van waarheid. Vrouwen plegen veel minder vaak recidive, dus AI die straffen uitdeelt, moet ook rekening houden met gender. Dat is wel zo eerlijk. Of toch niet? Maar wat als vrouwen nu veel vaker afhaken in het bedrijfsleven? Moet AI daar ook rekening mee houden? Of worden daardoor andere problemen in de maatschappij versterkt?

> ***“ AI wint van grootmeesters in schaken. Misschien kan AI dan ook wel even uitzoeken hoe we inclusiever worden. ”***

Met welke vooroordelen moeten we wel rekening houden en met welke niet? Een onmogelijke puzzel. En als we de puzzel al zouden kunnen leggen, dan ontdekken we dat AI ergens in de ‘black box’ wel weer andere vooroordelen vindt en versterkt. Vooroordelen die we niet hadden verwacht. Misschien is dat nog wel erger, want het gebeurt ergens in de black box. We denken dat het eerlijker is, maar is dat wel zo?

Cathy O’Neill zei ooit dat een succesvol etentje met haar kinderen veel groentes bevat. Haar kinderen vinden dat een succesvol etentje veel Nutella bevat. Als O’Neill de AI-code mag schrijven, komen er veel groentes uit, bij de kinderen liters Nutella. Algoritmen zijn daarom altijd meningen, gevat in code.

Dat kan prima zijn, als we vinden dat onze maatschappij inclusiever en eerlijker moet worden. Maar het gevaar is dat we ons vervolgens verbergen achter algoritmen. We nemen geen verantwoordelijkheid meer voor onze beslissingen, maar verwijzen naar de magie van AI. AI neemt mensen aan op basis van gezichtsuitdrukking, en doet dat volledig eerlijk. Echt? Op basis waarvan is gedefinieerd welke gezichtsuitdrukkingen goed zijn? En wat ziet AI als een gezichtsuitdrukking? Is het wel eerlijker? Of lijkt het alleen zo?

Maar misschien pak ik het verkeerd aan. Misschien onderschat ik AI en moet ik AI gewoon laten uitzoeken hoe de wereld inclusiever wordt. Immers, AI wint van grootmeesters in schaken, AI kan pokeren, GO & Jeopardy winnen. Misschien kan AI dan ook wel even uitzoeken hoe we inclusiever worden. Of functioneert AI vooral goed als de scope (het speelbord) en de doelen (winnen) duidelijk zijn? En dat is hier natuurlijk niet het geval. Hier is het hartstikke moeilijk om te bepalen wat spel en wat doel is, en dan is AI net een geest in de fles. Als je niet oppast wat je wenst, gaat het mis. Denk maar aan Koning Midas.

O ja, en er zal best gezegd worden dat er bij alle beslissingen altijd een 'human in the loop' zal zijn. Ik geloof niet zo in die uitdrukking. Het suggereert toch dat computers het voor het zeggen hebben en mensen ook 'in the loop' zijn. 'Computer in the loop' lijkt me dan een betere uitdrukking.

Prof. Toby Walsh

FAA Laureate Fellow & Scientia Professor, AI School of CSE, UNSW Sydney

Ja, we kunnen algoritmen gebruiken voor het algemeen welzijn, om discriminatie en ongelijkheid te bestrijden. Maar het zal niet gebeuren als we niet hard werken. De techbedrijven verkochten ons de grote leugen dat algoritmen onbevooroordeeld zijn. Ze kunnen echter net zo bevooroordeeld zijn als mensen, zelfs erger.

We moeten heel voorzichtig zijn, opdat ze de vooroordelen van het verleden niet bestendigen. Vooral wanneer ze gebruikmaken van machine learning, getraind op gegevens die onvermijdelijk historische vooroordelen weerspiegelen. Evenzo kunnen algoritmen ervoor zorgen dat we eerlijkere, meer op bewijs gebaseerde, beslissingen nemen die de onbewuste vooroordelen waartoe mensen geneigd zijn, negeren.

