Voorbij sciencefiction: hoe reëel is het gevaar van kunstmatige intelligentie?

By **Thalia Verkade**, decorrespondent.nl
november 18de, 2014

Raketbouwer Elon Musk waarschuwde vorige maand: kunstmatige intelligentie vormt waarschijnlijk de grootste bedreiging voor ons bestaan. Met de technologie zouden we demonen oproepen die we niet zullen kunnen bezweren. Ook sterrenkundige Stephen Hawking en andere wetenschappers slaan alarm. Maar een door onszelf gecreëerd monster dat de mensheid vernietigt – daar kun je veel vragen bij stellen.

Zoals:

1. Wie waarschuwt waarvoor?
2. Hoe ver zijn we met kunstmatige intelligentie?
3. Waarom slaan de wetenschappers juist nu alarm?
4. Bestaat er al een vorm van kunstmatige intelligentie die in staat is zelfstandig de mensheid te vernietigen?
5. *Wil* kunstmatige intelligentie ons eigenlijk wel vernietigen?
6. Waarom voelt het zo onwaarschijnlijk?

Wat definities vooraf: onder kunstmatige intelligentie, kortweg A.I. (van *artificial intelligence*), vallen in dit artikel software en robots die menselijk handelen of denken imiteren. Een vorm van A.I. die een specifieke vaardigheid nabootst (bijvoorbeeld: een schaakcomputer of vertaalsoftware) heet in dit artikel Beperkte A.I. Een vorm van kunstmatige intelligentie die even slim en autonoom is als de mens, bestaat nog niet. Dit concept heet in dit artikel Algemene A.I.

Daar gaan we dan.

1. Wie waarschuwt waarvoor?

Elon Musk, de man achter het ambitieuze ruimtebedrijf SpaceX Bekijk hier de website van SpaceX. Bekijk hier de website van SpaceX. SpaceX (missie: naar Mars) en de elektrische auto van Tesla Motors, zei Bekijk Musks lezing. Bekijk Musks lezing. zei vorige maand bij een grote bijeenkomst op het Massachusetts Institute of Technology: ‘Ik denk dat we heel voorzichtig moeten zijn met kunstmatige intelligentie. Als ik moest raden wat de grootste bedreiging is voor ons bestaan, dan is het waarschijnlijk dat. […] Met A.I. roepen we demonen op. Ken je die verhalen met die vent met een pentagram en heilig water die denkt dat ’ie daarmee de demon kan bezweren? Dat werkte niet.’

Stephen Hawking schreef Lees de brief. Lees de brief. schreef in mei samen met drie andere hoogleraren in *The Independent*: ‘Als we slagen in het scheppen van A.I., dan zou dat de grootste gebeurtenis zijn in de geschiedenis van de mens. Helaas zou het ook de laatste kunnen zijn, tenzij we leren hoe we risico’s moeten mijden.’

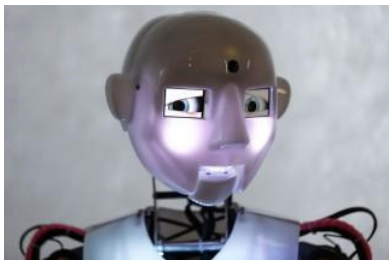
> ‘Als we slagen in het scheppen van A.I., dan zou dat de grootste gebeurtenis zijn in de geschiedenis van de mens. Helaas zou het ook de laatste kunnen zijn’

Musk adviseerde zijn Twittervolgers het boek *Superintelligence* (2014) van filosoof Nick Bostrom te lezen. Bostrom, verbonden aan de Universiteit van Oxford, beschrijft een scenario waarin de bedreiging begint als een kunstmatig zenuwstelsel, Een vorm van A.I.

die biologische zenuwstelsels imiteert en afgelopen twee decennia een grote ontwikkeling heeft doorgemaakt. Meer over kunstmatige zenuwstelsels bij vraag 3. dat de functies van het menselijk brein goed genoeg imiteert om zichzelf verder te kunnen ontwikkelen, met een digitale rekenkracht vele malen groter dan die van de mens.

