Verbetert de wereld met AI?

Wouter Beek, Master of Logic student aan de UvA
(wbeek@science.uva.nl)

In het jaar van onze Heer 1941, waarin de Tweede Wereldoorlog de vrije wereld in haar greep hield en in Duitsland de eerste concentratiekampen in gebruik werden genomen, beschreef Isaac Asimov de drie wetten waar robots zich aan zouden moeten houden om het bestaan van de menselijke soort veilig te kunnen stellen. De drie wetten kwamen voor in het verhaal 'Runaround', wat in 1942 verscheen in het tijdschrift Astounding Science Fiction. Men zou denken dat ten tijde van het verschijnen van dit verhaal nijpendere ethische kwesties denkbaar waren, zeker gezien het feit dat er nog helemaal geen robots waren waarvoor deze wetten bedoeld zouden kunnen zijn geweest. Wat dreef Asimov ertoe om zich te verdiepen in de morele wetten van entiteiten die op dat moment niet eens bestonden? Waarom richtte hij zich op het niet-bestaande; het fictieve?

Positronische robots

De wetten van Asimov zijn bedoeld voor zogenaamde *positronische* robots. Dit zijn robots die beschikken over een eigen bewustzijn. De wetten zijn dus al evenmin van toepassing op hedendaagse robots, die een dergelijk niveau van intelligentie nog niet hebben gereikt. Dat deze positronische robots nog steeds niet bestaan betekent niet dat er daarom ook minder over hun eventuele invloed op de mensheid wordt gespeculeerd. Volgens vele meer en minder gerenommeerde AI-onderzoekers is deze discussie zeer relevant voor de (nabije) toekomst van de mens. Want het zal niet lang meer duren, zo denken zij, voordat er daadwerkelijk robots met zulke complexe cognitieve functies zullen worden gefabriceerd.

Dit soort van voorspellingen, over het in de nabije toekomst ontstaan van zelfbewuste robots, worden al sinds de jaren '60 veelvuldig door mensen werkzaam in en rondom de AI gemaakt. Zo was het idee achter de HAL9000 supercomputer in Stanley Kubrick's *2001: A Space Odyssey*, dat complexe menselijke vaardigheden zoals het verstaan, interpreteren en spreken van natuurlijke taal (en zelfs liplezen) in het jaar 2001 geheel door computers zouden worden beheerst. Overigens waren niet alle vaardigheden van HAL louter overmoedige fantasieën, want enige andere van zijn eigenschappen zijn inmiddels wel gerealiseerd, zoals bijvoorbeeld het schaken. Ook zijn er bepaalde eigenschappen, zoals gezichtsherkenning, die in ieder geval gedeeltelijk gerealiseerd zijn. Maar de meest opmerkelijke eigenschap van HAL is wel dat hij emoties bezat. Deze kwamen tot uiting toen hij langzaam werd uitgeschakeld en bang was: "I am afraid, Dave."

Sterke en zwakke AI

Deze positie, dat er in de toekomst robots kunnen worden ontwikkeld die zoiets als een zelfbewustzijn bezitten, kan men vereenzelvigen met de gedachte van *strong AI*, waarin geclaimd wordt dat het onderzoek in de AI tot artefacten met echte intelligentie kan leiden. Alle aspecten van de menselijke geest zouden volgens dit idee in een kunstmatige entiteit na te maken zijn. De gedachte dat alle denkprocessen bestaan uit mechanische processen is natuurlijk al veel ouder en vindt haar oorsprong in het materialisme en *corpuscularisme*[1] van zeventiendeeeuwse natuurfilosofen zoals Francis Bacon en Thomas Hobbes.

Tegenover deze 'sterke' visie staat het idee van de zogenaamde *weak AI*, volgens dewelke AI-technieken enkel tot simulaties van cognitieve processen van de mens kunnen leiden. De aldus vervaardigde entiteiten bezitten zelf geen bewustzijn. Indien de zogenaamde 'zwakke' visie van AI juist is, zullen er nooit positronische robots à la Asimov ontstaan. Zijn de wetten van Asimov dan geheel nutteloos indien het sterke programma onjuist blijkt te zijn, en robots dus nooit over een bewustzijn zullen kunnen beschikken? In het geheel niet. Ook al beperkt Asimov in zijn verhaal de toepassing van de wetten op positronische robots, het is toch zeer goed denkbaar dat deze regels in niet-bewuste robots worden opgenomen om de gevolgen van eventuele fouten in hun programmatuur te kunnen beperken. Een niet-bewuste robot die een fout maakt kan namelijk op dezelfde wijze ongewenst gedag veroorzaken als een bewuste robot die naar eigen inzicht handelt.

