Laat **kunstmatige intelligentie** haar eigen gang gaan

Leren Neurale netwerken

Mensen eisen voorspelbaarheid van kunstmatige intelligentie, maar die presteert het best als ze vrij is om te leren. Haar makers hebben, om hun eigen creatie nog te doorgronden, een nieuwe wetenschap nodig.

tekst **Willem Schoonen**
illustratie **Fadi Nadrous**

Jelle Zuidema zal even laten zien wat hij bedoelt. De taal- en computerwetenschapper aan de Universiteit van Amsterdam gaat naar Google Translate, tikt in 'De dokter heeft haar handtas in het ziekenhuis achtergelaten', en laat dat vertalen naar het Turks. Turks kent hij niet, dus wat er nu op het scherm verschijnt zegt hem niets, maar hij laat het weer terugvertalen naar het Nederlands. Resultaat: 'De dokter heeft zijn tas in het ziekenhuis achtergelaten'.

De vrouwelijke arts is in de omweg langs Turkije een man geworden. Zuidema: "Het woord dokter kan in het Nederlands een vrouw of een man aanduiden. In sommige andere talen is dat niet zo. En dit vertaalalgoritme blijkt dan niet in staat om uit de rest van de zin af te leiden dat het hier om een vrouw gaat."

Nog een voorbeeld. In 2015 ontstond een roemruchte affaire rond Google Photos, toen een computerwetenschapper, Jacky Alciné, liet zien dat foto's van zwarte medemensen die hij uploadde door het algoritme van Google werden geclassificeerd als 'gorilla'. Schaamtevol beloofde Google de fout te herstellen. Drie jaar later bleek dat de fout helemaal niet was hersteld, maar dat Google simpelweg de fotocategorie 'gorilla' had geblokkeerd.

Zuidema: "Dat een wereldleider in slimme algoritmes als Google niet in staat is zo'n fout eruit te halen en dan maar een hele categorie blokkeert, laat zien waar we mee te maken hebben. De algoritmes voor vertalen of voor beeldherkenning zijn heel goed geworden, maar zo complex dat een seksistisch of raciaal vooroordeel er niet zo maar uit te halen is."

Probleem is dat zelfs de makers van die algoritmes niet precies weten hoe die tot hun vertaling of beeldkwalificatie komen. Het zijn neurale netwerken, zelflerende systemen. Die ontwikkelen zelf hun kwaliteiten én hun tekortkomingen.

Er heeft in de wereld van kunstmatige intelligentie lang een debat gewoed over de vraag of de maker van een algoritme moet kunnen uitleggen en garanderen wat zijn machine doet. *Explainable by design* wordt dat genoemd. En de eis duikt onder meer op in discussies over het toelaten van zelfsturende auto's. Wetgevers en publiek willen graag precies weten wat die auto gaat doen, ook als hij voor de onmogelijke keuze komt te staan of hij, uit de bocht vliegend, die voetgangers moet overrijden of zijn inzittenden te pletter moet laten storten.

Explainable by design, dus. En anders komt die auto de weg niet op. Dat is in veel landen uitgangspunt bij wetgeving voor kunstmatige intelligentie. Het is een voor de hand liggende eis, maar daarmee kom je er niet, zegt Zuidema. "Er is maatschappelijke druk en een sterke lobby, maar als je aan kunstmatige intelligentie die eis blijft stellen, worden de systemen gewoon niet goed genoeg."

Noeste arbeid

Een neuraal netwerk dat de vrijheid krijgt om te leren, komt veel verder. Dus laten we accepteren, zegt Zuidema, dat kunstmatige intelligentie een zwarte doos is. "We weten niet precies wat daarin gebeurt, maar we kunnen wel methoden ontwikkelen om dat te onderzoeken. Niet explainable by design, maar *explainable by hard work*."