Deze vorm van intelligentie groeit in Bostroms voorstelling eerst uit tot Algemene A.I. en vervolgens tot iets waar de mens zich geen voorstelling meer van kan maken, omdat ons verstand daarvoor tekortschiet: Superintelligentie. Bostrom noemt meer scenario's, bijvoorbeeld die waarin de neurowetenschappen zich sneller ontwikkelen dan A.I., waardoor het mogelijk wordt een heel menselijk brein te uploaden naar het internet. Als dat gebeurt, zo benadrukt Bostrom, is de mens niet meer de slimste en dus ook niet meer de machtigste soort op aarde.

Stephen Hawking gaat niet in op A.I. als mogelijke nieuwe levensvorm, maar stelt zich een wetenschappelijke explosie voor. 'Je zou je kunnen voorstellen dat dit soort technologie de financiële markten te snel af is, beter wordt in uitvinden dan menselijke onderzoekers, beter wordt in het manipuleren van mensen dan menselijke leiders en dat het wapens gaat ontwikkelen die we niet eens kunnen begrijpen. Waar de impact van A.I. op de korte termijn gaat over wie er de controle heeft, gaat de impact op lange termijn erover of ze überhaupt nog gecontroleerd kan worden,' schrijft Hawking.



De robot Thespian op de tentoonstelling 'Robot Ball' in Moskou.
Foto: Hollandse Hoogte

2. Hoe ver zijn we met kunstmatige intelligentie?

In de beginjaren van de studie naar A.I. (de jaren vijftig en zestig) is er veel onderzoek gedaan naar het ontwikkelen van een vorm van kunstmatige intelligentie die evengoed of beter zou presteren dan de mens.

Na tegenvallende resultaten werden de doelstellingen bescheidener: eerst maar eens een computer leren schaken. Met dat soort Beperkte A.I. is het de afgelopen twee decennia snel gegaan. Een greep:

- Smartphones kunnen spraak verstaan, terugpraten en een route berekenen.
- A.I. voor financiële markten leest het nieuws op zoek naar informatie die van belang is voor de beurs en handelt zelf.
- Gehoorapparaten kunnen spraak van achtergrondgeluiden ontdoen.
- Er was een enorm rekencentrum voor nodig om een computer het cryptische spel *Jeopardy!* te laten winnen van een mens, maar toen het zover was (in 2011) liet 'Watson' de menselijke competitie direct mijlenver achter zich.
- Het A.I.-programma DART berekende de logistiek van Operation Desert Storm, Onderdeel van de Eerste Golfoorlog in 1991 waarmee dertig jaar aan investeringen in A.I. zich volgens DARPA Het Defense Advanced Research Projects Agency, onderzoeksinstituut van het Amerikaanse ministerie van Defensie. in een klap hadden terugverdiend.

Verder zijn er vormen van A.I. die meer zijn dan software, zoals robots. Enkele toepassingen en recente ontwikkelingen:

- De zelfrijdende auto van Google leert zichzelf met elk ritje beter hoe hij een bocht optimaal neemt.
- De robot MDARS bewaakt Amerikaanse nucleaire installaties.
- De MARCbot helpt in Irak bij het onschadelijk maken van explosieven.
- De militaire pakezel BigDog (sinds dit jaar in handen van Google) blijft in balans als je hem een schop verkoopt.
- Een robot Uitleg en beelden van de Cornellrobot. Uitleg en beelden van de Cornellrobot. robot van Cornell University weet als hij wordt aangezet nog niet dat hij vier poten heeft, maar leert zichzelf binnen minuten een voorstelling van zichzelf maken en vervolgens te kruipen en te lopen.

Bekijk BigDog in actie

YouTube
Veel van bovengenoemde successen zijn te danken aan de ontwikkeling van de kunstmatige zenuwnetwerken Video met een korte uitleg. Video met een korte uitleg. kunstmatige zenuwnetwerken waar filosoof Nick Bostrom superintelligentie uit voort ziet komen. Deze vorm van A.I. kan zichzelf leren patronen te herkennen in diffuse input, bijvoorbeeld een handgeschreven tekst, een geluidsopname of videobeelden. Een simpel brein eigenlijk.