Singulariteit

Laten we eens verder ingaan op de claims van enige aanhangers van de 'sterke' visie van AI. Dit is immers de variant waarop de wetten van Asimov origineel gericht waren. Op dit moment zijn er nog niet veel aanwijzingen dat het ooit mogelijk zal zijn om daadwerkelijk positronische robots te vervaardigen. Er zijn veel onderzoekers die echter claimen dat de ontwikkeling van een zelf-bewuste robot op korte termijn zal worden gerealiseerd. Zij gaan vaak nog verder, door aan te geven dat deze robots de mens uiteindelijk qua intelligentie zullen voorbijstreven. Het bewijs voor deze voorspelling halen zij uit bepaalde wetmatigheden die aan technologisch onderzoek ten grondslag zouden moeten liggen. Ze voorzien in de nabije toekomst een punt waarop de robot de mens zal gaan 'inhalen'. Dit punt wordt aangeduid met de term 'singulariteit'. Een van de meest fervente publicisten aangaande het idee van de singulariteit is Raymond Kurzweil. In zijn in 2005 gepubliceerde boek *The Singularity Is Near: When Humans Transcend Biology*, beschrijft Kur-

1. Volgens het corpuscularisme komen de verschillende stoffen voort uit het aantal, de beweging en de positie van de primaire deeltjes: de corpuscules (red.)

zweil de zogenaamde 'Law of Accelerating Returns'. De wet stelt dat de technologische vooruitgang een exponentiële groei doormaakt. Wanneer de technologische ontwikkelingen elkaar steeds sneller zullen opvolgen, zal er uiteindelijk een punt komen waarop de techniek ver genoeg is om een positronische robot te ontwikkelen. Dit is dan het moment van de singulariteit. Daarna zullen deze robots zich echter blijven doorontwikkelen; hun intelligentie zal dan al zeer spoedig die van de mens overstijgen en overschaduwen. Deze wet wordt door Kurzweil ondersteund met de wet van Moore die stelt dat de complexiteit van geïntegreerde halfgeleider-circuits exponentieel groeit. Het karakter van exponentiële groei zou ons er dan toe moeten brengen om snel de wetten van Asimov te implementeren, want het ogenblik van de singulariteit staat voor de deur. Kurzweil interpoleert de groei van technologische kennis en komt dan uit in het jaar 2045, wat toch redelijk binnenkort is. (In het scenario van de singulariteit zouden de wetten van Asimov overigens misschien niet eens meer zinvol zijn. Dit aangezien de cognitie van de machine die van de mens op zulk een exorbitante wijze zal voorbijstreven dat het in staat zal zijn om toepassingen van deze wetten te bedenken waarop de mens nooit zou kunnen anticiperen.)

Moore's Law

Maar laten we eens wat beter kijken naar de wet van Moore, die aan Kurzweil's wet ten grondslag ligt. Deze wet is gebaseerd op een aantal uitspraken van Gordon Moore (mede-oprichter van *Intel*). Op verschillende momenten in zijn carrière heeft hij verschillende (onzorgvuldig gedefinieerde) claims aangaande de exponentiële groei van de complexiteit van geïntegreerde circuits gemaakt. Deze prognoses baseerde hij op de door hem gemaakte observatie dat er voorlopig geen principiële belemmeringen waren die het plaatsen van meerdere componenten op hetzelfde processor-oppervlak zouden kunnen belemmeren. Naarmate *Moore's Law* in haar onduidelijke vorm meer bekendheid verwierf, werd de wet op nog meer verschillende wijzen geïnterpreteerd. Zij zou ook op gaan voor de omvang van harde schijven, geheugenmodules en het aantal pixels in vergelijking met de prijs van beeldschermen.[2] Van belang bij al deze uiteenlopende wetten van Moore is dat de exponentiële toename enkel mogelijk is wanneer er voorlopig geen principiële belemmeringen zijn die de toename in capaciteit tegenhouden. Zulke principiële belemmeringen zijn op langere duur echter zeer reëel, aangezien er een duidelijke limiet is aan de mogelijke transistor-grootte, die niet voorbij de grootte van een atoom kan gaan (zoals door Moore zelf geobserveerd).