Hier is een heel nieuwe wetenschap aan het ontstaan, zegt Zuidema. Een wetenschap die kunstmatige intelligentie onderzoekt als was het een levend organisme, of het brein van een mens. En het gaat hard met die wetenschap, zegt Zuidema: "Ik ben midden veertig, maar voel me knap oud als ik zie hoe jonge medewerkers bijna dagelijks met dingen komen die voor mij volstrekt nieuw zijn."

Een van de jonge medewerkers is Dieuwke Hupkes, die net bij Zuidema haar promotieonderzoek heeft afgerond. Hupkes onderzocht of neurale netwerken die zijn getraind om met taal om te gaan, kunnen verklaren hoe taal werkt in het menselijk brein (zie kader). Ze is uit de natuurkunde naar de taalwetenschap gekomen, maar werkt ook met psychologen en hersenwetenschappers.

Hupkes: "De manier waarop neurale netwerken leren is heel goed te vergelijken met hoe wij mensen leren. Om kunstmatige intelligentie te begrijpen, gebruiken we technieken van hersenwetenschappers en psychologen. En omgekeerd gebruiken zij onze kennis van de wiskunde en logica van vertaalsysteem voor hun onderzoek aan het menselijk brein."

Een van die technieken heet Representational Similarity Analysis (RSA), een wiskundige techniek die iets meer dan een decennium geleden werd geïntroduceerd om verschillende representaties te vergelijken, zoals het licht dat op het netvlies valt aan de ene kant en de neurale patronen in hersenen aan de andere. Twee representaties in verschillende ruimtes, met verschillende dimensies. RSA is een techniek om te toetsen hoe goed ze overeenkomen.

Diezelfde techniek, zegt Hupkes, kun je gebruiken om de prestaties van een neuraal netwerk te meten, bijvoorbeeld netwerken die worden gebruikt om automatisch te vertalen. Het is een manier om te bepalen of twee neurale netwerken die onafhankelijk van elkaar zijn getraind tot min of meer dezelfde oplossingen komen voor vertalingen.

> 'Wie weet, levert onderzoek aan ki een goede verklaring van racisme op'

