Over
Als iemand dit bouwt, gaat iedereen dood:

‘Heel leesbaar en meeslepend.’
– The Observer

‘Een angstaanjagende waarschuwing waar je niet omheen kunt.’
– Publishers Weekly

‘Het onvoorstelbare wordt ineens voorstelbaar gemaakt.’
– Kirkus Reviews

‘Meeslepend en geloofwaardig, ondanks de apocalyptische these.’
– The Times

‘Yudkowsky en Soares laten in heldere, begrijpelijke taal zien waarom de huidige race naar steeds krachtigere AI heel gevaarlijk is.’
– Emmett Shear, voormalig CEO van OpenAI

‘Dit is misschien wel het belangrijkste boek van deze tijd. Een urgent pleidooi om onszelf te redden, zolang dat nog kan.’
– Tim Urban, auteur van *Wait But Why*

‘We leven in de laatste periode van de geschiedenis waarin wij mensen de dominante soort zijn. De mensheid mag blij zijn met Yudkowsky en Soares, die ons eraan herinneren om de korte mogelijkheid aan te grijpen om beslissingen over onze toekomst te maken.’
– Grimes

Eliezer Yudkowsky & Nate Soares

ALS IEMAND DIT BOUWT, GAAT IEDEREEN DOOD

Waarom superintelligente AI ons einde kan betekenen

Uit het Engels vertaald door Jan Willem Reitsma

MAVEN
PUBLISHING

Oorspronkelijke titel *If anyone builds it, everyone dies. Why superhuman AI would kill us all*

Nederlandse vertaling

www.mavenpublishing.nl

Omslag DPS
Zetwerk Elgraphic
Drukwerk Wilco

ISBN boek 978 94 9343 440 0
ISBN e-boek 978 94 9343 433 2
NUR 801

Voor alle mensen die ooit zijn gestorven,
terwijl onze soort op weg was naar waar we nu zijn,
voor iedereen die nog onder de levenden is
en voor alle kinderen die misschien nog zullen komen.

Inhoud

DEEL III

Het probleem onder ogen zien

INLEIDING
Moeilijke en makkelijke voorspellingen

'Het verkleinen van de kans op uitroeiing door AI zou een mondiale prioriteit moeten zijn, naast andere gevaren die de hele maatschappij raken, zoals pandemieën en kernoorlogen.'

Begin 2023 ondertekenden honderden wetenschappers die zich met artificiële intelligentie bezighielden een open brief die bestond uit deze ene zin. Onder de ondertekenaars bevonden zich enkele van de meest gelauwerde onderzoekers uit het vakgebied, onder wie Nobelprijswinnaar Geoffrey Hinton en Yoshua Bengio, die samen de Turing Award ontvingen voor de uitvinding van *deep learning*.

Wij, Eliezer Yudkowsky en Nate Soares, hebben de brief ook ondertekend, ook al vonden wij het een gigantisch understatement.

Over de AI's van 2023 maakten wij en de andere ondertekenaars ons geen zorgen. We zijn ook niet bezorgd over de AI's die er zijn terwijl we dit begin 2025 schrijven. De AI's van nu voelen nog oppervlakkig, op een diepgaande manier die lastig is te beschrijven. Ze hebben hun beperkingen, bijvoorbeeld dat ze geen langetermijngeheugen kunnen ontwikkelen. Vooralsnog zijn die tekortkomingen groot genoeg om te voorkomen dat die AI's wetenschappelijk onderzoek van enige betekenis doen of erg veel mensen werkloos maken.

Onze bezorgdheid gaat over wat daarna komt: artificiële intelligentie die écht intelligent is, intelligenter dan welke levende mens dan ook, intelligenter dan de hele mensheid bij elkaar. Wij maken ons zorgen

over AI die uitstijgt boven het menselijk vermogen om na te denken, die kan generaliseren op basis van ervaringen, die wetenschappelijke raadsels kan oplossen en nieuwe technologieën kan uitvinden, die kan plannen, strategieën ontwikkelen en lijnen uitzetten, en op zichzelf kan reflecteren en zichzelf kan verbeteren. Als zulke AI elke mens overtreft bij bijna elke intellectuele taak, zou je die 'artificiële superintelligentie' (ASI) kunnen noemen.

