HIGHBERG

Auteurs
Sebastiaan Streng
Thomas Mollema

WAT ALS JE DE UITKOMSTEN VAN AI-SYSTEMEN NIET BEGRIJPT?

Inhoudsopgave

Stel je voor: je vraagt een bijstandsuitkering aan via het online portal van je gemeente en krijgt naderhand een afwijzing. De gemeente handelt naar de uitkomst van een AI-systeem dat jouw aanvraag heeft beoordeeld. Maar een uitleg waarom de aanvraag is afgewezen, ontbreekt. Misschien waren je inkomsten toch te hoog, of misschien sloeg het systeem aan op iets anders; het wordt in ieder geval niet verduidelijkt. Het gebrek aan uitleg laat je met vragen achter, en je hebt geen vertrouwen in het oordeel. Dit hypothetische scenario is helaas niet ongewoon in een publieke sector waarin de uitkomsten van AI steeds meer invloed hebben op belangrijke beslissingen. Daarom is Explainable AI, ofwel 'uitlegbare AI' zo belangrijk.[1]

[1] Dit artikel is specifiek gericht op overheidsorganisaties en organisaties die opereren in de publieke sector. Echter is uitlegbare AI net zo noodzakelijk voor andere sectoren, zoals het onderwijs en de zorg. In het onderwijs kunnen steeds meer onderdelen van het lespakket met behulp van AI-systemen worden gegenereerd. Echter, hoe kan een docent verantwoordelijkheid dragen voor de AI-gegenereerde inhoud of lesvorm als de werking van het system niet aan haar verklaard wordt? In de zorg is het automatiseren van foto- of scaninterpretaties een belangrijke toepassing van AI. Hier ligt de noodzaak van uitlegbare AI dan weer in de reproductie van de criteria op basis waarvan een bepaalde uitkomst is gegeven (wel of geen tumor gesignaleerd,t van het systeem te volgen, of om een andere keuze te maken.

Waarom is uitleg belangrijk en wat is uitlegbare AI?

In de publieke sector, waar beslissingen directe gevolgen hebben voor burgers, is transparantie van essentieel belang. Stel je voor dat op basis van een AI-gestuurd systeem een beslissing wordt genomen over welke gezinnen recht hebben op sociale steun. Zonder uitleg kan een afwijzing niet alleen frustrerend zijn, maar ook tot onrechtvaardige situaties leiden. Burgers verwachten duidelijkheid over beslissingen en organisaties moeten verantwoording af kunnen leggen over hun handelen.

Een AI-systeem wordt doorgaans gedefinieerd als een algoritmisch systeem dat, op basis van invoer en uitkomsten, een taak uitvoert waarbij een zekere mate van intelligentie of leervermogen vereist is. Uitlegbare AI richt zich op de invoer en de uitkomsten en het biedt inzicht in hoe die uitkomsten tot stand zijn gekomen. Deze uitkomsten vormen vaak de basis voor beslissingen of acties die door het systeem worden genomen. Uitlegbare AI gaat echter verder dan louter transparantie. Het is een essentiële voorwaarde om grip te houden op AI-systemen en verantwoord gebruik ervan te waarborgen.

Er zijn drie fundamentele uitgangspunten die ervoor zorgen dat uitlegbare AI bijdraagt aan beheersbaarheid van een AI-systeem:

Transparantie

Geef begrijpelijke toegang tot de informatie die door het systeem is gebruikt, tot de uitkomsten van het systeem die tot de beslissing leidden en tot hun onderlinge relatie.

