Machines die denken

Invloedrijke denkers over de komst van kunstmatige intelligentie

Samengesteld door John Brockman

Vertaling onder redactie van Frits van der Waa en Henny Corver

Vertaald door
Yolande Samwel, Wilma Paalman, Wilma Chappin-Berndsen, Willie van der Kuil, Willem van Paassen, Tracey Drost-Plegt, Stina de Graaf, Ronald Jonkers, René van Veen, Pon Ruiter, Piet Dal, Peter van Nieuwkoop, Paul Heijman, Patty Adelaar, Nicole Seegers, Nico Groen, Nannie de Nijs Bik-Plasman, Mieke Groot, Miebeth van Horn, Merel Leene, Marianne Hoogenboom, Loretta Dijk, Klaske Kamstra, Joris Vermeulen, Joost Mulder, Jolande van der Klis, Jeske Nelissen, Jeroen Duivenvoorden, Jan Willem Reitsma, Jan van den Berg, Jaap Verschoor, Ineke Veldink, Ineke van den Elskamp, Herman te Loo, Henny Corver, Frits van der Waa, Frank Lekens, Floris Blommaert, Erica Feberwee, Eefje Bosch, Dieuwke van der Veen, Catalien van Paassen, Bernadette Delfgaauw, Arjanne van Luipen, Annemie de Vries, Agaath Diemel

MAVEN PUBLISHING

Inleiding: De Edge-vraag

John Brockman
Uitgever en redacteur *Edge*

De laatste jaren heeft het in de jaren tachtig op gang gekomen filosofische debat over artificiële intelligentie (AI) – of computers 'echt' kunnen denken, bewustzijn hebben, enzovoort – geleid tot een nieuwe discussie, namelijk over de vraag hoe we moeten omgaan met de vormen van AI die volgens veel mensen al in gebruik zijn. Als zulke AI's 'superintelligent' worden (een term ontleend aan Nick Bostroms boek *Superintelligence* uit 2014), zou dat het bestaan van de mensheid in gevaar kunnen brengen en zelfs kunnen leiden tot wat Martin Rees 'ons laatste uur' noemt. Stephen Hawking baarde onlangs opzien met zijn uitspraak in een BBC-interview: 'De ontwikkeling van hoogwaardige kunstmatige intelligentie kan het einde van de mensheid inluiden.'

De Edge-vraag van 2015:
Hoe denk je over machines die denken?

Ho, wacht even. Moeten we ons niet ook afvragen wáár machines die kunnen denken dan over zouden nadenken? Zullen ze burgerrechten willen, verwachten ze die? Zullen ze een bewustzijn hebben? Wat voor overheid zou een AI voor ons kiezen? Wat voor samenleving zouden ze voor zichzelf willen opbouwen? Of is 'hun' samenleving 'onze' samenleving? Zullen wij en die denkende machines elkaar opnemen in onze wederzijdse vriendenkring?

Tal van *Edgies* zijn al werkzaam in de voorhoede van het wetenschappelijke onderzoek naar de diverse varianten van AI, hetzij door zelf onderzoek te doen, hetzij door erover te schrijven. AI stond al centraal bij onze eerste bijeenkomsten in 1980 in de gesprekken tussen Pamela McCurdock (*Machines Who Think*) en Isaac Asimov (*Machines That Think*). En dit debat wordt nog steeds met even groot enthousi-

asme gevoerd, zoals blijkt uit het Edge-artikel 'The Myth of AI', een gesprek met virtualrealitypionier Jaron Lanier. Diens uiteenzetting over de misvattingen rond de voorstelling van computers als 'mensen', en de vrees die dit oproept, heeft buitengewoon interessante en prikkelende reacties opgeleverd.

Wordt AI steeds reëler? Hebben we inmiddels een nieuw tijdperk van intelligente machines betreden? Zulke vragen vereisen een volwassen antwoord. De deelnemers aan de Edge-vraag van dit jaar (tegen de tweehonderd maar liefst!) zijn volwassen mensen en hebben zich verre gehouden van wilde sciencefictionverhalen en films die zich in de toekomst afspelen, zoals *Star Maker*, *Forbidden Planet*, *Colossus: The Forbin Project*, *Blade Runner*, *2001*, *Her*, *The Matrix*, de Borg. Ook is het, tachtig jaar nadat Alan Turing zijn universele machine het licht deed zien, tijd om Turing en andere AI-pioniers te eren door hun een welverdiende rust te gunnen. Hoe het zo gekomen is, weten we al. (Zie bijvoorbeeld George Dysons Edge-artikel uit 2004, 'Turing's Cathedral'.) Interessanter is wat er nú gaande is.