Zover is AI nog niet. Maar AI's zijn nu intelligenter dan ze in 2023 waren, en veel intelligenter dan in 2019. Onderzoek heeft gezorgd voor de ene na de andere sprong voorwaarts wat betreft de vermogens van AI: in 2012,* in 2016,† in 2020,‡ in 2022§ en in 2024.¶ We weten niet of die vooruitgang zal stagneren en deze vooruitgang een tijdlang zal stoppen tot er weer nieuwe methoden en technologieën worden uitgevonden. We weten niet hoeveel sprongen er nog gemaakt moeten worden vóór AI zich heeft ontwikkeld tot het gevaarlijke niveau van uitroeiing waarvoor de ondertekenaars van de brief waarschuwden. Maar de geschiedenis heeft keer op keer laten zien hoe AI-onderzoekers nieuwe methoden uitvinden en oude hindernissen uit de weg ruimen. Die vooruitgang gaat vaak verrassend snel. In 2015 zouden de meeste informatici je hebben gezegd dat artificiële gesprekken op het niveau van ChatGPT pas over zo'n dertig of vijftig jaar haalbaar zouden zijn.

We wisten niet wanneer artificiële superintelligentie werkelijkheid zou worden, maar we waren het erover eens dat ze een wereldwijde prioriteit zou moeten zijn. Sterker nog: volgens ons wordt het probleem in de open brief nog veel te zwak uitgedrukt.

* In 2012 loste AlexNet het probleem op van het herkennen van objecten in afbeeldingen.

† In 2016 versloeg AlphaGo de allerbeste go-speler.

‡ In 2020 werd het (zuiver voorspellende) taalmodel GPT-3 gelanceerd.

§ In 2022 arriveerde het (algemeen bruikbare) ChatGPT.

¶ In 2024 begonnen redeneermodellen wiskundige, visuele en programmeervraagstukken op te lossen.

We werden uitgenodigd om die open brief van één zin te ondertekenen omdat wij samen de leiding voerden over het Machine Intelligence Research Institute (MIRI), een non-profitorganisatie. MIRI had zich sinds 2001 beziggehouden met vragen rond superintelligentie, lang voordat deze vraagstukken veel publiciteit of financiering kregen. Heel simpel gezegd: het selecte gezelschap dat zich decennia met deze kwestie heeft beziggehouden erkent MIRI als de organisatie die hier het langst mee bezig is geweest. Een van ons, Yudkowsky, heeft MIRI opgericht; de ander, Soares, is momenteel de directeur ervan.

MIRI was de eerste georganiseerde groep die zei: 'Het is vrijwel zeker dat op een bepaald moment superintelligente AI ontwikkeld zal worden, en dat lijkt ons een extreem serieuze kwestie. Waarschijnlijk zal het technisch moeilijk zijn om superintelligentie zo vorm te geven dat ze de mensheid helpt, in plaats van haar te schaden. Moet er niet iemand onmiddellijk met dat probleem aan de slag gaan, in plaats van te wachten tot er later een gigantische crisis ontstaat?'

In onze begintijd zeiden we dat nog niet. In het jaar 2000 probeerde Yudkowsky nog artificiële superintelligentie te bouwen. Maar in 2001 besefte hij dat die niet per se vriendelijk zou uitpakken. En in 2003 besefte hij dat het een complex probleem is.

In de eerste twee decennia van zijn bestaan was MIRI een technisch onderzoeksinstituut dat zich niet erg bezighield met politiek beleid. De organisatie gaf voornamelijk workshops voor geïnteresseerde wetenschappers en bood onderdak aan een paar veelbelovende onderzoekers. We probeerden te berekenen hoe we bovenmenselijke artificiële intelligentie zouden kunnen begrijpen en vormgeven, en hoe we konden voorspellen hoe dit alles fout zou kunnen gaan.

Op sommige uitkomsten van MIRI kijken we nu met een dubbel gevoel of zelfs spijt terug. Tijdens een congres dat wij organiseerden, hebben we Demis Hassabis en Shane Legg, de oprichters van het latere Google DeepMind, aan hun eerste grote geldschieter voorgesteld. En Sam Altman, de CEO van OpenAI, heeft ooit beweerd dat Yudkowsky

'bij velen van ons de interesse voor AGI heeft gewekt'* en 'van groot belang was bij het besluit om OpenAI te beginnen'.†

MIRI heeft een complexe geschiedenis, maar je zou onze relatie met het bredere vakgebied als volgt kunnen samenvatten: het was algemeen bekend dat als je wilde werken aan het bouwen van echt slimme AI, je de waarschuwingen van MIRI over de kans op uitsterving in de wind moest slaan. En dit was jaren voordat ook maar een van de huidige AI-bedrijven bestond.

