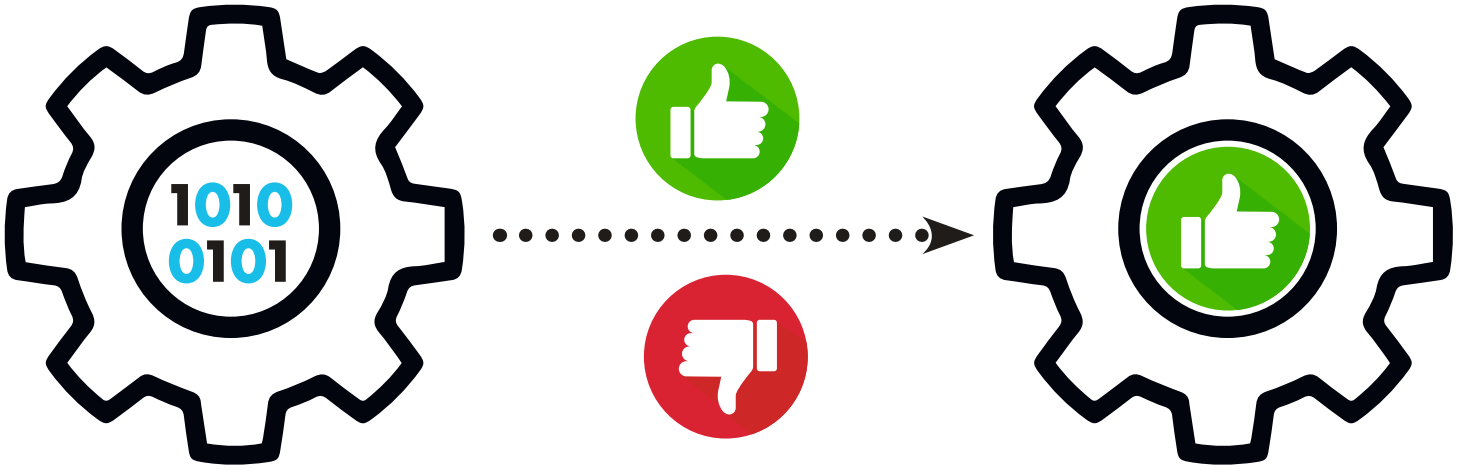
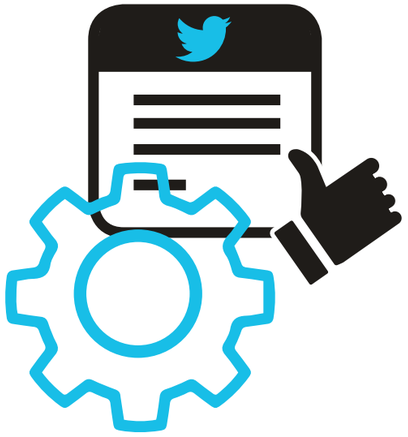
RECOMMENDER SYSTEMS: ALS DE KIP HET EI AANBEVEELT

De hoeveelheid data en informatie die één muisklik van ons verwijderd is, neemt al decennia lang exponentieel toe. Misschien is er daarom wel zoveel aandacht voor de gevaren van de *filter bubbels* en *echo chambers* die mensen weghouden van tegengestelde perspectieven. Het wegwerken van die ongewenste tendensen is alleen knap ingewikkeld.

door Mona de Boer en Frank van Praal beeld Shutterstock



IN HET PUBLIEKE DEBAT OVER DE FILTER BUBBLE IS VEEL AANDACHT voor *recommender systems* en hun potentieel om mensen intellectueel te isoleren. Deze automatische aanbevelers zijn gebouwd op algoritmes die de interesses en voorkeuren van gebruikers proberen te voorspellen, om daarmee nieuwe aanbevelingen te doen. De meeste recommender systems maken daarbij gebruik van *machine learning*. In het geval van recommender systems is daarbij sprake van actief leren: de algoritmes in recommender systems zijn gericht op het doen van voorspellingen die als aanbeveling dienen voor een gebrui-