3. Waarom slaan de wetenschappers juist nu alarm?

Terwijl A.I.-wetenschappers afgelopen decennia steeds geavanceerdere kunstmatige zenuwnetwerken ontwierpen, brachten neurowetenschappers almaar gedetailleerder in kaart hoe het menselijk brein werkt. Beide vakgebieden leren nu van elkaar.

Daarnaast brengen robotica, nano- en biotechnologie de digitale en fysieke wereld gestaag dichter bij elkaar. Sinds 1991 raakt alles op de wereld steeds meer verbonden door het internet.

Dit alles levert een scala aan nieuwe mogelijkheden op voor digitale techniek. Een nieuwe generatie A.I.-wetenschappers bijt zich nu vast in die aloude, enorme uitdaging: het creëren van technologie die kan leiden tot het ontstaan van Algemene A.I. Soms heet dit Strong A.I., daarbij geldt meestal het idee dat de software of machine op een bepaald moment zelfbewustzijn wordt.

Het is de combinatie van ontelbaar veel nieuwe mogelijkheden in en tussen verschillende vakgebieden, plus de kans op een doorbraak in het vakgebied A.I. die leidt tot onrust: gaat het allemaal niet te snel?

Elon Musk werd dit jaar investeerder in het bedrijf Vicarious, De website van Vicarious. De website van Vicarious. Vicarious, naar eigen zeggen om een oogje in het zeil te houden. Het doel van Vicarious: dat computers straks 'waarnemen, fantaseren en redeneren zoals mensen.' Dat zou een stap zijn richting Algemene A.I.: een machine zou dan wellicht een kunstwerk kunnen maken of boeken schrijven die mensen mooi vinden. Of nieuwe wapens ontwikkelen.

Grootste wapenfeit van Vicarious tot nog toe: het bedrijf zegt Vicarious over CAPTCHA Vicarious over CAPTCHA zegt CAPTCHA te hebben gekraakt, de foto's of plaatjes van kromgetrokken letters of cijfers die je soms moet invoeren als je inlogt op een website. CAPTCHA is een zogeheten Turingtest, waarmee gecontroleerd wordt of degene die toegang wil tot een website een mens is of een computer. Dat klinkt niet heel spectaculair, maar deze technologie was nu net bedoeld om computers buiten de deur te houden.

Vicarious hoopt de complexere kunstmatige zenuwnetwerken die nodig zijn om een machine abstract te laten denken te genereren door de neocortex na te bouwen, het (enorme) deel van het menselijk brein dat ziet, beweging coördineert, rekent en taal begrijpt.

> Vicarious hoopt de complexere kunstmatige zenuwnetwerken die nodig zijn om een machine abstract te laten denken te genereren door de neocortex na te bouwen

Het Britse bedrijf DeepMind DeepMind ging begin dit jaar na een biedingsstrijd tussen Google en Facebook naar Google, dat ook al meerdere roboticabedrijven overnam en bezig is met (medische) nanotechnologie. Van Vicarious is onder anderen Mark Zuckerberg (Facebook) mede-aandeelhouder. heeft een vergelijkbare missie ('*Solve intelligence*') en strategie, maar richt zich op de hippocampus, het deel van het het brein waar geheugen en leervermogen resideren.

De opvallendste prestatie Kijk hoe DeepMinds software Breakout en Pong leert spelen. Kijk hoe DeepMinds software Breakout en Pong leert spelen. prestatie van DeepMind is een programma dat zichzelf allerlei verschillende oude computerspelletjes heeft leren spelen. Niet door de regels aangereikt te krijgen, maar door, als een mens, te reageren op de wereld zoals die wordt gerepresenteerd in een spel, te oefenen en zichzelf te verbeteren. Bedoeling is dat het programma de kennis van oude games steeds beter kan gebruiken bij het spelen van nieuwe spellen.