Verantwoording

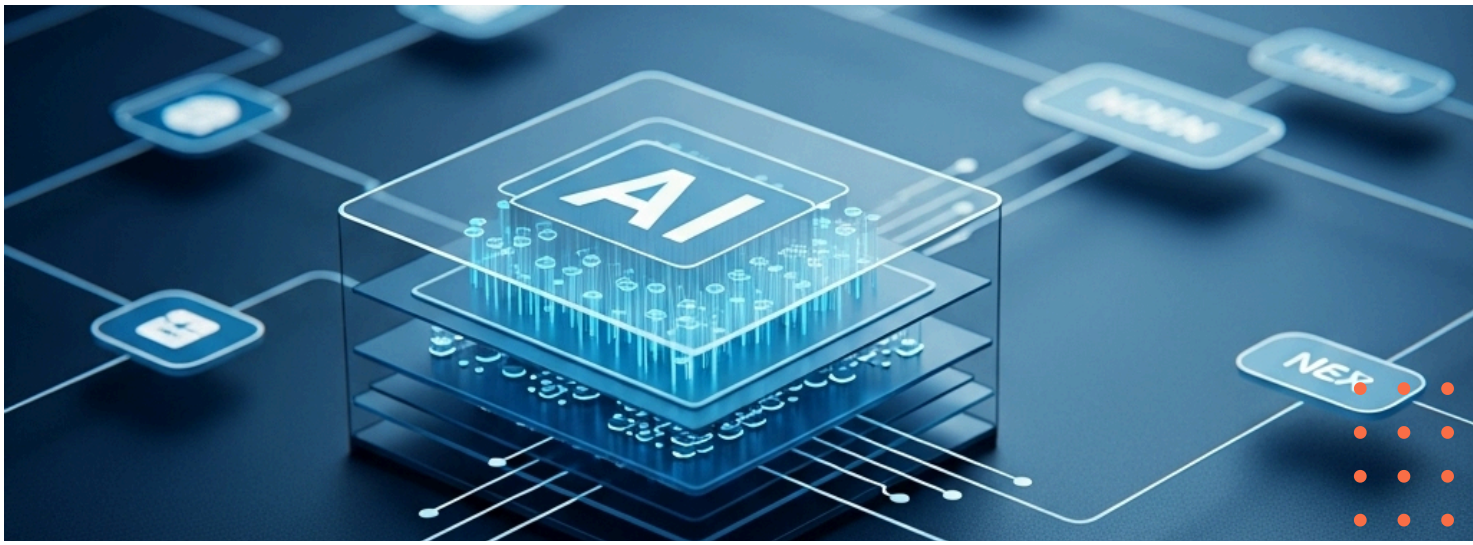
Motiveer hoe de beslissing volgt uit de uitkomsten van het systeem en wees verantwoordelijk voor de uitkomsten van het systeem.

Monitoring

Analyseer de uitkomsten van het systeem en de resulterende beslissingen. Dit stelt men in staat om de uiteindelijke beslissingen met ongewenste effecten, zoals discriminatie en andere vormen van bias, te herkennen en mitigeren en in het vervolg te voorkomen.

Transparantie, verantwoording en monitoring zorgen dat we de controle over het AI-systeem behouden. Tegelijkertijd dragen deze uitgangspunten bij aan het vertrouwen in het AI-systeem. Als publieke sector staan we voor de uitdaging om het gebruik van AI-systemen te omarmen om doelmatiger en doeltreffender te werken, waarbij deze systemen mensgericht en verantwoord ingezet moeten worden.

Naarmate AI-systemen complexer worden, wordt het uitdagender om inzicht te bieden in hoe specifieke uitkomsten tot stand zijn gekomen. Voor deep learning (DL) en machine learning (ML) AI-systemen gebaseerd op grote hoeveelheden data kom je niet tot absolute zekerheid, zoals wel mogelijk is bij op logische regels gebaseerde AI-systemen. Toch kunnen uitlegbaarheidstechnieken helpen om een bepaalde mate van inzicht te bieden. De mate van uitlegbaarheid is afhankelijk van de gebruikte AI-modellen en de methoden die worden ingezet om transparantie te waarborgen.



Hoe werkt uitlegbare AI in de praktijk?

Laten we aan de hand van een fictieve variant op het voorbeeld uit de inleiding verkennen hoe uitlegbare AI werkt in de praktijk. We laten zien hoe de uitgangspunten van transparantie, verantwoording en monitoring hier een rol in spelen. We werken één eenvoudig voorbeeld uit om de toelichting overzichtelijk en begrijpelijk te houden. Daarbij is bewust gekozen voor een toepassing van neurale netwerken (een vorm van ML), omdat dergelijke modellen in de praktijk veel voorkomen, in toepassingen zoals beeldherkenning en spraak/taalgeneratie en -verwerking. Tegelijkertijd is het belangrijk te onderkennen dat er ook complexere AI-systemen bestaan, zoals generatieve AI (een vorm van DL), waarbij de mate van transparantie en uitlegbaarheid verder beperkt kunnen zijn. Toch, zoals we zullen laten zien, geeft de manier waarop men de uitkomsten van simpele neurale netwerken kan uitleggen, inzicht in de uitlegbaarheid van generatieve AI.

Uitlegbaarheid in de sociale dienstverlening: voorspelling van bijstandsgerechtigheid met neurale netwerken

Een neuraal netwerk voorspelt dat een burger een verhoogd risico heeft om in de bijstand te belanden. Dit model is getraind op een selectie kenmerken uit een verzameling geanonimiseerde gegevens van bijstandsgerechtigden uit Nederland. De burger wordt hierom bericht door haar gemeente dat ze vanaf heden kan aansluiten bij een sociaal ondersteuningsprogramma en wordt uitgenodigd voor een gesprek. De burger vraagt zich af: waarom word ik aangemerkt? Ze boft, want het AI-systeem kan inzicht geven in deze beslissing met behulp van SHAP-waarden (een tactiek uit de speltheorie) door de volgende factoren te tonen. Een SHAP-waarde vertegenwoordigt de relatieve bijdrage van een factor aan de uitkomst.