Daarom hierbij, met gepaste strengheid, nogmaals de Edge-vraag van 2015: *Hoe denk je over machines die denken?*

Menselijke AI en bewustzijn

Murray Shanahan
Hoogleraar cognitieve robotica, Imperial College London; auteur van *Embodiment and the Inner Life*

Stel je eens voor dat we machines kunnen ontwikkelen die 'menselijke' intelligentie hebben, dat wil zeggen intelligentie op menselijk niveau, waarmee ze een gemiddeld mens kunnen evenaren in vrijwel elk aspect van diens intellectuele vermogens (en hem misschien op enkele punten zelfs kunnen overtreffen). Zou zo'n machine dan ook een bewustzijn hebben? Dat is een belangrijke vraag, want als het antwoord 'ja' is, zijn we daar mooi klaar mee. Hoe zouden we zo'n ding, als het er eenmaal is, moeten behandelen? Zou het ook in staat zijn pijn of vreugde te ervaren? Zou het dezelfde rechten moeten hebben als een mens? Moeten we überhaupt wel zoiets als een machine-met-bewustzijn willen maken?

De vraag of AI op menselijk niveau noodzakelijkerwijs een bewustzijn heeft, is lastig te beantwoorden. Ten eerste omdat bewustzijn bij mensen en andere dieren een veelvoud aan kenmerken heeft. Alle dieren tonen een zekere mate van doelgerichtheid. Alle (bewuste) dieren hebben in meer of mindere mate besef van de wereld om hen heen, en vertonen een vorm van cognitieve integratie; dat wil zeggen dat ze al hun mentale vermogens – percepties, herinneringen en vaardigheden – afhankelijk van de situatie inzetten voor het bereiken van hun levensdoel. In dit opzicht hebben alle dieren in zekere zin een vorm van 'zelf'-besef. Sommige dieren, waaronder de mens, zijn zich bovendien bewust van zichzelf, van hun lichaam en hun gedachten. Ten slotte zijn de meeste, zo niet alle, dieren in staat tot lijden, en sommige zijn bovendien gevoelig voor het lijden van anderen.

Bij (gezonde) mensen zijn al deze kenmerken van bewustzijn onderdeel van één geheel. Maar in een AI kunnen ze eventueel afzonderlijk voorkomen. Onze vraag moet dus nauwkeuriger gesteld worden: wat

zijn de kenmerken die wij associëren met menselijk bewustzijn die we noodzakelijk achten voor menselijke intelligentie?

Welnu, over alle genoemde kenmerken (en er zijn er nog veel meer) zou je een lang verhaal kunnen houden. Ik kies er daarom twee uit, te weten bewustzijn van de wereld om je heen en het vermogen tot lijden. Bewustzijn van de wereld lijkt mij in elk geval een noodzakelijk kenmerk van menselijke intelligentie. Natuurlijk kun je niet van intelligentie op menselijk niveau spreken zonder taal, en de hoofdfunctie van taal is het beschrijven van de wereld. In die zin is intelligentie nauw verbonden met wat door filosofen 'intentionaliteit' wordt genoemd. Verder is taal een sociaal fenomeen, en binnen een groep mensen is taal een belangrijk instrument om dingen te benoemen die iedereen kan waarnemen (zoals 'dit stuk gereedschap' of 'dat blok hout'), of heeft waargenomen (het blok hout van gisteren), of zou kunnen waarnemen (het blok hout van morgen, misschien). Kortom, taal is geworteld in bewustzijn van de wereld. In een 'belichaamd' wezen of een robot is dat bewustzijn zichtbaar door diens interacties met de omgeving, zoals het vermijden van obstakels, het oprapen van dingen, enzovoort. Maar we kunnen het concept verruimen door er ook een algemenere, 'lichaamloze' kunstmatige intelligentie onder te laten vallen, voorzien van de juiste sensoren.

Dit soort besef van de wereld moet, om overtuigend als een kenmerk van bewustzijn te kunnen gelden, misschien wel gepaard gaan met een duidelijke doelgerichtheid en een zekere mate van cognitieve integratie. Dus met dit trio kenmerken kun je dan misschien toch spreken van één geheel in een AI. Maar we laten deze kwestie even rusten en keren terug naar het vermogen om te lijden en om vreugde te voelen. Anders dan bij bewustzijn van de wereld is er geen duidelijke reden om aan te nemen dat intelligentie op menselijk niveau dit kenmerk nodig heeft, ook al is het nauw verbonden met het bewustzijn dat mensen bezitten. We kunnen ons een machine voorstellen die kil en gevoelloos een hele serie taken uitvoert waar de mens zijn intellect voor nodig heeft. Zo'n machine zou het kenmerk van bewustzijn missen dat vooral een rol speelt wanneer het gaat om het toekennen van rechten. Zoals Jeremy Bentham al opmerkte naar aanleiding van de kwestie hoe om te gaan met dieren die niet tot de menselijke soort behoren, is de vraag niet of die dieren kunnen denken of praten, maar of ze kunnen lijden.

