AUDIT MAGAZINE

ARTIKELEN | 31 MEI 2023

AI: BIASED OR NOT BIASED?!

Auteur: Damian Borstel RA

Leestijd: 8 min

Bias is een thema dat iedereen raakt. Niet alleen omdat elk mens biases heeft, maar ook om dat deze in artificial intelligence (AI) systemen kunnen zitten. Wat biases zijn, welke biases voor kunnen komen en welke thema's/gebieden voor een AI-systeem relevant kunnen zijn, wordt in dit artikel beantwoord aan de hand van een conceptueel model (BIASED2).

Omdat AI-systemen door mensen worden ontwikkeld, moet je er altijd rekening mee houden dat er biases in terecht kunnen komen. Of het nu in data, het model zelf of in de interpretatie van de output zit, het is lastig om alle biases te identificeren (als dit al mogelijk is). Het is belangrijker om als organisatie transparant te zijn over welke biases gepoogd zijn te mitigeren, dan wel aan te geven welke biases er bewust aanwezig zijn in bijvoorbeeld de selectie van data. Een stakeholder kan dan zelf bepalen of hij een interactie aangaat met deze organisatie en haar AI-systemen. In de komende AI-Verordening van de EU wordt hier op ingegaan met een transparantieverplichting (artikel 52). Indien een AI-systeem interactie heeft met een natuurlijk persoon dan moet dit kenbaar worden gemaakt en duidelijk gedocumenteerd.

Het niet onderkennen van een bias in een AI-systeem kan leiden tot grote consequenties voor de (AI-onbewuste) gebruikers, zoals uitsluiting of willekeur

Commmotie

De afgelopen maanden is er veel commotie geweest over het gebruik van large language models (LLMs), bijvoorbeeld

door studenten. OpenAI lanceerde namelijk op 30 november 2022 ChatGPT (tijdelijk), een interessante tool en gebaseerd op een LLM, namelijk het GPT-3 model. Inmiddels maakt de betaalde versie van ChatGPT gebruik van het vernieuwde GPT-4 'large multimodal model'. Op een lay-outvriendelijke manier krijg je als gebruiker een antwoord op een vraag. Een vraag die ik als schrijver gesteld heb aan ChatGPT is: wat is het grootste risico van ChatGPT? Volgens ChatGPT zelf is dat biases. Op 2 februari 2023 twitterde Sam Altman, CEO OpenAI, dat ChatGPT tekortkomingen heeft omtrent biases en dat ze deze zullen verbeteren. Dit klinkt logisch aangezien ook websites en online fora als trainingsdata worden gebruikt.

Conceptueel model

Bias is dus een belangrijk onderwerp in deze dynamische wereld. Het niet onderkennen van een bias in een AI-systeem kan leiden tot grote consequenties voor de (AI-onbewuste) gebruikers, zoals uitsluiting of willekeur. Om dit onderwerp op een eenvoudige manier te kunnen bespreken, is in dit artikel gebruikgemaakt van een door mij ontwikkeld conceptueel model, genaamd BIASED2. Aan de hand van dit model zal het thema verder toegelicht worden. In mijn toekomstig proefschrift zal dit model en de keuze voor de onderwerpen uitgebreider beschreven worden.

De letters van BIASED2 staan voor:

- Bias
- Integrity
- Audit
- Security
- ESG
- Disclosure
- 2

B van Bias

Wat is een bias eigenlijk? Om antwoord te kunnen geven op deze vraag staan we eerst stil bij artikel 21 van het Handvest van de grondrechten van de Europese Unie. Dit artikel gaat over non-discriminatie en geeft aan dat iedere vorm van discriminatie (onder andere op grond van geslacht, ras, kleur of nationaliteit) verboden is. In 2019 schreef de Europese Commissie in haar 'high level expert group guidelines for trustworthy AI' over het bestaan van unfair biases die mogelijk kunnen leiden tot discriminatie (artikel 21). Dit zijn dus vooral discriminatoire biases die vaak voorkomen, maar er zijn ook andere categorieën mogelijk zoals 'human bias', 'systemic bias' en 'statistical/computational bias'.