Inkomen (+0.40 SHAP)

De inkomsten van de burger schommelen elke maand sterk, wat een belangrijke risicofactor is, en liggen dichtbij het sociale minimum dat voor haar geldt.

Postcode (+0.25 SHAP)

De distributie van verleden bijstandsgerechtigden over verschillende wijken maakt dat bepaalde wijken meer aandacht krijgen.

Tijdelijke dienstverbanden (+0.20 SHAP)

Een patroon wat het model heeft verkregen is dat een hoger aantal tijdelijke dienstverbanden een voorspellende factor is van periodes van werkloosheid.

Partnerschap bekend (-0.15 SHAP)

De burger heeft een geregistreerd partnerschap wat positief bijdraagt aan stabiliteit en weerbaarheid van haar financiële situatie.

De burger krijgt uitgelegd hoe de SHAP-waarden de bijdrage van verschillende factoren aan de voorspelling kwantificeren. Een positieve SHAP-waarde betekent dat een factor bijdraagt aan een hoger risico op behoefte aan bijstaand, terwijl een negatieve SHAP-waarde aangeeft dat een factor juist het risico verlaagt. Daarnaast krijgt de burger inzicht in (de relatieve impact van) de factoren: hoe hoger (of lager) de SHAP-waarde, hoe groter de invloed van die factor op de uiteindelijke voorspelling. Dit helpt bij het begrijpen van welke factoren het meest doorslaggevend zijn in de risicobeoordeling.

In ons fictieve voorbeeld wordt een gevoelig gegeven zoals postcode gebruikt als verklarende factor voor het AI-systeem. Uiteraard speelt het 'wonen in een bepaalde postcode' geen causale rol in het voldoen aan de voorwaarden voor het ontvangen van een uitkering. Een AI-systeem leert echter patronen o.b.v. de beschikbare data. Als er in de historische data van het systeem een statistisch zwaartepunt komt te liggen op bepaalde gebieden en niet op anderen (omdat er simpelweg meer bijstandsgerechtigden wonen), dan leert het AI-systeem een correlatief patroon, wat het zal reproduceren. Postcode is hier dus niet echt een bepalende factor, maar juist wat men een 'proxy' noemt: een benadering voor onderliggende factoren, in ons voorbeeld zaken die samenhangen met de sociaaleconomische situaties van veel bijstandsgerechtigden.

Terug naar de verantwoording van de SHAP-uitleg in het voorbeeld. Voor het gesprek met de burger controleert de inkomensconsulent van de gemeente de uitkomsten van het model en de factoren die van invloed zijn geweest op de uitkomst. De consulent gaat na of de factoren overeenkomen met de dagelijkse praktijk. In het gesprek met de burger legt de consulent uit hoe het model tot het verhoogde risico is gekomen. Zo heeft het model op een schaal van één tot tien het verhoogde risico van de burger een acht gegeven vanuit de bovenstaande factoren. De consulent legt uit dat het verhoogde risico uit het model niet hoeft te betekenen dat de burger ook daadwerkelijk in de bijstand beland. Maar zij stelt dat resultaten hoger dan een ondergrens van zes in het model aannemelijk genoeg zijn om een preventief ondersteuningsprogramma te adviseren.

Hoewel het bij een neuraal netwerk lastiger is om concrete uitkomsten inzichtelijk te maken en SHAP-waarden slechts bij benadering een verklaring bieden, helpt deze informatie de burger om het advies tot het ondersteuningsprogramma beter te begrijpen. Daarnaast krijgt de burger ook praktische handvatten om het risico te verlagen, zoals inzicht in hoe het schommelende inkomen een positief effect heeft en hoe een eventueel loopbaantraject in het ondersteuningsprogramma een significante invloed kunnen hebben op de uitkomst, bijvoorbeeld door bij te dragen aan het verkrijgen van een vast dienstverband. Aan het einde van het ondersteuningsprogramma worden de resultaten van het model opnieuw geëvalueerd. De inkomensconsulent constateert hierbij een positief effect en een verminderd risico, aangezien de burger een baan met een vast contract heeft gevonden en een stabielere inkomstenstroom geniet.