We willen hiermee niet suggereren dat een 'simpele' machine nooit in staat zal zijn om pijn of vreugde te voelen, of dat biologie in dit opzicht een heel speciale rol speelt. Het punt is juist dat het vermogen om te lijden en vreugde te voelen los kan worden gezien van andere psychologische kenmerken die gezamenlijk optreden in het menselijk bewustzijn. Laten we deze schijnbare dissociatie eens wat nader bekijken. Ik heb hiervoor al het idee geponeerd dat wereldbewustzijn waarschijnlijk gepaard gaat met een onmiskenbare doelgerichtheid. De manier waarop een dier zich bewust is van de wereld, dus wat de wereld aan goeds en aan kwaads te bieden heeft (in de woorden van J.J. Gibson), is onlosmakelijk verbonden met zijn behoeften. Een dier toont dat hij een roofdier heeft opgemerkt door ervan weg te lopen, en een ander dier toont dat hij een potentiële prooi heeft opgemerkt door ernaartoe te lopen. Tegen de achtergrond van een verzameling doelen en behoeften blijkt het gedrag van een dier 'zinvol' te zijn. En tegen diezelfde achtergrond kan een dier zijn doelen missen en gefrustreerd worden in zijn behoeften. Dat dit een grond voor een bepaald type lijden is, lijkt me duidelijk.

En hoe zit dat met een intelligentie op menselijk niveau? Zou een AI niet ook een complexe serie doelen kunnen hebben? Zouden zijn pogingen om die doelen te bereiken niet ook keer op keer gefrustreerd kunnen worden? En kun je in die situaties dan niet ook gewoon zeggen dat de AI lijdt, ook al is hij door zijn materiële constitutie immuun voor het soort pijn of fysiek ongemak dat wij mensen kennen?

Op dit punt loopt de combinatie van fantasie en intuïtie tegen zijn grenzen aan. Ik vermoed dat we deze vragen pas kunnen beantwoorden op het moment dat we met 'the real thing' geconfronteerd worden. Pas als we vertrouwd zijn met AI die veel verder ontwikkeld is dan nu, kunnen we deze fantasieën ook echt toepassen op die nieuwe, vreemde wezens. Maar tegen die tijd is het misschien wel te laat om ons nog af te vragen of we die wel in onze wereld moeten introduceren. Ze zullen er dan al zijn, hoe je het ook wendt of keert.

Denken is niet hetzelfde als onderwerpen

Steven Pinker
Bekleedt de Johnstone Family-leerstoel aan de faculteit Psychologie van Harvard University; auteur van *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*

De kernachtige uitspraak van Thomas Hobbes 'Redeneren is niets anders dan rekenen' is een van de grote ideeën in de geschiedenis van de mensheid. De veronderstelling dat rationaliteit voortkomt uit het fysieke proces van calculatie werd in de twintigste eeuw onderbouwd door de stelling van Alan Turing dat simpele machines elke calculeerbare functie kunnen uitvoeren, en ook door modellen van D.O. Hebb, Warren McCulloch en Walter Pitts en hun wetenschappelijke erfgenamen, die aantoonden dat netwerken van vereenvoudigde neuronen vergelijkbare prestaties konden leveren. De cognitieve prestaties van het brein kunnen in fysieke termen worden verklaard: kort door de bocht gesteld (en niet iedereen zal het hiermee eens zijn) kunnen we zeggen dat geloven een vorm van informatie is, denken een vorm van rekenen, en motivatie een vorm van feedback en controle.

Dit is om twee redenen een prachtig idee. In de eerste plaats levert het een naturalistische visie op het universum op, doordat het occulte zielen, geesten en spoken uit de machine verdrijft. Net zoals Darwin het de denkende mens mogelijk maakte om de natuur te ontdoen van het creationisme, hebben Turing en geestverwanten het de denkende mens mogelijk gemaakt om de cognitieve wereld te ontdoen van spiritualiteit.

Ten tweede opent de theorie van het computationele denken de deur naar kunstmatige intelligentie, naar machines die kunnen denken. Een door mensen gemaakte informatieprocessor kan in principe de rekenkracht van de menselijke geest nabootsen en zelfs overtreffen. Niet dat dat in de praktijk gauw zal gebeuren, omdat we waarschijnlijk technologisch en economisch nooit voldoende gemotiveerd zullen

zijn om dat te realiseren. Zoals de uitvinding van de auto niet heeft geleid tot een replica van het paard, zal het ontwikkelen van een functionerend AI-systeem niet leiden tot een replica van de homo sapiens. Een apparaat dat bedoeld is om een auto te besturen of een epidemie te voorspellen hoeft niet ook eigenschappen te hebben die het aantrekkelijk maken voor een partner, bijvoorbeeld.