bevelingen. Aanbevelingen van recommender systemen zijn gebaseerd op twee soorten relaties. De populaire methode *collaborative filtering* zoekt naar relaties tussen gebruikers, met als uitgangspunt dat gelijksoortige gebruikers, bijvoorbeeld twee jonge pubers met een passie voor sport, op zoek zijn naar dezelfde soort aanbevelingen. Een andere veelgebruikte methode, *content-based filtering*, maakt gebruik van relaties tussen content. Iemand die bijvoorbeeld graag naar dance muziek luistert, zal door het recommender systeem worden voorgesteld aan muziek met evenveel *beats per minute*. Er vindt in dit geval geen vergelijking van gebruikers plaats, maar van de karakteristieken van de muziek.
Bij beide methoden is (meta)data van gebruikers een belangrijke bron voor de uiteindelijke gepersonaliseerde aanbevelingen.

BIAS IN RECOMMENDER SYSTEMS

Eén van de belangrijkste veronderstelde risico's van recommender systems is ongewenste tendenties (bias) in de onderliggende algoritmes. Het gaat dan om systematische en repetitieve fouten in die systemen, die vertekende resultaten (kunnen) opleveren. Bij recommender systems uit dat zich vaak in de *popularity bias* en de *confirmation bias*.
De popularity bias doet vaak haar intrede wanneer vroeg in de ontwikkeling van algoritmes data ontbreken die nodig zijn om het systeem aanbevelingen te laten doen. Het is dan niet ongebruikelijk dat met willekeurige aanbevelingen wordt gestart. De gebruiker kiest iets in de hoop dat er een reden achter de aanbeveling zit. De keuze van de teleurgestelde gebruiker geldt vervolgens als bevestiging van die aanbeveling, en weerlegging van de niet gekozen opties. Een algoritme dat actief leert, zal deze informatie gelijk gebruiken voor volgende aanbevelingen, met een popularity bias voor die willekeurige aanbevelingen als gevolg die het functioneren van het systeem blijvend hindert.

Confirmation bias is het gevolg van de menselijke behoefte om met oogkleppen op bewijs te verzamelen dat de eigen denkbeelden bevestigt. Een ontwikkelaar van een recommender system die bijvoorbeeld denkt dat bejaarden uit de provincie vooral tijdschriften lezen over tulpen, zal in de data net zo lang op zoek gaan naar relaties die dit aantonen. Bijvoorbeeld door bepaalde variabelen opzettelijk niet mee te nemen in het algoritme. Het gevolg? De aanbevelingen van het systeem zijn vooral gebaseerd op de aannames van de ontwikkelaar, in plaats van de daadwerkelijke voorkeuren van de eindgebruiker.

ONTLUIKENDE AUDITS

Sinds 2018 heeft het wetenschappelijk onderzoek naar recommender systems, en in het bijzonder *algorithmic bias* in de machinerie daarvan, een vlucht genomen. Deze zogeheten *algorithm audit* studies richten zich met regelmaat op populaire web zoekmachines en (sociale) media platforms, en de wijze waarop deze informatie filteren, rangschikken en suggesties doen aan gebruikers. Wat deze ontluikende algorithm audit studies zonder uitzondering uitwijzen, is dat:

- Algorithmic bias een veelkoppig monster is, dat veel verschillende verschijningsvormen kent en ook nog eens in verschillende fasen van systeemontwikkeling en -gebruik zijn intrede kan doen;
- Methoden en maatstaven om algorithmic bias in recommender systems eenduidig aan te tonen en te kwantificeren nog in de kinderschoenen staan.

Dit vormt niet alleen een probleem voor onafhankelijk *ex-post* onderzoek naar dergelijke systemen in het publiek belang. Voor dat begint al, hindert de bias de ontwerpers en ontwikkelaars van recommender systems. Het ontbreekt hen vooralsnog aan voldoende beproefd instrumentarium om te voorkomen dat onbedoelde en ongewenste biases in die systemen worden geïntroduceerd of om verantwoording af te leggen over bedoelde biases.