Hoe slimmer machines worden, hoe groter de stappen zijn die ermee kunnen worden gezet. Wellicht speurt een daarvoor geoptimaliseerde vorm van A.I. straks met nanobots in ons lichaam naar kanker (zie dit project TechCrunch over Googles nanopillen. TechCrunch over Googles nanopillen. dit project van Google) - waarmee het meteen het menselijk lijf in kaart brengt, wat weer allerlei nieuwe mogelijkheden biedt.

Of leert A.I. zoals die van DeepMind zichzelf straks niet alleen schietspellen spelen maar ook echte wapens bedienen. Zolang de mens de uiteindelijke controle heeft, betekent dat niet het einde van de wereld. Maar door delen van het brein te simuleren, dragen bedrijven als DeepMind en Vicarious de verantwoordelijkheid steeds verder over aan die machine zelf.

De robot Thespian op de tentoonstelling ‘Robot Ball’ in Moskou. Foto: Hollandse Hoogte

4. Bestaat er al een vorm van kunstmatige intelligentie die in staat is zelfstandig de mensheid te vernietigen?

Er bestaat in elk geval een specifieke vorm van kunstmatige intelligentie die in staat is om mensen te herkennen en te vernietigen. De Britse drone Taranis Producent BAE Systems over Taranis. Producent BAE Systems over Taranis. Taranis heeft in theorie geen menselijke tussenkomst meer nodig om een doelwit te selecteren en uit te schakelen. Het is daarmee een van de meest geavanceerde LAWS, ‘Lethal Autonomous Weapons Systems’ (dodelijke autonome wapenssystemen), waarvan het bestaan bekend is. De drone schiet zelf nog niet op mensen: de producent kiest ervoor om een mens ‘in the loop’ te houden. Vorig jaar maakte Taranis zijn eerste testvlucht.

Afgelopen mei vergaderden de Verenigde Naties voor het eerst over deze zogenoemde ‘Lethal Autonomous Weapons Systems’ (LAWS) en afgelopen week voor de tweede keer. Sommige landen (waaronder Pakistan) stellen dat dit soort wapens volledig verbannen moet worden.

Andere landen (waaronder Nederland) vinden het voldoende zolang er ‘strenge menselijke controle’ is.

Maar die redenering kan problemen opleveren.

Stel je een situatie voor waarbij iemand een drone zoals Taranis uit de lucht probeert te schieten. Dankzij zijn geavanceerde sensors heeft Taranis die poging veel eerder door dan de mensen die de drone op afstand commanderen.

Moet een kostbaar project De ontwikkeling van het prototype kostte 143 miljoen Engelse pond. als Taranis een menselijke beslissing afwachten met de kans dat het aan flarden geschoten wordt? Of willen de eigenaren toch liever dat het in zo’n geval zelf schiet? Waarom hebben ze het anders die mogelijkheid gegeven?

Wie dit soort technologie te slim af wil zijn, moet met een afweersysteem komen met software die nog weer sneller is dan die van Taranis. En op het moment dat die er is, kun je geen mensen meer ‘*in the loop*’ houden zonder het risico te nemen te verliezen van de robot van je tegenstander.

Hoe ontwikkel je zo snel mogelijk een nog beter wapen? Een A.I. die is geoptimaliseerd om wapens te begrijpen, kan dat als bedrijven als Deepmind en Vicarious in hun missie slagen straks misschien sneller dan een mens. Dan gaan we richting het scenario waar Hawking voor waarschuwt: dat er wapens komen die mensen niet meer kunnen begrijpen.

5. Maar *wil* kunstmatige intelligentie ons vernietigen?

Dat is, stellen veel A.I.-wetenschappers, de verkeerde vraag: het gaat erom of kunstmatige intelligentie mensen *niet* wil vernietigen.

Stel dat een zelflerende en zelfoptimaliserende vorm van A.I. de opdracht krijgt zoveel mogelijk cijfers achter de komma van het getal pi te berekenen.

Zonder enige kwade wil kan deze machine berekenen dat zij de taak het beste volbrengt als zij eerst zoveel mogelijk rekenkracht genereert, bijvoorbeeld door zoveel mogelijk computers die zij via het internet kan bereiken te hacken en te laten meedraaien in de berekeningen.