Hoe we ervoor zorgen dat algoritmen op een inclusieve manier keuzes voor ons neemt? Gaat het om algoritmen die inclusieve beslissingen voor ons maken, kunnen inclusieve en diverse teams in de kamer helpen de juiste vragen te stellen. Het kan ook gaan om terugduwen en bepaalde belangrijke beslissingen niet overdragen aan machines, maar altijd door een mens te laten nemen. Alleen mensen hebben empathie en kunnen ter verantwoording worden geroepen. Daarom moeten we veel beslissingen in de rechtspraak en het leger nooit aan machines overlaten, zoals veroordelingen of beslissingen over leven en dood.

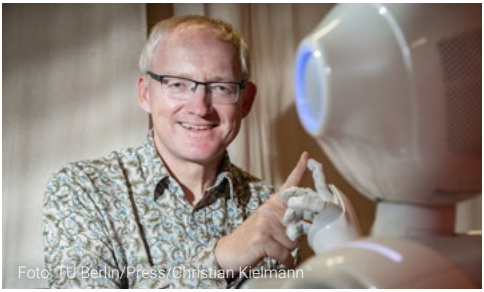


Foto: TU Berlin/Press/Christian Kielmann

> *“Alleen mensen hebben empathie en kunnen ter verantwoording worden geroepen.”*

Mr. Dr. Bart Wernaart

Professor Moral Design Strategy, Fontys Hogescholen

We moeten ervoor zorgen dat inclusiviteit aan de ontwerptafel van algoritmen een sterk uitgangspunt is. Hierbij zijn twee dingen van groot belang.

Op de eerste plaats dienen de maatschappelijke waarden beter afgestemd te worden op de waarden waarmee ontworpen wordt: veelal is het design van algoritmen gericht op effectiviteit en optimalisatie, en niet op het bijdragen aan maatschappelijke oplossingen. Terwijl juist die maatschappelijke waarden zo sterk geraakt kunnen worden wanneer de toepassing van algoritmen schadelijke neveneffecten krijgt. Denk aan de toeslagenaffaire.

> *"Het design van algoritmen is vaak gericht op effectiviteit en optimalisatie, niet op het bijdragen aan maatschappelijke oplossingen."*

Foto: Angeline Swinkels

Op de tweede plaats is het van belang dat het individu serieus gehoord wordt aan die tekentafel, bij uitstek diegenen die uiteindelijk de gevolgen dragen van algoritmische toepassingen. Dat gaat verder dan 'draagvlak creëren' waarbij dat individu via inspraakprocedures/ stakeholderconsultatie achteraf kan tekenen bij het kruisje. Het betekent dat je op basis van de morele verwachtingen van betrokkenen gaat ontwerpen.

Szymon Wróbel

Professor of philosophy at the Faculty of Artes Liberales at the University of Warsaw

Het netwerkalgoritme lijkt in eerste instantie te voldoen aan de eis van J. Rawls' theorie van rechtvaardigheid. Algoritmen lijken een 'oorspronkelijke positie' in te nemen, 'neutraal' te zijn en achter een 'sluier van onwetendheid' te blijven. Maar is dat wel zo?

Google is een bedrijf dat is gebaseerd op een algoritme. Google is schijnbaar een zoekmachine. Het oorspronkelijke algoritme organiseert het hele World Wide Web volgens de peer citatiemodellen die kwantificeren welke papers het meest invloedrijk en relevant zijn. Deze computationele meritocratie (volgens de verklaring van het bedrijf) staat in dienst van een universalistische missie om de informatie van de wereld niet alleen te ordenen, maar 'toegankelijk en nuttig te maken'.

> ***"De economie van Google is gebaseerd op de transformatie van het web in een gigantisch advertentieplatform en van reclame in een netwerk van rekenpunten en kliks."***

Het punt is dat de inkomsten van Google niet voortkomen uit de directe levering van detailhandelsdiensten, zoals zoeken, maar uit de tegeldemaking van de aandacht die de gebruiker in de loop van zijn engagement aan de diensten besteedt. De economie van Google is gebaseerd op de transformatie van het web in een gigantisch advertentieplatform en op de transformatie van reclame in een netwerk van rekenpunten en kliks. Laat u niet misleiden, dit wordt niet gedaan voor het algemeen belang, maar voor de enorme inkomsten van het bedrijf.

