Verantwoord toepassen van AI vereist goede uitleg van geautomatiseerde besluitvorming

EXPLAINABLE AI: HET RECHT OP EEN GOEDE UITLEG

VOOR HET VERTROUWEN VAN DE MENS IN TECHNOLOGIE, EN SPECIFIEK ARTIFICIAL INTELLIGENCE (AI), IS HET VAN HET GROOTSTE BELANG DAT EEN GOEDE UITLEG KAN WORDEN GEGEVEN OVER HOE BESLUITEN TOT STAND ZIJN GEKOMEN. EEN RECENT VOORBEELD VAN DE BELASTINGDIENST LAAT ZIEN DAT HET VERTROUWEN EN DAARMEE OOK DE REPUTATIE EEN ENORME DEUK KRIJGEN ALS ER GEEN OF GEEN GOEDE UITLEG WORDT GEGEVEN. DE CONTEXT VAN EEN AI-TOEPASSING IS CRUCIAAL VOOR EEN GOEDE UITLEG. BIJ HET ONTWERP EN DE BOUW VAN ZO'N TOEPASSING IS HET DAAROM ZAAK OM DAAR VANAF HET BEGIN REKENING MEE TE HOUDEN. UITEINDELIJK MOET AI DE MENS DIENEN EN NIET ANDERSOM.

door Martin van den Berg en Stefan Leijnen beeld Shutterstock

DE TOESLAGENAFFAIRE BIJ DE BELASTINGDIENST HEEFT VOOR VEEL OPHEF GEZORGD EN DOET DAT NOG STEEDS. De nieuwe staatssecretaris, Alexandra van Huffelen, moest recentelijk Kamervragen beantwoorden over de kwestie. Daaruit kwam een aantal interessante inzichten naar voren [1]. Gedupeerden pikken het niet langer en willen inzage in hun dossier. Maar het blijkt dat dit wel maanden tot een halfjaar kan duren, omdat het samenstellen van een dossier gemiddeld 400 uren kost.

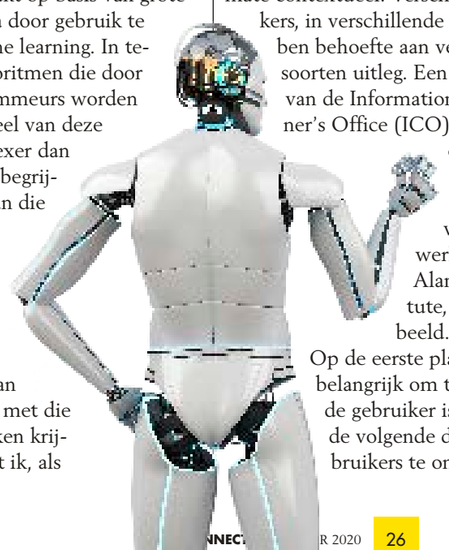
Daarnaast blijkt dat de informatiebehoeften van gedupeerden niet altijd aansluiten bij het ingediende verzoek 'inzage dossier'. De Belastingdienst gaat nu telefonisch contact opnemen met gedupeerden om te achterhalen aan welke informatie zij het meeste behoefte hebben. Kortom, mensen hebben behoefte aan uitleg van besluiten die hen raken. Die uitleg moet snel en gemakkelijk beschikbaar kunnen komen en moet specifiek gemaakt kunnen worden. Deze inzichten bevatten belangrijke lessen voor ontwerpers en bouwers van toepassingen op het gebied van AI. Deze toepassingen worden immers steeds complexer, waarbij het niet altijd meer duidelijk is hoe keuzes of besluiten tot stand zijn gekomen: de machine als black box. Het is van groot belang dat die uitleg wel aansluit bij de belevingswereld van

Het mensgericht ontwerpen is onderbelicht

de gebruiker, zoals een klant, aanvrager of patiënt. Gebeurt dat niet, dan stapelen affaires zoals bij de Belastingdienst zich op, zeker als in de nabije toekomst steeds meer besluiten op basis van AI tot stand komen.

