
BRAM BELLONI

‘Een AI-systeem moet kunnen zeggen: dat is geen goed idee’

Hoe meer we uitbesteden aan AI, hoe belangrijker het is dat we AI-systemen kunnen vertrouwen. AI-onderzoeker Pınar Yolum stelt dat betrouwbare AI-systemen bezwaar moeten kunnen maken tegen opdrachten en dat ze het moeten zeggen als ze iets niet weten.

Tekst: Fenna van der Grient

Als we in de toekomst steeds grotere taken willen uitbesteden aan kunstmatige intelligentie (AI), dan moeten we ervoor zorgen dat deze systemen betrouwbaar zijn. Volgens Pınar Yolum, hoogleraar betrouwbare AI aan de Universiteit Utrecht, betekent dat dat AI-systemen onder andere moeten kunnen aangeven dat ze iets niet weten. ‘Neem ChatGPT. Daar krijg je soms heel goede antwoorden van, maar soms ook incorrecte antwoorden. Als een systeem zegt ‘ik weet het niet’ in plaats van dat het een fout antwoord geeft, dan creëert dat betrouwbaarheid.’

Wat moet een AI-systeem nog meer kunnen?
‘Een AI-systeem moet kunnen laten zien welke denkstappen het heeft gemaakt om tot een bepaald antwoord te komen. Daarnaast moet het systeem taken autonoom kunnen uitvoeren, en in onverwachte situaties bij je terugkomen en vragen: hoe wil je dat ik dit verder aanpak? Het moet ook in staat zijn om te accepteren dat andere systemen uitkomsten geven die beter zijn dan de eigen uitkomst. In die gevallen moet het systeem een stap terug kunnen doen. Tot slot moet een AI-systeem kunnen weigeren of weerstand kunnen bieden tegen opdrachten die indruisen tegen de normen en waarden die het systeem zijn aangeleerd. Als mens kunnen we alles aan AI vragen, zoals het maken van deepfakes om politieke tegenstanders in diskrediet te brengen. Maar stel dat een AI-systeem kan zeggen: ‘Ik denk dat het niet goed is om dit te doen, om deze redenen.’ Dan zou dat zo’n systeem in mijn ogen betrouwbaarder maken.’

Zo’n systeem zou dan voor degene die deze dingen vraagt juist onbetrouwbaarder zijn. Hoe beslis je bij tegenstrijdige belangen welke morele principes het systeem moet hanteren?
‘Uiteindelijk is vertrouwen inderdaad subjectief. De mensen die jij vertrouwt, zijn misschien niet de mensen die ik vertrouw. Verschillende mensen zullen verschillende AI-systemen bouwen en vertrouwen, omdat zij normen en waarden al dan niet expliciet in de systemen inbouwen. Er zullen AI-systemen zijn die welwillender zijn dan andere. Maar ik denk dat het desondanks belangrijk is om het technisch mogelijk te maken dat ze weigeren. Dan kunnen we er tegelijkertijd goed over nadenken in wat voor gevallen we willen dat die functionaliteit er is.’

Bestaan er objectieve tests of criteria die kunnen bepalen of een AI-systeem betrouwbaar is?
‘We hebben op dit moment nog geen goede maatstaven om te zeggen: dit AI-systeem is betrouwbaarder dan dat andere. Die maatstaven proberen we te ontwikkelen. Wel kun je de prestaties van een AI-systeem meten. Dan kijk je bijvoorbeeld naar of het systeem goed handelt in een aantal vooraf gedefinieerde situaties. Je ziet dan dat het heel moeilijk is om vertrouwen op te bouwen, maar heel makkelijk om het af te breken. Vertrouwen komt te voet en gaat te paard. Dat is bij mensen net zo. Als een systeem in acht van de tien testscenario’s goed handelt, maar in twee niet, dan zijn die twee gevallen dodelijk voor de betrouwbaarheid. Wij als onderzoekers moeten methoden ontwikkelen om een AI-systeem betrouwbaar te maken. Die kunnen we dan aan ontwikkelaars van AI-systemen geven, zodat zij ze kunnen toepassen.’

> ‘We moeten methoden ontwikkelen om AI betrouwbaar te maken’

Maar kunnen we er wel op vertrouwen dat ontwikkelaars die methoden toepassen als ze voor bedrijven werken met belangen die soms indruisen tegen die van de gebruikers?
‘Ja, dat is altijd een probleem. In een ideale wereld leggen we in regelgeving vast dat ontwikkelaars die methoden moeten toepassen, en handhaven we die regels. Als je een auto koopt, denk je er niet over na of de fabrikant aan elke veiligheidsstandaard heeft voldaan. Je vertrouwt niet per se de auto zelf, of de autoverkoper, maar het hele ecosysteem: de auto-industrie en de regels die daarin gelden. Uiteindelijk zou ik dat ook willen zien voor AI-systemen.’

Kunnen we in de toekomst grote taken delegeren aan AI, zonder menselijke controle?
‘Ja, ik denk dat dat kan. Als we de juiste eigenschappen hebben ingebouwd, als we het systeem hebben getest, als we weten waar de grenzen liggen, dan is er geen reden om het systeem niet te vertrouwen. Het systeem gaat niet zomaar iets geks doen. Natuurlijk is er nog een hoop onderzoek nodig: hoe gaan we al deze ideeën technisch uitvoeren? En hoe gaan we bijvoorbeeld een zelflerend systeem testen? Maar ik denk dat het in de toekomst mogelijk gaat zijn.’ ■