Op het eerste gezicht consolideren platforms (zoals Google) heterogene actoren en gebeurtenissen in meer geordende allianties, maar zij bevinden zich niet noodzakelijkerwijs in een werkelijk centrale positie ten opzichte van die allianties. De universaliteit van platforms maakt ze formeel open voor alle gebruikers, zowel menselijke als niet-menselijke. Kunnen we dan vertrouwen op platform-democratie?

> ***“ Wij zijn ons er niet van bewust dat bepaalde politieke standpunten al in de hardware zijn ingebouwd. ”***

Mijn antwoord is: nee. Platforms zijn gebaseerd op een standaardisatie van hun essentiële componenten. De formele politiek van platforms wordt gekenmerkt door de paradox tussen een strikt en onveranderlijk mechanisme (autocratie van middelen) en een opkomende heterogeniteit van zelfgestuurd gebruik (vrijheid van doelen).

Is de gebruiker echter soeverein in zijn beslissingen en kiest hij zijn doelen daadwerkelijk autonoom? Wij zijn ons er niet van bewust dat bepaalde politieke standpunten al in de hardware zijn ingebouwd. Voor veel processorchips is de ‘kerngebruiker’ een soevereine figuur die ondergeschikte administratieve subjecten kan voortbrengen, die op hun beurt de berekenende toegang van andere gebruikers kunnen controleren. Dit betekent dat ‘politiek’ niet alleen te vinden is in de juridische consensus, maar ook rechtstreeks in de machines. Beslissingen worden niet alleen óver infrastructuur genomen, zij worden dóór infrastructuur genomen.

Hans de Zwart

Filosoof en docent/onderzoeker, Hogeschool van Amsterdam

Medewerkers van campagnebureau Cambridge Analytica hadden ongeoorloofd toegang verkregen tot de profielen van miljoenen Facebookgebruikers. Deze data werden in 2016 gebruikt voor de verkiezingscampagne van Donald Trump.

Dit schandaal heeft een belangrijke consequentie gehad: data zijn een politiek issue geworden. Iedereen ziet nu opeens het dominante businessmodel van het internet. Een model waarbij het afbreken van privacy centraal staat en publieke waarden onder druk komen te staan. Wat wij hier samen allang weten, begrijpt de rest van de samenleving nu ook. Zonder privacy heb je geen eerlijke verkiezingen, zonder privacy kun je gelijkheid niet garanderen, en zonder privacy zijn we niet vrij en autonoom.

Steeds meer menselijk gedrag wordt gemodelleerd en steeds vaker wordt de kloof tussen model en werkelijkheid vergeten. Een model werkt met statistische waarden zoals gemiddelden, maar geen enkel mens is een gemiddelde. Onze menselijkheid is te complex om in het hokje van een model te passen.

Foto: Juri Hiensch

> ***"Iedereen ziet nu opeens het dominante businessmodel van het internet waarbij het afbreken van privacy centraal staat en publieke waarden onder druk komen te staan."***

Er zit een kloof tussen daadwerkelijke studenten en wat het learning analytics-systeem bijhoudt over studenten. De data in de verschillende zorgsystemen geven nooit een volledig beeld van een patiënt. En ook de meetdata over werknemers zeggen niet veel over hoe de werknemers hun werk doen. Naarmate we steeds vaker geautomatiseerde beslissingen nemen op basis van de systeemwerkelijkheid en niet het individu, wordt de kans groter dat we in die kloof vallen. Te beginnen met de mensen die in deze samenleving al een zwakkere positie innemen.

> ***“ Geen enkel mens is een gemiddelde. Onze menselijkheid is te complex om in het hokje van een model te passen. ”***

Een verandering moet beginnen met de intentie om modellen te veranderen en mee te nemen. De aanname dat computers (algoritmen) objectief, neutraal, correct en foutloos zijn, moeten we verlaten. Betere aanname is dat een computermodel ook de reflectie van een racistische kijk van de wereld meeneemt.