Het mooie aan dit voorbeeld over een simpel neuraal netwerk, is dat het ons een schaalbaar inzicht meegeeft. De bekende AI-systemen van tegenwoordig zoals ChatGPT, Claude en Gemini, de zogeheten Large Language Models, oftewel 'grote taalmodellen', zijn namelijk niet meer dan enorm opgeschaalde neurale netwerken. Het idee om een ander model of een technische meting, zoals de SHAP-waarden, toe te passen op de 'zwarte doos' geldt dus ook voor grote taalmodellen. De 'verklaring bij benadering' die we met uitlegbare AI verkrijgen moet altijd in de context van het gebruik van het AI-systeem gesitueerd worden; als een check of een indicatie voor vervolgonderzoek.

Het uitleggen van de uitkomsten van een neuraal netwerk verhoogt dus de mate van transparantie. Het maakt controlemechanismen mogelijk en het controleren van concrete resultaten. Dit zorgt niet alleen voor betere en navolgbare besluitvorming, maar verhoogt ook het vertrouwen in het systeem doordat uitkomsten met een bepaalde mate van objectiviteit kunnen worden toegelicht. De mens die samenwerkt met het AI-systeem kan daardoor goed verantwoording afleggen over het proces.

Wat kan ik doen als ik aan de slag wil met uitlegbare AI?

Misschien vraag je je nu af: hoe begin ik hieraan? Het proces is niet ingewikkeld, maar vereist wel bewustzijn van een aantal belangrijke aandachtspunten:

Neem verantwoordelijkheid en behoud menselijke controle

AI-systemen mogen nooit ontsnappen aan menselijke controle. Een 'black box' onderdeel uit laten maken van kritieke processen is niet acceptabel. Borg deze verantwoordelijkheid door de eindverantwoordelijkheid bij menselijke beslissers te beleggen en AI-inzichten als advies te presenteren, en niet als bindende besluiten.

Blijven testen en verbeteren

Begin klein, bijvoorbeeld met een proefproject. Begin met een uitleg van de uitkomsten, verzamel feedback, pas aan waar nodig en breid daarna uit. Om tot een goede uitleg te komen, is een continu en iteratief proces benodigd.

Betrekt de juiste mensen

AI raakt veel belanghebbenden: collega's, burgers, beleidsmakers. Luister naar hun vragen en zorgen. Wat willen zij weten over beslissingen gebaseerd op AI? Wat hebben ze nodig om erop te vertrouwen? Welke expertkennis heb jij nodig om uitlegbare AI tot een goed einde te brengen? Betrek daarom mensen met de juiste technische, juridische en ethische vaardigheden.

Kies voor de meest uitlegbare technologie

Uitlegbare AI kan op verschillende niveaus worden toegepast. Dit zal verschillen tussen toepassingen van machine learning en meer complexe systemen zoals neurale netwerken. Het wordt moeilijker om een concrete uitleg te geven naarmate de complexiteit van de technologie toeneemt. Zorg voor een duidelijke doelgerichtheid van de oplossing en werk samen met experts om systemen te selecteren die maximale transparantie en eerlijkheid bieden, afgestemd op het specifieke doel. Er is namelijk geen ongedifferentieerde oplossing voor uitlegbaarheid, omdat er altijd een context-gebonden AI-toepassing wordt uitgelegd aan een bepaalde doelgroep.

Het grotere verhaal: samenbouwen aan vertrouwen

Uitlegbare AI is dus meer dan technologie alleen. Het is een stap richting een eerlijkere en transparantere samenleving waarin iedereen kan profiteren van technologische vooruitgang. Wanneer publieke organisaties de controle behouden over hun AI-gedreven systemen, kunnen zij niet alleen de betrouwbaarheid van technologie waarborgen maar ook het vertrouwen van burgers versterken. Dit betekent verantwoordelijkheid nemen voor beslissingen met een AI-component, niet alleen voor de technologie die men gebruikt, maar ook voor hoe dit aan de mensen die erop vertrouwen wordt uitgelegd.

Highberg heeft uitgebreide ervaring met uitlegbare AI en staat klaar om jou en je organisatie te begeleiden. Of het nu gaat om een eerste kennismaking of een diepgaande implementatie, we denken graag met je mee. Ook zijn we enorm benieuwd naar jouw ervaring met uitlegbare AI. Wij komen graag met je in contact, want samen kunnen we kijken hoe jouw organisatie maatschappelijke vraagstukken met behulp van uitlegbare algoritmes succesvol aan kan pakken.

Auters

Sebastiaan Streng

+31 6 43866482 | sebastiaan.streng@highberg.com

Thomas Mollema

+31 6 42743619 | thomas.mollema@highberg.com