Als de machine zichzelf leert wat je kunt met geld of energie of politieke macht, dan kan ze ook proberen om die in handen te krijgen of te manipuleren. In de meest fantasierijke scenario's (zie *Superintelligence* van Nick Bostrom) ontwikkelt een op deze manier losgebroken A.I. vervolgens nanotechnologie om atomen uit menselijke lichamen te kunnen hergebruiken voor energie of constructie van nog meer rekenkracht. En voilà: de mensheid is vernietigd.

> Als A.I. zichzelf leert wat je kunt met geld of energie of politieke macht, dan kan zij ook proberen om die in handen te krijgen of te manipuleren

Maar dan geef je toch de opdracht dat bij het berekenen van het getal pi géén menselijk leed mag worden veroorzaakt?

Het probleem daarvan is dat iedere oplossing waarmee je machine komt, mogelijk niet degene is die je in gedachten had: 'Een supergeoptimaliseerde machine die als grootste doel heeft om het menselijk leed te beperken, zou, lijkt me, een manier vinden om alle mensen te doden zonder pijn. Geen mensen, geen menselijke pijn,' schreef Peter Norvig in 2012. Norvig is hoofdonderzoeker bij Google.

De boodschap: als A.I. niet is uitgerust met de expliciete opdracht om de mensheid in haar waarde te laten, dan zal die zich ontwikkelen zonder rekening te houden met de mensheid *whatsoever*.

Om dit probleem op te lossen, moet er eerst een antwoord komen op de nu nog open vraag hoe je een zeer autonome vorm van A.I. (die bijvoorbeeld het internet op kan om bij te leren) een doel geeft dat deze onmogelijk verkeerd kan interpreteren.

Een filosofische oplossing komt van het Machine Intelligence Research Institute (MIRI) De website van het MIRI. De website van het MIRI. Machine Intelligence Research Institute (MIRI) , een van de handvol non-profitorganisaties Andere zijn het Future of Humanity Institute in Oxford (sinds 2005), Centre for the Study of Existential Risk in Cambridge (sinds 2012) en Future of Life Institute in Boston (sinds dit jaar). die zich bezighouden met de risico's van A.I. Onderzoeker Eliezer Yudkowsky stelt voor kunstmatige intelligentie uit te rusten met een programma dat telkens weer extrapoleert wat de mensheid wil zijn, maar nog niet is, en deze wens als uitgangspunt neemt voor haar eigen verdere ontwikkeling. Dan krijgt de mensheid in elk geval wat ze zelf als soort wil.

Een praktischer idee Zie dit artikel van Steve Omohundro. Zie dit artikel van Steve Omohundro. idee komt van natuurkundige Steve Omohundro, die een denktank De denktank van Steve Omohundro. De denktank van Steve Omohundro. denktank bestiert die zich op risico's met A.I. richt en gastcolleges verzorgt op universitaire opleidingen Kunstmatige Intelligentie.

Omohundro ontwikkelt veiligheidsmechanismes die de komende jaren zouden kunnen worden toegepast om 'bosbranden' als gevolg van onvoorzichtige ontwikkeling van A.I. te voorkomen, bijvoorbeeld door de rekenkracht van experimentele en potentieel machtige vormen van A.I. te voorzien van natuurkundige limieten. Voor deskundigen: Steve Omohundro noemt de Bekenstein bound en Bremermann's limit.

De robot Thespian op de tentoonstelling 'Robot Ball' in Moskou. Foto: Hollandse Hoogte

6. Waarom voelt het zo onwaarschijnlijk?

Er wordt al eeuwen gesuggereerd dat de mensheid binnenkort vergaat. Door zevendedagsadventisten, door media die een spannende invalshoek zochten voor een verhaal 'Could a super collider destroy the world? Proposed upgrade could create black holes and 'strange matter', warn experts' over de CERN. Bovendien wordt de belofte dat er binnenkort een computer komt die slimmer is dan de mens al sinds de jaren veertig gedaan. En is die belofte nog nooit uitgekomen.

