Johannes Fahrenfort • cognitief neurowetenschapper aan de Vrije Universiteit en de Universiteit van Amsterdam BEELD AP

Robot moet voorspelbaar blijven

…De angst voor kunstmatige intelligentie die ons overvleugelt, is reeel. Dan moeten we op tijd de stekker er nog uit kunnen trekken.

De Chinees Ke Jie, de nummer 1 van de wereld in het eeuwenoude spel Go, kijkt niet blij als hij voor de tweede keer door een computer wordt verslagen, in mei van dit jaar.

In 1996 slaagde voor het eerst een schaakcomputer erin de regerend wereldkampioen, Garry Kasparov, te verslaan. Deze computer rekende simpelweg alle mogelijke zetten door en won – op basis van pure rekenkracht.
Tot voor kort waren mensen nog superieur in Go, een duizenden jaren oud bordspel uit China: dat kent simpelweg te veel mogelijkheden om al-

Een zelfrijdende auto moet soms de regels negeren. Maar wat gaat hij dan doen?

lemaal door te rekenen. Sinds kort echter kunnen de nieuwste artificiële netwerken die gebruikmaken van deep learning , met gemak zelfs de beste Go-speler ter wereld verslaan.
Sommigen redeneren dat we hiermee ons voordeel moeten doen: geef Kunstmatige Intelligentie (AI) het vermogen mensen beter te gaan begrijpen, dan kan ze ons nog beter van dienst zijn. Het idee is dat AI dan problemen kan oplossen die we zelf anders niet opgelost krijgen.
AI neemt ons ontegenzeggelijk gemakkelijke taken uit handen en levert daarmee een zinvolle bijdrage aan ons dagelijks bestaan. We moeten echter wel degelijk oppassen welke taken we in eigen hand houden, zeker als we niet precies meer begrijpen hoe AI tot oplossingen komt.

kaders oprekken

Angst voor het idee dat AI intelligenter wordt dan wij, wordt soms weggeredeneerd met het idee dat mensen altijd bang zijn voor nieuwe technologie (de stoomtrein, het vliegtuig, de computer). De uitdaging is dan de technologie te perfectioneren, zodat die ongevaarlijk wordt.
Het probleem is echter dat de kaders waarbinnen AI zich beweegt, opgerekt moeten worden om AI effectief te laten zijn. Een voorbeeld: zelfrijdende auto's wordt geleerd tussen de doorgetrokken strepen van de snelweg te rijden, om zo een ongeluk te voorkomen. Een kunstenaar verfde onlangs een doorgetrokken streep in een cirkel om een geparkeerde zelfrijdende auto heen, waardoor deze niet meer kon wegrijden.
Om de AI van de auto effectief te laten zijn, zal haar (het?) dus geleerd moeten worden dat de regels rekbaar zijn. Als er een mens op de snelweg springt en er is niet genoeg tijd om te remmen, mag/moet de auto wel degelijk over de doorgetrokken streep rijden om een botsing te voorkomen. (Ingewikkelde morele keuzes tussen het redden van de inzittenden of van de mensen op de snelweg, laat ik nog maar even buiten beschouwing.)

Geavanceerde AI gebruikt de regels dus als richtlijn en niet als wet – net als mensen. Het probleem met buigzame regels is dat ze beide kanten op kunnen buigen. Sommige mensen buigen de regels om levens te redden, anderen om iemand om het leven te brengen. Hetzelfde zal gebeuren bij AI. Hoe beter AI wordt, hoe moeilijker het wordt te voorspellen wat ze doet of wat haar motieven zijn. Deze eigenschap is onlosmakelijk verbonden met intelligentie zelf, in tegenstelling tot de veiligheid van stoomtreinen, vliegtuigen en computers – die al hun kracht ontlenen aan voorspelbaarheid.
Hier zit de angel. De beroemde sf-schrijver Isaac Asimov voorzag dit al en legde robots de drie wetten der robotica op, waarvan de eerste luidt: 'Een robot mag een mens geen letsel toebrengen, of door niet te handelen toestaan dat een mens letsel oploopt.' Maar hoe beter de AI, hoe minder 'hard' zij deze regel kan hanteren, en hoe moeilijker het wordt het motivatie- en waardensysteem van de AI te doorgronden.

beurskrach

Kunstmatige intelligentie die we kunnen begrijpen en waarvan we het gedrag kunnen voorspellen, is niet gevaarlijk. Maar kunstmatige intelligentie die ons begrijpt terwijl wij het zelf maar matig begrijpen, is het summum van gevaar – een gevaar dat we niet moeten bagatelliseren.
De uitdaging is ons hiervan terdege bewust te zijn en op tijd wetgeving te ontwikkelen. Het feit dat computers bijvoorbeeld de gelegenheid wordt gegeven mee te spelen op de beurs, is gevaarlijk als we niet precies weten wat die dingen doen. Ze zouden een beurskrach kunnen veroorzaken als dat winst oplevert, of nog slinkser te werk kunnen gaan door op een voor ons ondoorgrondelijke manier koersen te manipuleren.
Voor sommige problemen kan AI heel nuttig zijn, zoals het doorrekenen van efficiënte transportroutes, verbetering van landbouwtechnieken etcetera. Maar zodra een machine haar eigen motivatie ontwikkelt, is het belangrijk goed te doorgronden wat die motivatie is en wat voor invloed dat op onze levens heeft. En als we dat niet precies kunnen duiden, is het misschien beter de stekker eruit te trekken. <