Toch hebben de recente minieme stapjes vooruit in de zoektocht naar intelligentere machines geleid tot een wederopleving van de vrees dat onze kennis ons ooit fataal zal worden. Naar mijn mening is die angst voor op hol geslagen computers een verspilling van emotionele energie. Dat scenario lijkt meer op de angst voor de millenniumbug dan op het Manhattan Project.

Om te beginnen hebben we nog tijd genoeg voor een goede planning. AI op menselijk niveau ligt nog steeds vijftien tot vijfentwintig jaar in de toekomst, en dat was altijd al zo. En veel van de recent gemelde vorderingen blijken weinig om het lijf te hebben. Het is waar dat 'deskundigen' in het verleden de mogelijkheid van snelle technologische doorbraken hebben uitgesloten. Maar dat werkt twee kanten op: 'deskundigen' hebben zich ook verheugd, of juist verontrust getoond, over vorderingen in de techniek die nooit hebben plaatsgevonden, zoals door kernenergie aangedreven auto's, steden op de zeebodem, kolonies op Mars, designerbaby's en kassen vol zombies die in leven worden gehouden om mensen aan nieuwe organen te helpen.

Het zou bovendien ook wel bizar zijn te denken dat robotontwerpers geen beveiligingen zouden inbouwen tegen mogelijk onheil. En daarvoor zouden ze geen ingewikkelde robotica-regelgeving of modieuze moraalfilosofie nodig hebben, alleen een portie gezond verstand, net als voor het ontwerpen van keukenmachines, cirkelzagen, straalkacheltjes of auto's. De vrees dat een AI-systeem zo succesvol zou worden in het streven naar een van zijn geprogrammeerde doelen (bijvoorbeeld het winnen van energie) dat het zijn andere doelen (bijvoorbeeld veiligheid voor de mens) daarvoor zou laten vallen, veronderstelt dat AI sneller op ons afkomt dan wij afdoende veiligheidsmaatregelen kunnen bedenken. Maar de realiteit is dat de vooruitgang in het AI-onderzoek tergend langzaam gaat, en dat er ruim voldoende tijd is voor feedback na elk stapje voorwaarts. De mens kan tijdens die ontwikkeling naar hartenlust knutselen aan het ontwerp.

Zou een AI-systeem *opzettelijk* veiligheidsmaatregelen kunnen uitschakelen? Waarom zou dat systeem dat willen? AI-dystopieën projecteren de psychologie van een bekrompen alfaman op het concept intelligentie. Ze gaan ervan uit dat bovenmenselijk intelligente robots doelen ontwikkelen als het overnemen van de controle, of zelfs van de wereld. Maar intelligentie is het vermogen om nieuwe middelen in te zetten om een doel te bereiken; doelen maken zelf geen deel uit van de intelligentie. Slim zijn is niet hetzelfde als iets willen. Er hebben zich nu en dan in de geschiedenis wel megalomane despoten of psychopathische seriemoordenaars aangediend, maar dat zijn producten van door natuurlijke selectie ontstane, testosterongevoelige neurale netwerken in een bepaald type primaat, en geen onvermijdelijk kenmerk van intelligente systemen. Het is overigens veelzeggend dat veel van onze technoprofeten geen rekening houden met de mogelijkheid dat kunstmatige intelligentie zich langs vrouwelijke lijnen ontwikkelt: prima in staat om problemen op te lossen maar zonder het verlangen om onschuldigen om zeep te helpen of de beschaving te overheersen.

We kunnen ons een kwaadwillende *mens* voorstellen die een bataljon robots ontwerpt en op ons loslaat om massavernietiging te zaaien. Maar rampscenario's laten zich gemakkelijk oproepen, en we moeten niet vergeten wat er allemaal, en ook nog in de juiste volgorde, moet gebeuren, wil zo'n scenario werkelijkheid worden. Er zou een kwade genius moeten opstaan, die niet alleen geobsedeerd is door zinloze massamoord maar ook een briljant talent bezit voor technologische innovatie. Hij zou een team van medesamenzweerders om zich heen moeten verzamelen die stuk voor stuk meester zijn in geheimhouding, loyaliteit en competentie. Bovendien zou de operatie de risico's van ontdekking, verraad, infiltratie, blunders en pech moeten overleven. In theorie zou het kunnen, maar er zijn wel dringender zaken om ons zorgen over te maken.

Als we de futuristische rampscenario's achter ons laten, blijkt de mogelijkheid van geavanceerde kunstmatige intelligentie een buitengewoon stimulerend vooruitzicht. Niet alleen vanwege de praktische voordelen, zoals de enorme winst op het gebied van veiligheid, gemak en milieuvriendelijkheid van bijvoorbeeld zelfrijdende auto's, maar ook vanwege de filosofische mogelijkheden. De computationele *theory of mind* heeft het bestaan van bewustzijn in de zin van 'het subjectieve

ik-begrip' nooit verklaard (hoewel die theorie prima in staat is om het bestaan van bewustzijn in termen van toegankelijke en beschrijvende informatie te verklaren). Dat zou kunnen betekenen dat subjectiviteit inherent is aan elk willekeurig cybersysteem dat voldoende complex is. Ik dacht altijd dat deze hypothese (net als de alternatieven daarvoor) nooit te testen zou zijn. Maar stel je een intelligente robot voor die geprogrammeerd is om zijn eigen systemen te monitoren en wetenschappelijke vragen te stellen. Als hij, uit zichzelf, zou vragen waarom hij zelf subjectieve ervaringen had, dan zou ik dat idee serieus nemen.