Kurzweil volgt echter een geheel andere redeneertrant. Volgens hem zal er tegen de tijd dat een fundamentele limiet wordt bereikt een nieuwe technologie worden uitgevonden die de exponentiële groei mogelijk blijft maken. Daar waar de gedachte van Moore nog was dat een exponentiële groei kon optreden door de afwezigheid van nabije limieten, is de gedachte van Kurzweil precies de tegenovergestelde: doordat exponentiële groei moet plaatsvinden zullen eventuele limieten worden overwonnen. Het is niet duidelijk waarom er altijd sprake moet zijn van een ongelimiteerde exponentiële groei. En zelfs wanneer de groei van technologische ontwikkelingen exponentieel blijft toenemen, dan is het toch nog geheel onduidelijk waarom een louter kwantitatieve groei automatisch aanleiding zou moeten geven tot kwalitatieve verschillen in de cognitieve capaciteiten van robots. Er zijn al een heleboel gebieden waarop de AI de cognitieve capaciteiten van de mens voorbijstreeft, denk bijvoorbeeld aan schaken en rekenen. Wanneer de technologische ontwikkelingen toenemen zullen robots nog sneller berekeningen kunnen maken en nog beter kunnen schaken. Maar of ze daarmee ook het gebied van het zelfbewustzijn binnentreden is zeer onwaarschijnlijk.

Calculus van de ethiek

Zullen robots dan nooit bovenmenselijk worden? Het ligt erg aan de definitie van 'bovenmenselijk'. Indien het alleen om de rekenkundige facetten van het denken gaat, is de zaak snel besloten: machines zijn in dit opzicht superieur aan de mens. Hetzelfde geldt voor andere deelgebieden van de cognitie waar AI-technieken met succes zijn toegepast, zoals in het bovengenoemde schaken, maar ook in het vervaardigen van logistieke planningen en het assembleren van auto's. Wanneer de bovenmenselijke intelligentie van de robot zich nog tot enige andere gebieden zou uitstrekken, zoals het bewijzen van wiskundige stellingen (iets wat overigens al steeds meer aan het gebeuren is), blijft de relatie tussen mens en machine vrijwel onveranderd. Het verschil gaat pas optreden wanneer machines in staat zijn om bovenmenselijke intelligentie te bereiken in gebieden die met politieke, sociale en ethische keuzes te maken hebben. We hebben het hier dan over wat ik het handelingsgerichte denken zou willen noemen. De essentie van het handelingsgerichte denken bestaat uit een vorm van cognitie die veel minder makkelijk in regels lijkt te vatten dan de zonet besproken processen. Het is wellicht ook geen toeval dat er geen uitgebreid formeel vakgebied bestaat dat zich bezighoudt met politieke beslissingen. Er is geen standaard-calculus voor ethisch handelen. Er bestaan weliswaar verschillende deontisch logische systemen[3], maar deze berusten op een groot aantal intuïties omtrent wat wel en niet een ethische aanname mag heten en welke zaken wel en niet wenselijk afleidbaar zijn. Je zou kunnen zeggen dat

2. Voor een historisch overzicht van de ontwikkeling van deze misinterpretaties van de Wet van Moore, zie het artikel 'The Lives and Death of Moore's Law' van Ikka Tuomi. (Online beschikbaar op http://www.firstmonday.org/issues/issue7_11/tuomi/)

3. De deontische logica houdt zich bezig met verplichtingen, permissies en gerelateerde concepten. (Red.)

een ethische theorie een ethisch gevoel vooronderstelt bij degene die de ethische theorie opstelt en aanwendt. Wanneer de machine zou komen te beschikken over handelingsgerichte cognitieve vaardigheden, zou het ook beschikken over een aantal eigenschappen die traditioneel niet in computersystemen kunnen worden aangetroffen. Schaalvergroting alleen brengt deze cognitieve functies waarschijnlijk niet tot stand. Meer geheugen of meer rekenkracht kan zulke lastige redeneerprocessen uiteraard ondersteunen en tot dienst zijn, maar zij lijken hier echter geen volledige basis voor te vormen. Wat een machine nodig lijkt te hebben om handelingsgerichte cognitieve activiteiten te voltrekken is een bepaalde houding, bepaalde overtuigingen, een bepaalde levenswijze; een dispositie om te handelen. Een intuïtie van wat goed en slecht is, wat wenselijk is en wat niet. Pas wanneer deze set van basisvoorwaarden in min of meer heldere mate is gedefinieerd, kan een calculus de overbrugging tussen het doel en de huidige toestand bewerkstelligen.