Voor de definitie zijn verschillende bronnen en voorbeelden mogelijk en op basis van mijn biasedselectie breng ik de volgende onder de aandacht:

- De ISO-standaard 24028 definieert bias als een voorkeur voor dingen, mensen of andere groepen. Hoewel in deze standaard enkele voorbeelden gegeven worden, bestaan er in totaal wel meer dan honderd verschillende biases.
- In de NIST Special Publication 1270 *A proposal for Identifying and Managing Bias in Artificial Intelligence* wordt in de woordenlijst het begrip bias zelf niet beschreven, maar wel enkele voorbeelden van biases met hun definities. Deze NIST-publicatie gaat wel weer in op drie AI-biascategorieën, namelijk human bias, systemic bias en statistical/computational Bias.
- In de laatste versie van de EU-concept AI-Verordening (versie 25 november 2022) wordt geen definitie gegeven van de term bias, maar wel voorbeelden en mogelijke resultaten.

- Maken we gebruik van Google Translate, dan geeft deze aan dat biases vooroordelen zijn. Volgens Van Dale online betekent dit: ‘op een gebrek aan kennis berustende mening of afkeer’, wat weer in lijn is met de eerder benoemde ISO-standaard.

I van Integrity

Hoever moet je gaan om de integriteit van een AI-systeem vast te stellen? Dit is een lastige vraag om te beantwoorden, omdat je bijvoorbeeld als internal auditor veel verschillende soorten assurancewerkzaamheden kunt uitvoeren op het thema bias. Ook de diepgang van de werkzaamheden kan invloed hebben op de wijze hoe je met integrity omgaat (zie ook de volgende letter van het acroniem).

Ook de diepgang van de werkzaamheden kan invloed hebben op de wijze hoe je met integrity omgaat

Volgens de website Titlemax zijn er vijftig verschillende cognitieve biases, ChatGPT kan er makkelijk meer dan honderd benoemen en even zoeken op internet levert mooie visuals op met bijvoorbeeld een lijst van 188 cognitieve biases. Is meer biases identificeren dan beter voor de integriteit? Waarschijnlijk wel, definities en uitleg van mogelijke biases kunnen hierbij helpen, maar waar je dient te stoppen hangt sterk af van het beoogde doel van de organisatie, bijvoorbeeld enkel gericht op discriminatoire biases. Ook de scope van de assurancewerkzaamheden is hierbij van belang, bijvoorbeeld wanneer ‘third-partymanagement’ een rol speelt.

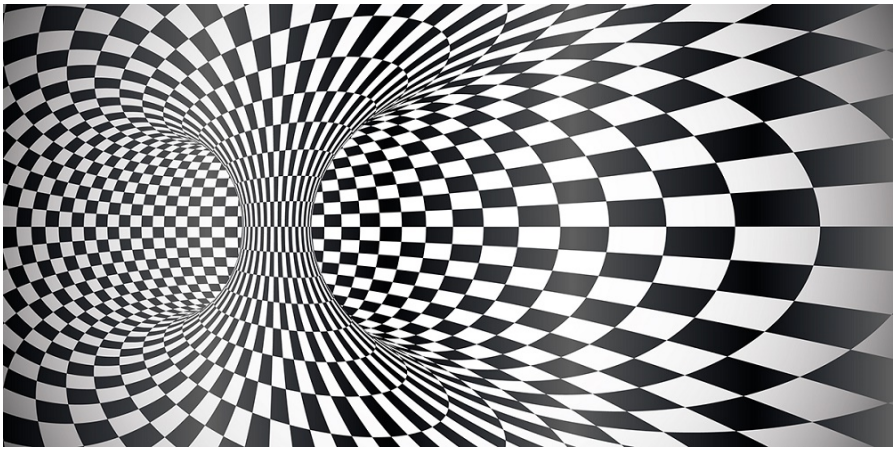
Volgens de AI-Verordening is datakwaliteit een belangrijke vereiste en is deze daarom expliciet opgenomen (artikel 10) en op hoofdlijnen toegelicht. Alle training, validatie en testdata dienen zo goed mogelijk vrij van fouten en volledig te zijn. Wat nu exact goed is, is ten tijde van dit artikel nog onduidelijk, maar raakt wel de integriteit van de data. Integriteit kan ook toepasbaar zijn op (uitgangspunten ten aanzien van) het algoritme, van personen (bijvoorbeeld data scientists) en andere zaken.