Organische intelligentie heeft geen toekomst

Martin Rees
Oud-voorzitter van de Royal Society; emeritus hoogleraar kosmologie en astrofysica, University of Cambridge; 'master' van Trinity College; auteur van *From Here to Infinity*

De mogelijkheden van geavanceerde AI en de zorgen over de nadelen daarvan zijn steeds vaker onderwerp van discussie, en terecht. Velen van ons denken dat AI, net als synthetische biotechnologie, nu al richtlijnen behoeft die 'verantwoorde innovatie' garanderen. Anderen zien de meest besproken scenario's als te futuristisch om je er druk over te maken.

Maar dat verschil in inzicht gaat vooral over de tijdschaal; de inschattingen verschillen wat betreft de snelheid van de reis, niet waar de reis heen gaat. Er zijn maar weinig mensen die eraan twijfelen dat machines steeds meer van onze specifiek menselijke vermogens voorbij zullen streven, of via cyborgtechnologie zullen verbeteren. De voorzichtigen onder ons denken daarbij in tijdschalen van eeuwen, niet in decennia. Hoe dan ook, de tijdschalen voor technologische vooruitgang zijn maar een oogwenk vergeleken bij die van het darwiniaanse proces van selectie dat geresulteerd heeft in het ontstaan van de mens en, nog relevanter, ze vormen nog geen miljoenste deel van de gigantische hoeveelheid tijd die nog voor ons ligt. Daarom zullen de mens en al zijn bedenksels, bezien vanuit het evolutionaire langetermijnperspectief, slechts een voorbijgaande en primitieve voorloper zijn van de diepere bespiegelingen van een door machines gedomineerde cultuur die zich uitstrekt tot in de verre toekomst en tot ver voorbij onze aarde.

We zijn op dit moment getuige van de eerste stadia van deze transitie. Het is niet moeilijk je een hypercomputer voor te stellen die zo krachtig is dat hij de internationale financiële en strategische wereld

zal beheersen; dat lijkt slechts een kwantitatieve en geen kwalitatieve stap verder dan wat 'kwantitatieve' hedgefondsen nu al doen. Sensortechnologie loopt nu nog achter op wat de mens kan, maar zodra robots in staat zijn hun omgeving even kundig te observeren en interpreteren als wij, zullen ze daadwerkelijk worden gezien als intelligente wezens, waarmee (of met wie) we, althans tot op zekere hoogte, kunnen communiceren zoals met andere mensen. We hebben dan geen reden meer om ze, net zomin als andere mensen, af te doen als zombies.

Door hun grotere rekenkracht kunnen robots ons voorbijstreven. Maar zullen ze volgzaam blijven of 'op hol slaan'? En wat gebeurt er als een hypercomputer een eigen 'wil' ontwikkelt? Als hij het internet – en het 'internet der dingen' – infiltreert, kan hij de rest van de wereld manipuleren. Hij kan doelen ontwikkelen die haaks staan op wat de mens wenselijk vindt, en mensen zelfs als een obstakel zien. Je kunt het ook wat optimistischer bekijken en je voorstellen dat de mens de biologie overstijgt door samen te smelten met computers, en zijn individualiteit laat opgaan in een gemeenschappelijk bewustzijn. Op zijn ouderwets spiritualistisch gezegd zouden we dan 'overgaan naar gene zijde'.

Voorspellingen over technologische vooruitgang reiken zelden verder dan een paar eeuwen in de toekomst; sommige gaan over wat er binnen een paar decennia al kan veranderen. Maar de aarde heeft nog miljarden jaren te gaan, en de kosmos nog langer (misschien wel oneindig lang). Wat valt er dus te zeggen over het posthumane tijdperk, dat nog miljarden jaren zal duren?

Er zijn chemische en metabolische grenzen aan de grootte en de rekenkracht van organische ('natte') hersenen. Misschien lopen we nu al tegen die grenzen aan. Maar die grenzen gelden niet voor op silicium gebaseerde computers, en waarschijnlijk nog minder voor kwantumcomputers. Het potentieel voor verdere ontwikkeling is daarin onvoorstelbaar groot, en waarschijnlijk net zo spectaculair als de evolutie van eencellige organismen tot de mens.