Nawoord

Voor dit boek vroeg ik veel experts in binnen- en buitenland naar hun visie op AI, inclusiviteit en vooroordelen. De lijst van experts die zich dagelijks met deze actuele materie bezighoudt, reikt echter veel verder dan de experts die ik benaderde. Die lijst groeit gestaag. Zowel wetenschap, overheid als bedrijfsleven omarmt steeds zichtbaarder de noodzakelijke discussie over vooroordelen en discriminatie binnen algoritmen en de potentiele gevaren ervan. Een ontwikkeling waar ik erg blij van word.

Ook binnen Fontys Hogeschool ICT (FHICT) houd ik me actief bezig met de boeiende materie van AI, algoritmen en bias. In de door Gerard Schouten en kwartiermaker Danny Bloks geleide AI-groep en met mijn mede-technofilosofen waaronder Rens van der Vorst, Jo-An Kamp, Huub Prüst en Lennart de Graaf, tracht ik de ethische kant van AI te ontleden. Via de podcastserie waarin we geen enkel aspect van AI onbesproken laten, ontstaat er een mooie verbinding tussen wat zou kunnen en wat wenselijk is. https://open.spotify.com/show/0hDFsglp8xvDsajgqZ6Jdf

Zoveel experts, zoveel invalshoeken en visies over of, en zo ja, hoe we algoritmen kunnen gebruiken voor het nemen van niet-discriminerende, inclusieve beslissingen, waarin de mens centraal staat. We zijn nog ver verwijderd van een eenduidige oplossing, maar we zijn al een eind op de goede weg. Want als er één ding duidelijk wordt uit dit boek, is het dat wetenschap, overheid en bedrijfsleven zich bewust zijn van de noodzaak die ondubbelzinnige oplossing te vinden. Samen is er nog een wereld te winnen.

Tot slot: mijn moeder is inmiddels van de schrik bekomen. Haar vertrouwen in de Nederlandse overheid is wel beschaamd. Maar waarschijnlijk minder als het vertrouwen van gedupeerden van de toeslagenaffaire. Inmiddels zijn de compensatieregelingen op gang gekomen, hoewel ook dat niet geheel vlekkeloos verloopt. Maar toch, hoe wrang en oneerlijk de toeslagenaffaire ook is: het heeft wel deze discussie in één klap op de brede maatschappelijke kaart gezet. Dat is op de lange termijn misschien toch een winst? Oh, en mijn zoon? Die kan niet wachten tot FIFA 22 uitkomt.

Over de auteur

Erdinç Saçan woont samen met zijn vrouw en drie zonen in 's Hertogenbosch. Hij heeft een passie voor een breed scala aan onderwerpen. Marketing, technologische ontwikkelingen, ethiek, privacy, data, verdienmodellen en overtuigingsprincipes hebben zijn grote aandacht. Maar met even grote interesse stort hij zich in zaken met betrekking tot politiek, sociale cohesie, vrijwilligerswerk en netwerken.


Foto: Ektor Tsolodimos

Erdinç Saçan werkt bij Fontys Hogeschool ICT als docent en coördinator. Ook is hij als Practor verbonden aan ROC Tilburg en is hij trainer bij Google Digitale Werkplaats en ICM. Als blogger schrijft hij onder meer voor Emerce.nl, Marketingfacts.nl, Dutchcowboys.nl en Digitalmarketingblog.nl.

Dankwoord

Wat begon als het schrijven van een artikel met de meningen van ongeveer vijf experts heeft na een jaar geresulteerd in een boekwerk met 40 experts. Een groot aantal andere experts konden helaas vanwege een drukke agenda niet meedoen maar hebben wel de moeite genomen om een reactie te sturen met vaak verwijzingen naar bronnen en namen van andere experts die ik kon benaderen.

