Robot in lotushouding. © Getty Images for The New Yorker

Het summum van gevaar: onvoorspelbare kunstmatige intelligentie

Artificial Intelligence

ARTIKEL Hoe beter AI wordt, hoe moeilijker het wordt om te voorspellen wat ze doet of wat haar motieven zijn.

Door: Johannes Fahrenfort 20 november 2017, 02:00

Professor Bronkhorst betoogt dat we niet bang moeten zijn voor kunstmatige intelligentie (AI), maar het voor ons moeten laten werken (O&D, 16 november). Hij redeneert dat mensen altijd bang zijn voor nieuwe technologie (eerst de stoomtrein, toen het vliegtuig, toen de computer), en dat de uitdaging eruit bestaat de technologie te perfectioneren zodat die ongevaarlijk wordt.

De nieuwste artificiële netwerken die gebruikmaken van deep learning kunnen met gemak de beste go-speler ter wereld verslaan, en we zouden hier ons voordeel mee moeten doen door ze het vermogen te geven ons beter te begrijpen en ons zo beter van dienst te kunnen zijn.

Rekbare regels

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 verkeersstrepen 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 (waarbij ik ingewikkelde morele keuzes tussen het redden van de inzittenden of het redden van de mensen op de snelweg nog maar even buiten beschouwing laat).

> Het probleem met buigzame regels is dat die beide kanten op kunnen buigen

Onvoorspelbaar

Johannes Fahrenfort is cognitief neurowetenschapper aan de VU en de UvA. ©

Goede AI gebruikt de regels dus als een richtlijn en niet als een wet (net als mensen). Het probleem met buigzame regels is dat die beide kanten op kunnen buigen. Sommige mensen buigen de regels om levens te redden, anderen buigen de regels om iemand om het leven te brengen. Hetzelfde zal gebeuren bij AI. Hoe beter AI wordt, hoe moeilijker het wordt om te voorspellen wat ze doet of wat haar motieven zijn. Dit probleem is intrinsiek aan intelligentie zelf, in tegenstelling tot de veiligheid van stoomtreinen, vliegtuigen en computers (die alle hun kracht ontlenen aan voorspelbaarheid).

Hier zit dus de angel. De beroemde sf-schrijver 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 zoals zojuist betoogd: hoe beter de AI, hoe minder 'hard' die een dergelijke regel kan hanteren om problemen op te lossen, en hoe moeilijker het wordt om het motivatie- en waardensysteem van de AI te doorgronden. Kunstmatige intelligentie die ons begrijpt maar die wij zelf maar matig begrijpen is het summum van gevaar, een gevaar dat een hoogleraar op het gebied van AI niet zou moeten bagatelliseren.

Johannes Fahrenfort doet onderzoek naar de neurale processen die ten grondslag liggen aan bewustzijn.

Volg en lees meer over:
TECH | OPINIE

AANBEVOLEN ARTIKELEN

Wie wint de grote dopingshow. Degene die in elkaar zakt, of degene die blijft staan?
20 november 2017

Waarom moet ik geld uitgeven om spammers te stoppen?
20 november 2017