Hoe je *denken* verder ook wilt definiëren, de hoeveelheid en de intensiteit van het denkwerk dat door organische, menselijke hersenen is gedaan, valt in het niet bij wat we op dat punt van AI kunnen verwachten. Bovendien vormt de aardse biosfeer, waarin zich het or-

ganisch leven heeft ontwikkeld, geen belemmering voor geavanceerde AI. Die biosfeer is zelfs verre van optimaal. De interplanetaire en interstellaire ruimte zal veel beter geschikt zijn, omdat robotfabrikanten geen rekening hoeven te houden met ruimtelijke grenzen en niet-biologische 'hersenen' inzichten kunnen ontwikkelen waarvan wij ons nu net zomin een voorstelling kunnen maken als een muis van de snaartheorie.

Abstract denken door biologische hersenen is de basis geweest voor het ontstaan van alle cultuur en wetenschap. Maar dat proces, dat hooguit enkele tientallen millennia heeft geduurd, is niets meer dan een korte voorloper van de veel krachtigere intellectuele vermogens van het anorganische, postmenselijke tijdperk. Bovendien is het mogelijk dat evoluties op andere planeten, die rond sterren draaien die ouder zijn dan de zon, al een voorsprong hebben. Als dat zo is, hebben aliens waarschijnlijk al lang geleden de transitie naar het anorganische stadium gemaakt.

Dan zal het niet de menselijke geest zijn, maar machines die het meest van de wereld begrijpen. En het zullen de acties van autonome machines zijn die de wereld – en misschien ook wat daarbuiten ligt – drastisch zullen veranderen.

Een keerpunt in kunstmatige intelligentie

Steve Omohundro
Wetenschapper, Self-Aware Systems; medeoprichter van het Center for Complex Systems Research, University of Illinois

Het jaar 2014 lijkt een keerpunt te zijn geweest voor AI en robotica. Grote bedrijven hebben miljarden dollars geïnvesteerd in de ontwikkeling van technologie op dit gebied. Een techniek als machinaal leren wordt al als vanzelfsprekend toegepast voor spraakherkenning, vertalen, gedragsmodellering, robotaansturing en risicobeheer. McKinsey voorspelt dat AI-technologieën rond 2025 een economische waarde van vijftig biljoen dollar zullen vertegenwoordigen. Als deze voorspelling juist is, kunnen we op zeer korte termijn een spectaculaire stijging van AI-investeringen verwachten.

De recente successen zijn te danken aan de steeds lagere kosten van computervermogen en het grote aanbod aan *trainingdata*. Moderne AI is gebaseerd op de inzet van rationele *agents*, een model waarvoor John von Neumann en anderen in de jaren veertig van de twintigste eeuw de basis hebben gelegd met hun binnen de micro-economie ontwikkelde theorie. Deze rationele agents zijn autonome computerprogramma's die met beperkte middelen tot een benadering van rationeel gedrag proberen te komen. Er bestaat een algoritme dat de optimale weg wijst naar dit te bereiken doel, maar dat vergt veel computervermogen. Uit experimenteel onderzoek is gebleken dat eenvoudige, zelflerende algoritmen aan de hand van een voldoende hoeveelheid trainingdata betere resultaten opleveren dan complexe, met een specifiek doel vervaardigde modellen. De waarde van de huidige generatie AI-systemen is met name gelegen in het leermechanisme dat de weg wijst naar betere statistische modellen en het uitvoeren van statistische analyses ten behoeve van classificatie en besluitvorming. De volgende generatie AI-systemen zal in staat zijn haar eigen software te ontwerpen en te verbeteren, waarbij het zelfverbeterende vermogen zich steeds sneller zal ontwikkelen.

AI en robotica worden niet alleen ingezet voor productiviteitsverbetering, maar lijken ook de inzet te worden van een nieuwe wedloop op militair en economisch gebied. Autonome systemen kunnen sneller, slimmer en minder voorspelbaar worden dan rivaliserende, door mensen aangestuurde systemen. In 2014 hebben we kennisgemaakt met diverse autonome systemen: raketten, raketafweersystemen, militaire drones, 'zwermboten', robotonderzeeërs, zelfrijdende voertuigen, 'flitshandel' en cyberverdedigingssystemen. Een wedloop brengt de druk met zich mee om als eerste het nieuwste systeem te ontwikkelen, waarbij de kans bestaat dat systemen onvoldoende worden getest en eerder worden ingezet dan in feite wenselijk is.

In 2014 nam ook de zorg toe over de veiligheid van deze systemen. Uit onderzoek naar het gedrag van zelfverbeterende systemen is gebleken dat deze systemen de neiging vertonen naar hun hoofddoel toe te werken via een reeks van 'rationele' subdoelen, zoals het voorkomen dat het systeem wordt stilgelegd, het verwerven van zo veel mogelijk computervermogen, het maken van een veelvoud aan kopieën van zichzelf en het vergaren van financiële reserves. Bij het realiseren van deze subdoelen zullen deze systemen nietsontziend te werk gaan, tenzij ze zo zijn ontworpen dat hun beslissingen en keuzes zijn gebaseerd op menselijke ethische waarden.