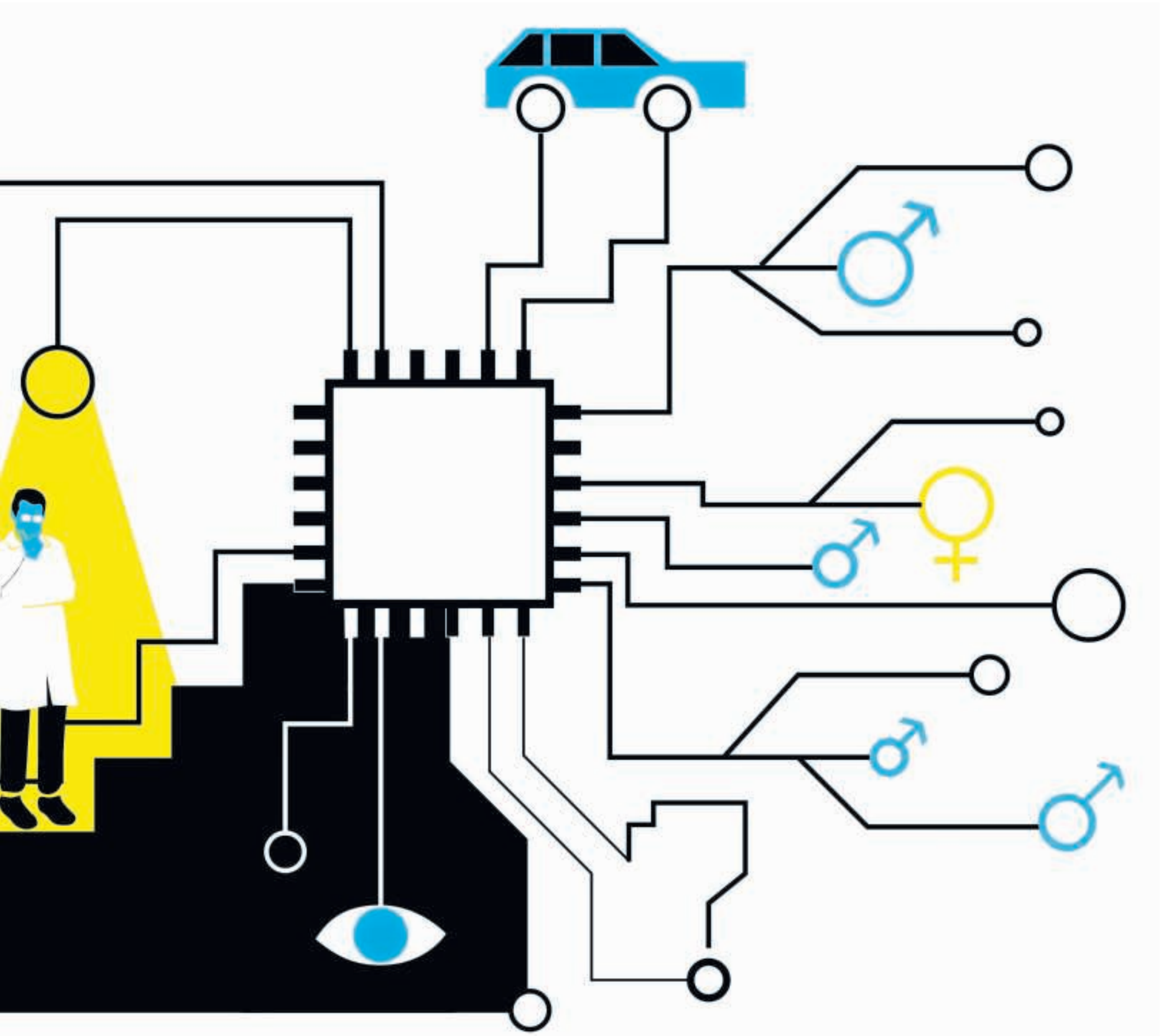
Voor gewone stervelingen blijft het lastig te accepteren dat je een computer die je zelf hebt geprogrammeerd, niet meer kunt volgen en dat je niet gewoon aan een knopje kunt draaien als die computer gorilla's en zwarte mensen door elkaar haalt.

Zuidema: "Je praat bij kunstmatige intelligentie, zelflerende systemen, over enorme aantallen parameters, miljarden soms. Dat zijn er veel te veel om te weten aan welk knopje je moet draaien en om te kunnen voorzien wat er verandert als je ergens aan draait. Je moet bestuderen hoe ze werken en daarna is het proberen en leren om fouten eruit te halen. Net zoals een farmaceut doet bij de ontwikkeling van een medicijn. Hoe virusremmers werken weten we ook niet tot in detail, maar we kunnen wel effectieve virusremmers ontwikkelen."

Bijwerkingen

En die racistische en seksistische vooroordelen van vertaal- en zoekmachines? Moeten die verholpen worden met een medicijn? Zuidema: "Inderdaad. En net als bij menselijke medicijnen krijg je bijwerkingen. Herstel van dat soort fouten is heel complex. Je weet niet precies waar ze vandaan komen. Het zijn vooroordelen net zoals de mens heeft. En hoe menselijke vooroordelen ontstaan weten we ook nog niet."

Hupkes: "We ontdekken wel meer en meer dat de mechanismen vergelijkbaar zijn. Dus wie weet, levert het onderzoek aan kunstmatige intelligentie nog een goede verklaring op van racisme."