ARTIFICIAL INTELLIGENCE

AI is een sleuteltechnologie die de wereld net zo ingrijpend gaat veranderen als de industriële revolutie dat in de 18e en 19e eeuw heeft gedaan[2]. Met AI worden steeds meer uitkomsten zoals aanbevelingen, besluiten en acties automatisch bepaald op basis van algoritmen. Zo'n algoritme wordt veelal geautomatiseerd gemaakt op basis van grote hoeveelheden data door gebruik te maken van machine learning. In tegenstelling tot algoritmen die door menselijke programmeurs worden geschreven, zijn veel van deze algoritmen complexer dan wij in staat zijn te begrijpen. Sommigen van die algoritmen zijn zo complex dat niet meer te herleiden is hoe uitkomsten tot stand zijn gekomen. Dit vergroot de argwaan van gebruikers die met die uitkomsten te maken krijgen. Waarom moet ik, als inwoner van Utrecht, meer betalen voor dezelfde autoverzekering als mijn broer die in Middelburg woont? Waarom kom ik niet in aanmerking voor een lening? Waarom krijg ik geen toeslag? Gelukkig is het besef wereldwijd doorgedrongen dat het verantwoord toepassen van AI een goede uitleg van geautomatiseerde besluitvorming vereist [3, 4]. In de AI-wereld is daarvoor een stroming ontstaan onder de noemer 'explainable AI', kortweg 'XAI' genoemd. Binnen deze stroming wordt onderzoek gedaan naar methoden en technieken om de werking van AI-toepassingen te interpreteren om daar een uitleg uit te destilleren. Maar over wat voor soort uitleg hebben we het, wie zit daarop te wachten en wat is een goede uitleg?



XAI

Binnen het lectoraat Artificial Intelligence van de Hogeschool Utrecht wordt onderzoek gedaan naar XAI. Het lopende literatuuronderzoek toont aan dat de meeste onderzoeken hun focus hebben op methoden, technieken en software om de black box open te maken om een uitleg te kunnen geven. Natuurlijk is dat nodig, maar de geringe aandacht voor de vraag wat feitelijk een goede uitleg is, geeft aan dat het mensgericht ontwerpen onderbelicht is. Een uitleg is in hoge mate contextueel. Verschillende gebruikers, in verschillende situaties hebben behoefte aan verschillende soorten uitleg. Een recent rapport van de Information Commissioner's Office (ICO), de Britse evenknie van de Autoriteit Persoonsgegevens, in samenwerking met het Alan Turing Institute, bevestigt dit beeld. [5]

Op de eerste plaats is het belangrijk om te weten wie de gebruiker is. Grofweg zijn de volgende drie typen gebruikers te onderscheiden.

1. Medewerkers van een organisatie wiens besluiten worden ondersteund door een AI-toepassing en die zinvolle informatie moeten doorgeven aan een persoon die geraakt wordt door het besluit van de AI-toepassing.
2. Degenen die door het besluit van de AI-toepassing worden geraakt, zoals aanvragers, klanten of patiënten.
3. Auditors of externe toezichthouders die belast zijn met de controle of het toezicht op de bouw en het gebruik van de AI-toepassing.

Het voorbeeld van een leningbeoordelingsproces laat zien dat de informatiebehoeften van deze gebruikers verschillen. Aanvragers waarvan een lening door een bank wordt afgewezen, willen weten waarom ze zijn afgewezen en voor welk bedrag ze wel in aanmerking komen. Als deze informatie niet rechtstreeks door de AI-toepassing wordt verstrekt maar door een medewerker, wil die medewerker begrijpen welke criteria door de AI-toepassing zijn toegepast en meer weten van de aanvrager om die adequaat te woord te kunnen staan. Een toezichthouder, ten slotte, wil begrijpen hoe het algoritme werkt en wat de impact daarvan is op het kredietrisico van de bank.

Naast het type gebruiker worden de volgende contextfactoren onderkend in het ICO- rapport [5].

- Het domein waar AI wordt toegepast. Een goede uitleg is veel noodzakelijker in het financiële domein dan voor gebruikers van bijvoorbeeld Spotify, die luisteraanbevelingen krijgen. Er zullen vast mensen zijn die zich afvragen hoe Spotify tot zijn aanbevelingen komt, maar het gros van de mensen zit niet te wachten op een uitleg. Maar in het financiële domein, als er geld in het geding is, zoals bij het aanvragen van leningen of toeslagen, moet ervan worden uitgegaan dat een goede uitleg noodzakelijk is, zeker in geval van een afwijzing.
- De impact die het besluit heeft op een individu of op de maatschappij. Besluiten in de zorg over leven en dood hebben een grote impact, in tegenstelling tot de aanbeveling om een kaartje te kopen voor een optreden van je favoriete artiest.
- De data die gebruikt zijn om de AI-toepassing te trainen en testen en de data die als input gebruikt worden om tot een bepaald besluit te komen. In bepaalde situaties zijn deze inputdata vast en kunnen niet worden gewijzigd, zoals bijvoorbeeld de uitkomst van een genetisch onderzoek. In andere situaties, zoals bij het toekennen van een lening, waarbij het toegekende bedrag afhangt van het bestedingspatroon van de aanvrager, liggen die inputdata helemaal niet zo vast. De aanvrager kan immers zijn bestedingspatroon wijzigen. In dat laatste geval houdt een goede uitleg in dat een aanvrager te horen krijgt bij welk bestedingspatroon hij wel in aanmerking komt voor de lening.
- De urgentie die samenhangt met de actie die op het besluit volgt. Hoe urgenter (en belangrijker) de actie voor een aanvrager, des te eerder wil die de uitkomst van het besluit weten en des te beter dient die uitkomst te worden onderbouwd.