Het materialisme en corpuscularisme hebben een calculerende denkrichting tot stand gebracht die de niet direct tot de natuur te herleiden aspecten van het leven, voorheen in de scholastieke traditie nog nauw verbonden met de dagelijkse wetenschapsbeoefening, heeft afgezonderd van de natuurwetenschappelijke praktijk. Het is daarom niet toevallig dat juist de handelingsgerichte deelgebieden van de cognitie in de moderne wetenschapsbeoefening onbehandeld blijven. Het is de aanname dat alle vormen van cognitie, dus ook het handelingsgerichte denken, uit mechanische operaties zouden bestaan, die de denkbeelden van de sterke AI en de singulariteit ondersteunen. Wanneer dit het geval zou zijn en het handelingsgerichte denken daadwerkelijk uit louter mechanische bewegingen zou bestaan, dan zou de toenemende kracht van computationele apparatuur ons toch een toenemend inzicht in ons handelingsgerichte denken moeten verschaffen? Dit is echter niet het geval. Ons ethisch bewustzijn is nog niet verruimd door betere kennis omtrent computationaliteit, of door de beschikbaarheid van geavanceerdere apparatuur. In de geschiedenis van de AI is misschien niet één techniek aan te wijzen die ons meer inzicht in het ethisch handelen heeft gegeven. Het is derhalve zeer onwaarschijnlijk dat er een essentieel verband bestaat tussen technologische vooruitgang en ethische vooruitgang. Want als het verband er zou zijn, dan zou het ook zichtbaar moeten zijn. Om ons heen zien we echter dat de techniek vooruitgang boekt, terwijl de ethiek op gelijk niveau achterblijft.

Indien de aanhangers van ideeën zoals 'sterke AI' en 'singulariteit' daadwerkelijk een menselijke of bovenmenselijke robot willen creëren, zouden zij zich dan ook beter op een formele theorie van de ethiek kunnen richten in plaats van de verdere exponentiële groei van computatie na te streven. ø

Agenda

Vrijdag 20 april 2007
[L.E.J. Brouwer Colloquium]
20 April a.s. organiseert de OZSL in samenwerking met de VvL en de Heyting Stichting een colloquium over het leven en werk van L.E.J. Brouwer. Brouwer wordt gezien als de stichter van de stroming binnen de wiskunde en logika die bekend is onder de naam intuïtionisme.
Voor meer informatie: http://ozsl.uu.nl/brouwer/

Dinsdag 8 mei 2007
[Imaging & Genetics in Cognition]
8 tot 11 mei zal in Amsterdam het congres "Integrating Imaging and Genetics in Cognitive Research" gehouden worden. Internationaal bekende wetenschappers zullen daar hun nieuwste onderzoeksresultaten presenteren op het gebied van brain imaging en genetica.
Voor meer informatie: http://www.imaginggenetics.org/

Vrijdag 1 juni 2007
[Symposium: 'Logic and Cognition']
Op vrijdag 1 juni 2007 zal in Groningen een symposium over logica en cognitie plaatsvinden. Verschillende sprekers zullen het deze dag onder andere hebben over logisch redeneren in het brein, formele modellen voor taalverwerving en representaties in robotica.

Donderdag 14 juni 2007
[Symposium: 'P=NP?']
Het jaarlijkse symposium van USCKI Incognito zal dit jaar over de beroemde vraag "P = NP" gaan. Verschillende sprekers zullen op de diepere betekenis van deze fundamentele vraag ingaan en interessante toepassingen laten zien van hoe de logische complexiteit van een probleem juist een voordeel kan zijn.
Voor meer informatie: http://symposium.uscki.nl/

Vrijdag 22 juni 2007
[Student Conference]
De nieuwe Nederlandse Studenten Vereniging voor KI (NSVKI) organiseert op 22 juni in Nijmegen een conferentie, waarbij studenten hun eigen werk kunnen presenteren aan mede-studenten. De deadline voor het insturen van artikelen is 1 mei 2007.
Voor meer informatie: http://www.nsvki.nl/sc

Ook een agendapunt voor De Connectie? Mail ons!
redactie@connectie.org