A van Audit

Is het mogelijk om een bias audit uit te voeren? Ondanks dat het voorbeeld hierna niet dichtbij huis is, wil ik dit als interessant inzicht graag delen. Als we namelijk naar Amerika kijken dan is dit volgens de New York City Counsel (NYC) wel degelijk mogelijk. NYC geeft de volgende definitie aan een bias audit (§ 20-870 Definitions): ‘The term ‘bias audit’ means an impartial evaluation by an independent auditor. Such bias audit shall include but not be limited to the testing of an automated employment decision tool to assess the tool’s disparate impact on persons …’

Vorig jaar is in NYC een wet aangenomen waarmee voor elk AI-systeem in een hr-proces, een bias audit door een derde partij verplicht is. Dit bracht nogal wat commotie met zich mee, veel vragen werden gesteld door belanghebbenden en er kwam volgens sommige juristen een afgezwakte nieuwe (concept)versie. De datum van inwerkingtreding is daarom verschoven van 1 januari 2023 naar 15 april 2023, en recentelijk weer verschoven naar 5 juli 2023. In New Jersey zijn nu ook plannen om NYC te volgen.

De Europese Commissie (EC) publiceerde recentelijk een conceptversie van de AI-Verordening (sinds 21 april 2021 zijn er diverse versies verschenen). Het doel van deze AI-Verordening is om de inwoners van de EU te beschermen op het gebied van mensenrechten, gezondheid en veiligheid. De EC verwoordde hiervoor een risicogebaseerde aanpak met in bijlage 3 een lijst van high-risk AI-systemen die moeten voldoen aan de AI-Verordening. In de AI-Verordening staan hr-systemen ook op de high-risk AI-systemenlijst en kunnen mogelijk naast de conformity assessment ook een ‘bias audit’ kunnen ondergaan.



Onder een conformity assessment wordt in de AI-Verordening verstaan: ‘het proces van verificatie van de naleving van de voorschriften van titel III, hoofdstuk 2, van deze verordening in verband met een AI-systeem.’ In hoofdstuk twee worden onder andere de volgende onderwerpen genoemd: een systeem voor risicobeheer, data en databeheer, technische documentatie, cyberbeveiliging en menselijk toezicht.

Volgens diverse publicaties (bijvoorbeeld van de SEO over digitale werving & selectie, juli 2022) gebruiken veel organisaties in een hr-proces al een AI-systeem (minimaal 27%) en afhankelijk van de stap in het werving- en selectieproces kan dit oplopen naar zelfs 96% voor het verspreiden van de vacature. Of zal dit ook (deels) onderdeel worden van de conformity assessment? Dat zal waarschijnlijk door CEN/CENELEC JTC21, een joint technical committee met diverse werkgroepen van de EC, uiterlijk in januari 2025 duidelijk worden (reeds vastgelegd in de concept Standardisation Request) via aanvullende guidance op de AI-Verordening van de EU.

S van Security

Heeft informatiebeveiliging (IB) van een AI-systeem effect op biases? Een goed werkend IB waarborgt bijvoorbeeld dat de gehanteerde data langer van goede kwaliteit is en integer blijft. Ditzelfde geldt uiteraard ook voor het AI-model zelf, waarbij het voor complexe modellen zoals neurale netwerken (ook wel black boxes genoemd) lastig is. Dit komt onder andere omdat de meerderheid van de modellen alleen correlaties als output hebben en deze zijn niet altijd te begrijpen voor een mens.

In 2021 hebben Chinese onderzoekers malware kunnen toevoegen aan een neuraal netwerk zonder dat de data scientists dit door hadden (Cornell University – EvilModel: Hiding Malware Inside of Neural Network Models). Het resultaat was een beïnvloeding van de output. Een bepaalde groep mensen zou door hackers bijvoorbeeld bewust kunnen worden uitgesloten bij de toepassing van AI in de praktijk.