VANAF HET BEGIN

Een goede uitleg is dus niet zomaar te produceren. Om toestanden als de toeslagenaffaire te voorkomen, moet bij het ontwerp van AI-toepassingen (en niet alleen AI-toepassingen) vanaf het begin rekening worden gehouden met het kunnen geven van een goede uitleg aan diverse typen gebruikers. Afhankelijk van de context van de AI-toepassing kan dat een standaarduitleg zijn die door AI-toepassing wordt gemaakt. Maar het kan ook zijn dat de AI-toepassing geen uitleg kan of mag geven en dat daar een mens voor nodig is ('human in the loop'). Voorbeelden hiervan zijn de zorg, waar een arts de uitleg geeft aan een patiënt, en de rechtspraak, waar een rechter de verdachte te woord staat. In het rapport van de ICO wordt een aanpak van zes stappen geschetst om vanaf de start uitlegbare AI-toepassingen te ontwerpen en te implementeren waarmee passende uitleg over besluiten kan worden gegeven aan diverse typen gebruikers. Kortom, 'explainability by design'. In meer algemene zin is ook het uitvoeren van een AI impact assessment (AIIA) een goed middel om te komen tot een mensgerichte AI-toepassing [6]. Met een AIIA kan de impact van een AI-toepassing vooraf helder in beeld worden gebracht, en vervolgens worden getoetst aan ethische en juridische kaders. Dat maakt het mogelijk om bij de ontwikkeling van de AI-toepassing rekening te houden met de belangen die gebruikers hebben bij uitlegbaarheid, en met de contextfactoren die daarbij een rol spelen. In Nederland pleit de Tweede Kamer voor kaders voor een ethisch en maatschappelijk verantwoorde inzet van AI [7]. Een aanpak voor uitlegbaarheid mag daarin beslist niet ontbreken. Een goede uitleg is immers een randvoorwaarde voor de acceptatie van AI. Ten slotte is het onze verwachting dat het ethische kader voor het verantwoord toepassen van AI op Europees niveau zal worden uitgebreid met een juridisch kader waarmee een goede uitleg, zoals dat ook met dataprivacy is gebeurd in de AVG, wettelijk wordt verankerd.

REFERENTIES

1. *AG Connect (2020). Belastingdienst doet 400 uur over elk dossier toeslagenaffaire. Geraadpleegd: www.agconnect.nl/artikel/belastingdienst-doet-400-uur-over-elk-dossier-toeslagenaffaire?tid=TIDP353191X88C-67F1E215146F4AFA6CA9FB7BD49C0Y15&utm_campaign=AGC_nieuwsbrief%20 dagelijks&utm_medium=email&utm_source=SMG&utm_content=752_do%20 28%20mei.*
2. *AINED (2018). AI voor Nederland: Vergroten versnellen en verbinden. Geraadpleegd: www.nwo.nl/documents/enw/rapport-ai-voor-nederland-vergroten-versnellen-en-verbinden.*
3. *HLEG (2019). High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI. European Commission. Geraadpleegd: ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.*
4. *Asilomar AI Principles (2017). Geraadpleegd: futureoflife.org/ai-principles/?cn-reloaded=1.*
5. *Information Commissioner's Office (ICO) and Alan Turing Institute (2020). Explaining decisions made with AI. Geraadpleegd: ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence.*
6. *ECP, Platform voor Informatiesamenleving (2018). AI Impact Assessment. Geraadpleegd: ecp.nl/actueel/artificial-intelligence-impact-assessment.*
7. *Tweede Kamer der Staten-Generaal (2020). Update vereist. Naar meer parlementaire grip op digitalisering. Geraadpleegd: www.tweedekamer.nl/sites/default/files/atoms/files/eindrapport_tijdelijke_commissie_digitale_toekomst_tweede_kamer_der_staten-generaal.pdf*

REACTIES EN BIJDRAGEN

Voor reacties en nieuwe bijdragen van IT-experts:
Tanja de Vrede
020-2356415
t.d.vrede@agconnect.nl

AUTEUR

MARTIN VAN DEN BERG
is werkzaam als onderzoeker bij het lectoraat Artificial Intelligence van de Hogeschool Utrecht en als enterprisearchitect in de financiële sector.

STEFAN LEIJNEN
is lector bij het lectoraat Artificial Intelligence van de Hogeschool Utrecht en werkt daarnaast als adviseur op het gebied van AI governance (stefan.leijnen@hu.nl).