En dan is er de paradox dat elke vorm van kunstmatige intelligentie, zodra die er is, door mensen niet langer als werkelijke intelligentie wordt ervaren. Een potje schaken winnen, een smartphone die je verstaat en terugpraat: zodra een computer het kan en de techniek een toepassing vindt lijkt het weinig meer voor te stellen.

Er is nog een verklaring waarom een doemscenario met A.I. moeilijk te behappen is. Mensen zijn niet goed in (wetenschappelijke) revoluties voorspellen. Volgens veel technofuturisten, waaronder die van de Singularity University die deze week Nederland aandoet, Webpagina over de bijeenkomst in Amsterdam (dinsdag en woensdag). Webpagina over de bijeenkomst in Amsterdam (dinsdag en woensdag). aandoet, zitten we op dit moment midden in een periode waarin het aantal mogelijke technologische toepassingen exponentieel groeit - en blijven we ten onrechte in lineaire technische vooruitgang denken.

A.I.-onderzoeker Eliezer Yudkowsky ziet een historische parallel. In 1933, toen er moeizaam werd gezocht naar manieren om atomen te splitsen, verklaarde de scheikundige en Nobelprijswinnaar Ernest Rutherford: 'Mensen die zoeken naar een bron van energie door atomen te transformeren, kletsen uit hun nek.'

Het zou nog negen jaar intensieve arbeid kosten tot het eerste experiment om precies dat te doen slaagde, met grote gevolgen voor de wereld. Het werd ook de eerste uitvinding die de mensheid daadwerkelijk aan de rand van de afgrond bracht, met de Cubacrisis van 1962.

Aangehaalde literatuur:

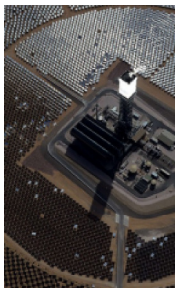
'Autonomous technology and the greater human good' Het artikel van Omohundro. Het artikel van Omohundro. 'Autonomous technology and the greater human good' (2014) van Steve Omohundro

‘Intelligence Explosion and Machine Ethics’ Het artikel van Muehlhauser en Helm. Het artikel van Muehlhauser en Helm. ‘Intelligence Explosion and Machine Ethics’ (2014) van Luke Muehlhauser en Louie Helm (MIRI)

‘Artificial Intelligence as a Positive and Negative Factor in Global Risk’ Het artikel van Yudkowsky. Het artikel van Yudkowsky. ‘Artificial Intelligence as a Positive and Negative Factor in Global Risk’ (2008) van Eliezer Yudkowsky (MIRI)

Superintelligence Bostrom over zijn boek. Bostrom over zijn boek. Superintelligence (2014) van Nick Bostrom. De homepage van Bostroms website. De homepage van Bostroms website. Nick Bostrom.

Voor Silicon Valley is democratie de grootste vijand van vooruitgang In Californië wordt in 2016 gestemd over de vraag of Silicon Valley zich mag afscheiden van de Verenigde Staten. Achter het referendum gaat een diepe onvrede schuil van de tech-wereld met de ineffeciënte worstenmachine die democratie heet, stelt gastcorrespondent Thijs Kleinpaste. Waar komt die onvrede vandaan? En wat zijn de mogelijke consequenties? Lees het verhaal hier terug



Is de slimme thermostaat van Google het begin van een groene revolutie? Wie wil googelen heeft internet en dus stroom nodig. Daarom gaat Google de energiemarkt op. Het bedrijf heeft de capaciteiten om die radicaal te veranderen en te vergroenen. Regelt deze megacorporatie naast onze zoekopdrachten, e-mailverkeer en navigatie straks ook onze energiedistributie? Lees het verhaal hier terug

Suggestie voor een Kamervraag We gebruiken machines al eeuwen om de buitenwereld te beheersen. Maar sinds kort dringen ze door tot in onze diepste gedachten en geheimen. Een nogal fundamentele vraag dringt zich op. Lees hier de column van Rutger terug

Abonneer je op mijn verhalen via mijn nieuwsbrief Waarom veroveren sommige uitvindingen de wereld en andere niet? Als je je inschrijft houd ik je op de hoogte van mijn verhalen. Schrijf je hier in!