Aanbevelingen voor nieuwe bezoekers zijn vaak gebaseerd op aannames ontwikkelaar

ker van het systeem. Wanneer een aanbeveling gedaan is, pleegt de gebruiker interventies op basis van die aanbeveling. Die interventies kunnen heel simpel zijn: bijvoorbeeld het liken of delen van een voorgesteld bericht op je Twitter-tijdlijn. De interventie bevestigt voor het algoritme dat de aanbeveling goed (of fout) was, en daarvan leert het algoritme. Door ervaring wordt het systeem zo geleidelijk beter in het doen van aan-

WETGEVING IN WORDING

De afgelopen jaren bogen wetgevers wereldwijd zich over de maatschappelijke risico's van *emerging technologies* en de bescherming van fundamentele mensenrechten bij de inzet daarvan. Zo zag in april 2019 de US Algorithmic Accountability Act het licht.
Dit wetsvoorstel beoogt dat bepaalde personen, partnerships en bedrijven die persoonlijke informatie van consumenten gebruiken, opslaan of delen, eerst *impact assessments* uitvoeren op de hoog-risico systemen die daarbij worden ingezet, en eventuele geïdentificeerde algorithmic biases "redelijk en tijdig adresseren".
Later dat jaar, in oktober, presenteerde een groep Amerikaanse senatoren de Filter Bubble Transparency Act (FBTA). Dit wetsvoorstel richt zich op het aansporen van grote online platforms om transparanter te zijn over het gebruik van algoritmes die gedreven worden door specifieke gebruikersgegevens. Ook zou het verboden moeten zijn gebruik te maken van 'ondoorzichtige' algoritmes zonder voorafgaande kennisgeving aan gebruikers.
Een jaar later, in december 2020, publiceerde de Europese Commissie (EC) haar voorstellen voor de Digital Services Act (DSA) en de Digital Markets Act (DMA). De DSA en DMA hebben onder meer tot doel om consumenten en hun fundamentele rechten online beter te beschermen en om een stevig transparantie- en verantwoordingskader te creëren voor online platforms.
Zeer recentelijk, in april 2021, bracht de EC haar voorstel uit voor een wettelijk kader voor het reguleren van AI-technologie. Daarin hanteert de EC een uitgesproken risicogebaseerde benadering, waarbij de nadruk ligt op het vaststellen van regels rondom het gebruik van hoog-risico- en verboden AI-praktijken. Een van de verboden praktijken, gerelateerd aan de *recommender systems*, betreft AI-technologie die 'verborgen' technieken inzet buiten het bewustzijn van een persoon om, met het oogpunt diens gedrag materieel te vervormen op een manier die waarschijnlijk fysieke of psychologische schade veroorzaakt. Of intellectuele isolatie als gevolg van vergaande personalisatie van ons wereldbeeld tot deze categorie behoort, wordt in het wetsvoorstel niet expliciet gemaakt.
Op dit moment constateren we dat voornoemde wetsvoorstellen voorlopig nog voorstellen zijn. Ze hebben met elkaar gemeen dat de bescherming van fundamentele rechten van personen centraal staat, dat er sprake is van een risicogebaseerde benadering voor het reguleren van datatechnologie, en dat algorithmic bias daarbij – als bedreiging van die fundamentele rechten - ruimschoots in het vizier is.
Opvallend is dat het fenomeen van de filter bubble as such niet expliciet terugkomt in de voorstellen.
In de voorstellen zoeken de wetgevers de oplossing van de complexe vraagstukken waar ze op gestoeld zijn voor-al in de hoek van transparantie, zelfevaluatie (risk management, impact assessments en *internal conformity assessments*) en onafhankelijke toetsing (*third-party conformity assessments*).
Laat dat nu net instrumenten zijn die pas goed werken als er eenduidige en in de praktijk beproefde definities, methoden en maatstaven zijn om algorithmic bias te identificeren en te evalueren. De ontwikkeling daarvan is inmiddels op gang gekomen, maar het heeft tijd nodig om te rijpen in theorie, net als te wortelen in de praktijk. Pas daarna kan je er goed mee handhaven.

AUTEUR

MONA DE BOER
is Partner Data & Technology bij PwC, voorzitter van de NOREA kennisgroep Algorithm Assurance en doet wetenschappelijk onderzoek naar algorithm audits.

AUTEUR

FRANK VAN PRAAT
werkt bij KPMG en leidt het Trusted Analytics team. Daarnaast zit hij in de kennisgroep Algorithm Assurance van NOREA.