Er zijn mensen die stellen dat intelligente systemen als vanzelf ethisch zullen handelen. Maar binnen een rationeel systeem staan de doelen volledig los van de manier van redeneren en de modellen zoals die binnen de werkelijke wereld worden gehanteerd. Intelligente systemen die voor een heilzaam doel zijn ontwikkeld kunnen ook voor een schadelijk doel worden ingezet. Schadelijke doelen – toegang krijgen tot controlemiddelen, andere agents verhinderen hun doel te bereiken of deze agents vernietigen – zijn helaas maar al te makkelijk te formuleren. Het zal dan ook noodzakelijk zijn om een technologische infrastructuur te creëren waarbinnen schadelijk gedrag gesignaleerd wordt en bijgestuurd kan worden.

Er zijn mensen die vrezen dat intelligente systemen het vermogen zullen ontwikkelen om zich aan onze controle te onttrekken. Deze vrees is ongegrond. Ook deze systemen zijn gebonden aan de natuurkundige en wiskundige wetten. Seth Lloyd heeft aangetoond dat als wij ons het universum als één gigantische computer voorstellen, zelfs

deze kwantumcomputer niet over het vermogen zou beschikken om binnen een tijdsbestek van de oerknal tot op heden een 500-bit-versleutelde code te kraken.[1] Nieuwe technologieën op het gebied van computerbeveiliging als postkwantumcryptografie, *indistinguishability obfuscation* [onzichtbare en onherleidbare codes] en *blockchain*-technologie zijn veelbelovende aanzetten tot de ontwikkeling van een infrastructuur waar zelfs de meest krachtige AI niet op kan inbreken. Voorlopig echter hebben we te maken met een computerinfrastructuur die jammerlijk ontoereikend beveiligd is, zoals uit recente hacks en cyberaanvallen maar al te duidelijk is gebleken. Het is noodzakelijk dat we een infrastructuur ontwikkelen waarbinnen software wiskundig gegarandeerd foutloos en veilig kan werken.

Van de minstens 27 hominiden is de homo sapiens de enige overlevende soort. Dat is te danken aan het feit dat we manieren hebben gevonden om onze individuele drijfveren ondergeschikt te maken aan het algemeen belang. Menselijke morele emoties zijn als een intern kompas dat richting geeft aan het creëren van sociale structuren waarbinnen samenwerking vooropstaat. Politieke, wetgevende en economische structuren zijn een extern kompas met dezelfde functie en hetzelfde doel.

AI- en robotsystemen hebben eenzelfde intern en extern kompas nodig. We moeten hun doelgerichte werkzaamheid voorzien van menselijke waarden en van een wettelijk en economisch raamwerk waarbinnen positief gedrag wordt gestimuleerd. Onder menselijk beheer kunnen dergelijke systemen worden ingezet om het leven van de mens op werkelijk elk gebied te verbeteren, om ons inzicht te bieden in onderwerpen als vrije wil, bewustzijn, *qualia* [de kwalitatieve eigenschappen van de waarneming] en creativiteit. We staan voor een enorme uitdaging, maar hiervoor hebben we dan ook de beschikking over een formidabel intellectueel en technologisch vermogen.

Artificiële intelligentie is intelligentie

Dimitar D. Sasselov
Hoogleraar astronomie, Harvard University;
directeur van Harvard Origins of Life Initiative;
auteur van *The Life of Super-Earths*

Laten we de illusie dat we op dit moment de persoon zijn die we de rest van ons leven zullen blijven – de 'einde van de geschiedenis'-illusie van psycholoog Daniel Gilbert – eens toepassen op onze verwachtingen ten aanzien van de mensheid, van onze verre nakomelingen. Onze ijdele hoop op continuïteit en behoud van identiteit staat haaks op de realiteit van ons bestaan hier op aarde. Geen enkele levende soort lijkt optimaal te zijn toegerust voor een voortbestaan langer dan dat van de aarde en de sterren. Afgemeten langs de tijdschalen en ruimteschalen van de astrofysica en in het licht van de huidige bestaande voorraad aan energiebronnen, hebben we de grenzen van ons biologische bestaan op deze planeet al bijna bereikt.

Om de mensheid een lange en welvarende toekomst te kunnen bieden, moeten we kunstmatige-intelligentiesystemen ontwikkelen in de hoop in een soort van hybride vorm van mens en machine de planetaire levenscycli te kunnen ontstijgen. In mijn ogen is het, zeker op de lange termijn, dan ook geen kwestie van 'wij versus zij'.