Fouten verbeteren zonder les in grammatica

De speler die zijn knie had verdraaid riep de trainer. Als je de woorden uit deze zin los knipt, vertaalt en weer bij elkaar zet, kun je komen te zitten met een speler die roept naar de trainer die zijn knie heeft verdraaid. In werkelijkheid gebeurt dat niet snel, omdat we op school hebben geleerd wat de structuur is van de zin. De speler is het onderwerp, de trainer lijdend voorwerp en de verdraaide knie is een betrekkelijke bijzin bij speler.

Die verschillende grammaticale categorieën waaruit de zin is opgebouwd, zou je kunnen weergeven met een symbool. Dat is precies wat er in de eerste vormen van kunstmatige intelligentie gebeurde: de computer kreeg de grammaticale kennis van de mens geprogrammeerd in de vorm van symbolen.

De kunstmatige intelligentie van vandaag kent geen symbolen. Die heeft geen idee van onderwerp, gezegde of lijdend voorwerp. Het zijn neurale netwerken die, bijvoorbeeld om iets te vertalen van Nederlands naar Engels, worden getraind met enorme hoeveelheden tekst. Ze leren van hun fouten en worden behoorlijk goede vertalers, zonder ooit een les grammatica gehad te hebben.

Het is voor taalkenners misschien teleurstellend om te zien dat die neurale netwerken steeds beter worden zonder dat ze de hulp en kennis van menselijke experts nodig hebben, schrijft Dieuwke Hupkes in het proefschrift waarop ze onlangs promoveerde aan de Universiteit van Amsterdam, maar het maakt die neurale netwerken voor taalwetenschappers juist bijzonder interessant.

Want wij hebben in de schoolbanken wel keurig leren zinsontleden, maar in onze hersenen gaat dat niet zo. Er zijn geen aparte neuronen voor onderwerp, gezegde en lijdend voorwerp. De hersenen zijn als die neurale netwerken; ze hebben geleerd met taal om te gaan zonder weet te hebben van categorieën of symbolen.

Wetenschappers als Hupkes willen nu weten of neurale netwerken iets kunnen zeggen over de taalvermogens van de mens, of neurale netwerken die hebben geleerd om met taal om te gaan een goed model zijn voor taalverwerking in het brein. Als dat zo is, dan kun je kunstmatige intelligentie gebruiken om te onderzoeken hoe dingen werken in de menselijke hersenen en wat daar fout kan gaan.

> Kunstmatige intelligentie is net echt

Het is een nieuwe tak in de wetenschap van kunstmatige intelligentie, maar Hupkes' voorzichtige conclusie is dat vertaalmachines inderdaad een goed model zijn voor taakverwerking in het brein. Niet alleen is hun architectuur vergelijkbaar – het zijn allebei neurale netwerken – maar ook hun omgang met taal.

Om dat te achterhalen moeten wetenschappers diep in die neurale netwerken duiken om te ontdekken hoe ze nu eigenlijk werken. En dat is ingewikkeld omdat die netwerken getraind zijn met gigantische datasets en ontelbare parameters hebben (zie hoofdverhaal).

Om een indruk te krijgen hoe dat onderzoek werkt, helpt de zin waarmee dit kader begon: De speler die zijn knie had verdraaid riep de trainer.

Met zijn grammaticale kennis kan de mens die zin weergeven als een boom. In die boom zit bij speler een zijtakje met de verdraaide knie, terwijl de stam doorgaat naar de trainer. In een neuraal netwerk wordt ieder woord van die zin weergeven in de vorm van een – unieke – rij getallen. Een wiskundige kan met die getallen ieder woord zijn plek geven in een ruimte. Dat kan in een diagram met maar twee dimensies, maar voor hetzelfde geld in een ruimte met honderden dimensies.

Die boom duikt daarin niet ineens op, maar de wiskunde kan zoeken naar iets wat erop lijkt. En dan blijkt in die ruimte de verdraaide knie dichter bij de speler te liggen dan bij de trainer. Dus zonder iets te weten van onderwerp en lijdend voorwerp, brengt het neurale netwerk wel de juiste hiërarchie aan in de zin.