In 2021 konden Chinese onderzoekers malware toevoegen aan een neuraal netwerk zonder dat de data scientists dit door hadden. Het resultaat was een beïnvloeding van de output

E van ESG

Wat is de link tussen ESG en biases? Het moeilijke is dat ESG uit drie verschillende onderwerpen bestaat, namelijk environment, social en corporate governance. Voor elk van deze onderwerpen kunnen biases ontstaan (zoals governance bias), worden versterkt of juist verminderd. Ik sta kort stil bij corporate governance. Het kan voorkomen dat de (corporate) governance van een organisatie (tone at the top) zo is ingericht dat er veel of juist weinig rekening wordt gehouden met het thema biases. Dan kan dit resulteren in een vermindering of vermeerdering van het aantal biases. De verschillende afdelingen binnen het three lines model hebben ook invloed op het aantal biases, denk hierbij aan de

aanwezigheid van een ethische commissie. Door diversiteit van een afdeling en met (externe) feedback kan het aantal en de impact van biases mogelijk verminderd worden.

D van Disclosure

Dienen organisaties transparant te zijn over de mate van biases? Dit zal per branche en organisatie verschillen. Toegeven dat je bepaalde biases hebt in een AI-systeem voelt als een behoorlijke stap, maar gebeurt bijvoorbeeld al als je kijkt naar gepersonaliseerde advertenties. Dit kan positief gevonden worden door de ontvanger zolang dat niet te vaak voorkomt, en na een aankoop nog steeds zichtbaar is. In China is er sinds 2022 al wetgeving voor deze recommendation systems, organisaties mogen daar geen onwettige of schadelijk informatie als trefwoorden invoeren in gebruikersinteresses of gebruikerstags.

Omgedraaid kan dit ook helpen voor de gebruikers en eigenaren van een AI-systeem om afgewogen keuzen te maken om wel of niet van een AI-systeem gebruik te maken. Niet voor elk AI-systeem, bijvoorbeeld bij de overheid, is een alternatief mogelijk voor de gebruiker. Dat zorgt voor een bepaalde afhankelijkheid waarbij transparantie nog belangrijker is. De Nederlandse overheid is wel gestart met een algoritmeregister en blijft deze aanvullen met nieuwe algoritmes en additionele toelichtingen. Mogelijk heeft dit ook een positieve werking op het verminderen van biases doordat de transparantie verhoogd wordt richting de gebruikers, en zij dan weer feedback kunnen geven aan de makers. Ook voor toezichthouders zal zo'n register inzichtverhogend werken en helpen bij hun (risicogebaseerd en dagelijks) toezicht.

2 van 'too'

2 staat hier voor 'ook' en betekent in het kort dat ook dit conceptueel model biased zal zijn door de perspectieven die ik heb gebruikt, mogelijk met enkele voorkeuren en ervaringen uit het verleden, heden en de toekomst. Dit conceptueel model is dan ook niet volledig en dat zal ook voor AI-systemen gelden. Mogelijk zijn er nu nog relatief veel biases in AI-systemen, omdat er weinig aandacht is voor biases. Wie zal het zeggen. Maar door ontwikkelingen en meer (wetenschappelijke) studies worden er meer varianten ontdekt en zal dit in ieder geval tot meer bewustzijn leiden bij de makers en het brede publiek.

Mogelijk zijn er nu nog relatief veel biases in AI-systemen, omdat er weinig aandacht is voor biases. Wie zal het zeggen

Conclusie

In dit artikel is beknopt stilgestaan bij het thema bias en waar een organisatie aan kan denken gebruikmakende van het BIASED2-model. Door de beperkte omvang en voorbeelden is dit artikel zeker niet volledig. Hopelijk zet het wel de lezer aan het denken. Het ligt namelijk niet altijd aan de data, ook door menselijke input bij de ontwikkeling kan een AI-systeem biases bevatten. Ofwel, wees continue alert!

BIAS-disclaimer: dit artikel is op persoonlijke titel en met de perspectieven van de schrijver geschreven in maart 2023. Er is geen gebruikgemaakt van een large language model om teksten te generen. Dit neemt echter niet weg dat ook dit korte artikel biases kan bevatten, de 'bias blind spot'.

AUDIT
MAGAZINE

Over
Damian Borstel RA is senior toezichthouder bij de Autoriteit Financiële Markten en houdt zich onder andere bezig met toezicht op AI. Daarnaast is hij betrokken bij de NOREA kennisgroep Algorithm Assurance, docent aan de Universiteit van Amsterdam en trainer bij het IIA.

Tags: biases

Bron url: https://auditmagazine.nl/artikelen/ai-biased-or-not-biased/