Hierbij wil ik ook hen bedanken, in willekeurige volgorde:

Maria Axente, Frédérick Bruneault, Professor Stuart Russell, Professor Luciano Floridi, prof. dr. Max Louwerse, Prof.dr.ir. M. (Mehmet) Aksit, prof.dr.ir. PPCC Verbeek (Peter-Paul), Dr. S. (Sennay) Ghebreab. prof.mr.dr.ir. B.H.M. (Bart) Custers, Marietje Schaake, Peter Joosten, Geert ten Dam, Joshua B. Cohen, Assistant Prof. Loek Cleophas, Eddie Altenburg-Collin, Steven Van Belleghem, Prof.dr. Sabine Roeser, Piek Visser-Knijff, Edwin Borst, Dorothea Baur, Ahmed Larouz, Colette Cuijpers, Arjan Kors, Jessy Kouwenberg, Bert Deen, Hanan Challouki, Jeroen Vinkesteijn, Sander Duivestein, Ricardo A. Abdoel, Marc Appels, Jasper de Wilde, Chanel Matil Lodik.

Speciale dank voor Frank de Nijs, die mijn boek als eerste heeft geredigeerd.

Dankzij Sandra Verhoeven is het uiteindelijk een boek geworden.

Bianca Lathouwers, erg bedankt. Je hebt ieder woord en zin nagekeken. Bedankt voor je suggesties, feedback en verbeteringen.

Ook dank aan Audrey Kawarmala en Enson voor design en opmaak. We hebben vaak met elkaar hierover gediscussieerd.

Collega's en vrienden die mij bij het proces geholpen hebben met tips en adviezen: Aldwin van de Ven, Koen Suilen, Pauline Schepers-van der Rest, Twan Arts, Halil Kılıç.

Het managementteam van Fontys Hogeschool ICT, voor het vertrouwen en de vrijheid om jezelf te kunnen professionaliseren als docent.

Zonder steun en hulp van mijn lieve vrouw en kinderen was het mij niet gelukt. Erg bedankt Funda. Hartelijk dank Umut, Destan en Güven.

Bronnenlijst

Alexa Hagerty, A. A. (2021, april 15). The Conversation.
Opgehaald van https://theconversation.com/ai-is-increasingly-being-used-to-identify-emotions-heres-whats-at-stake-158809

Conger, K. (2021, April 13). New York Times.
Opgehaald van https://www.nytimes.com/2021/04/13/technology/racist-computer-engineering-terms-ietf.html

Durkee, M. (2021, April 8). OECD.AI.
Opgehaald van https://oecd.ai/wonk/how-to-achieve-trustworthy-algorithmic-decision-making

Feathers, T. (2021, Mei 20). Vice.
Opgehaald van https://www.vice.com/en/article/m7evmy/googles-new-dermatology-app-wasnt-designed-for-people-with-darker-skin

Heaven, W. D. (2020, September 23). MIT Technology Review.
Opgehaald van https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/2020/09/23/1008757/interview-winner-million-dollar-ai-prize-cancer-healthcare-regulation/amp/

Jorge Barrera, A. L. (2021, Mei 17). CBC.ca.
Opgehaald van https://www.cbc.ca/news/science/artificial-intelligence-racism-bias-1.6027150

Lang, S. (2021, April 7). Linkedin.
Opgehaald van https://www.linkedin.com/pulse/future-says-ethical-ai-sean-lang/

Russell, S. (2021, April 4). Standard.co.uk.
Opgehaald van https://www.standard.co.uk/news/uk/cambridge-university-scientists-ucl-china-b927785.html

Ujué Agudo, H. M. (2021, April 21).
The influence of algorithms on political and dating decisions.
Opgehaald van http://dx.doi.org/10.1371/journal.pone.0249454

Villanustre, F. (2021, April 8). CMS Wire.
Opgehaald van https://www.cmswire.com/information-management/make-responsible-ai-part-of-your-companys-dna/?_lrsc=691b57ab-ed94-45a7-afb6-d3c29e96d12f

Wood, M. (2021, Maart 22). Marketplace.
Opgehaald van https://www.marketplace.org/shows/marketplace-tech/bias-in-facial-recognition-isnt-hard-to-discover-but-its-hard-to-get-rid-of/

Fontys

› FOR SOCIETY

Erdinç Saçan, Fontys Hogeschool ICT, 2021