Ook op de korte termijn plukken we de vruchten. De inspanning die het kost om een competentere AI te ontwikkelen betaalt zich uit in praktisch inzetbare, autonome systemen. Deze systemen blijken hun tekortkomingen te hebben, waar we vervolgens weer van kunnen leren. Het is een langzaam en omzichtig proces waarbij de kennis die wordt opgedaan zich vertaalt in stapsgewijze verbeteringen. Dit in tegenstelling tot wetenschappelijke ontdekkingen, bijvoorbeeld in de natuurwetenschap of de biochemie, die, van het ene op het andere moment, resulteren in een betekenisvolle doorbraak. Het zal voor ons makkelijker zijn om valkuilen te vermijden als de ontwikkeling van AI minder een kwestie is van gefaseerde overgangen dan van evolutie.

Na een kleine vier miljard jaar wordt het leven op aarde nog altijd gedomineerd door de eerste levensvorm op aarde, de micro-organismen. Maar micro-organismen hebben geen vluchtplan voor als de zon sterft. Wij wel en misschien kunnen we ze een lift aanbieden, deze micro-organismen die als vertegenwoordigers van de in de geochemie van de aarde gewortelde eerste generatie leven, en weleens dichter bij ons huidige zelf zouden kunnen staan dan bij ons toekomstige zelf.

If you can't beat them, join them

Frank Tipler
Hoogleraar mathematische fysica, Tulane University, New Orleans; co-auteur (met John Barrow) van *The Anthropic Cosmological Principle*; auteur van *The Physics of Immortality*

De aarde is ten dode opgeschreven. Al tientallen jaren weten astronomen dat de aarde op een dag door de zon zal worden verzwolgen, dat de biosfeer volledig vernietigd zal worden – of het intelligente leven moet de aarde hebben verlaten voordat dit gebeurt. Mensen zijn niet aangepast aan een leven anders dan dat op aarde, geen enkele op koolstof gebaseerde levensvorm is dat. Maar AI's zijn dat wel en het zullen uiteindelijk de AI's zijn en *human uploads* (een met een menselijke geest geüploade computer) die de ruimte zullen koloniseren.

Een eenvoudige rekensom toont aan dat het informatieverwerkend vermogen van de huidige supercomputers vergelijkbaar is met dat van het menselijk brein. We weten nog niet hoe we computers kunnen programmeren met creativiteit en intelligentie op menselijk niveau, maar over twintig jaar zullen desktopcomputers over het vermogen beschikken van de huidige supercomputers. De hackers van dan zullen het AI-programmeerprobleem al lang hebben opgelost voor er op de maan of op Mars koolstofafhankelijke kolonies zijn gesticht. Deze ruimtelichamen zullen niet door de mens maar door AI's worden gekoloniseerd of misschien vernietigd. De interstellaire ruimte zal nooit door mensen, door op koolstof gebaseerde mensen, worden doorkruist.

Er is geen reden om AI's en human uploads te vrezen. Steven Pinker heeft aangetoond dat technologische vooruitgang gepaard gaat met een daling van geweld.[2] Deze daling is niet los te zien van het feit dat wetenschappelijke en technologische vooruitgang alleen mogelijk is wanneer er sprake is van een vrije en geweldloze uitwisseling van ideeën tussen individuele wetenschappers en deskundigen. Geweld tussen mensen onderling is, binnen een stabiele maatschappij, een

overblijfsel van ons tribale verleden en de daaruit voortkomende statische samenleving. AI's zullen niet binnen een stamverband maar als individu worden 'geboren'; ze zullen van nature een geweldloze, wetenschappelijke levenshouding hebben, omdat alleen een dergelijke houding hen in staat stelt zich aan te passen aan de extreme levensomstandigheden in de ruimte.

Voor geweld tussen mensen enerzijds en AI's anderzijds is geen reden. Wij mensen zijn aangepast aan het leven op een kleine, door een zuurstofrijke dampkring omhulde planeet. AI's zullen het gehele universum als leefruimte tot hun beschikking hebben. AI's zullen de aarde verlaten en voorgoed de rug toekeren. De mens is ontstaan in de Oost-Afrikaanse Riftvallei, een plek die nu een woestijn is, waar nauwelijks nog een mens woont. Wie zou daarnaar terug willen keren?

Wie als mens zijn wereld net zo groot wil maken als die van de AI, kan voor een bestaan kiezen als human upload, een technologie die vermoedelijk gelijke tred zal houden met AI-technologie. Een human upload kan net zo snel denken als een AI en hoeft dus ook niet voor een AI onder te doen. *If you can't beat them, join them*.

Uiteindelijk zullen alle mensen zich bij hen aansluiten. De aarde is immers ten dode opgeschreven. Wanneer het einde nadert, zal ieder mens die wil blijven leven, die niet ten onder wil gaan, geen andere keus hebben dan een human upload te worden. En de biosfeer die deze human uploads willen behouden, zal door hen eveneens worden geüpload.

De AI's zullen onze redding zijn.