Nadat AI echt van de grond was gekomen, keken we bezorgd toe hoe een paar van de nieuwere mensen die AI-bedrijven hadden opgericht over artificiële superintelligentie begonnen te praten als een bron van enorme, geweldige krachten. Krachten waarvan ze aannamen dat ze die konden beheersen. Volgens veel van die oprichters was het grootste gevaar dat ASI 'in handen' zou komen van de verkeerde mensen. Ze spraken over de noodzaak om de 'AI-wapenwedloop' te winnen. De mogelijkheid dat niet jij een ASI 'in handen hebt', maar dat de ASI zélf een ASI 'in handen heeft' – dat de enige winnaar van een AI-wapenwedloop de ASI zélf zou zijn – tja, daarover spraken die oprichters niet.

We zagen dat de capaciteiten van AI heel snel toenamen.

We zagen dat het onderzoeksgebied waarin wij actief waren – gericht op het begrijpen van AI's en zorgen dat ze niet de fout in gingen – veel, véél langzamer vorderde.

De roekeloze race van de AI-bedrijven om bovenmenselijke AI te ontwikkelen – hun pogingen om die zo snel mogelijk te bouwen, en eerder dan hun concurrenten – werd in onze ogen een neerwaartse spiraal. De branche stormde op een ramp af: het soort ramp dat als schoolvoorbeeld zou dienen hoe je niét moet ontwikkelen. Alleen zou er niemand meer in leven zijn om die analyse te schrijven.

* 'AGI' staat voor *Artificial General Intelligence*, een term om een AI die intuïtief 'echt intelligent' is te onderscheiden van de AI's van weleer die slechts één doel hadden. In dit boek vermijden we deze term, omdat er sinds de opkomst van AI's zoals ChatGPT veel onenigheid bestaat over wat hij betekent.

† Als dat zo is, is het ondanks het bezwaar van Yudkowsky dat OpenAI een echt gruwelijk idee was.

Het leek ons niet langer realistisch dat de mensheid een manier om die catastrofe te voorkomen kon programmeren of onderzoeken. Niet in deze omstandigheden. Niet op tijd.

We concludeerden dat al ons werk tot dan toe was mislukt, wikkelden de meeste onderzoeksactiviteiten van MIRI af en verplaatsten de aandacht van het instituut naar het duidelijk maken van één enkel punt, de waarschuwing die de kern van dit boek vormt:

Als welk bedrijf of welke groep dan ook, waar ook ter wereld, een artificiële superintelligentie bouwt met gebruikmaking van iets wat ook maar enigszins op de huidige technieken lijkt, op basis van iets wat ook maar enigszins lijkt op het huidige begrip van AI, zal iedereen overal op aarde sterven.

Dat bedoelen we niet als overdrijving. We overdrijven niet uit effectbejag. Volgens ons is dit de meest directe gevolgtrekking uit de kennis, het bewijs en het gedrag van instituten die zich op dit moment bezighouden met artificiële intelligentie.

In dit boek zetten we ons betoog uiteen in de hoop dat we genoeg beslissingmakers en gewone mensen ervan kunnen overtuigen dat ze AI serieus moeten nemen. Als we niets doen is de uitkomst dodelijk, maar de zaak is niet hopeloos: er bestaat nog geen artificiële superintelligentie, en de ontwikkeling ervan kan nog steeds worden tegengehouden.

Hoe kan iemand zeker weten wat er met AI zal gebeuren? 'Voorspellingen doen is erg moeilijk, vooral over de toekomst,' zoals het aforisme luidt. De meeste dingen die we graag over de toekomst zouden willen weten zijn eigenlijk niet te voorspellen. Zo kunnen we je niet zeggen op welke lotnummers volgende week prijzen vallen. Het ene lotnummer lijkt net zoveel kans te hebben als het andere.

Maar sommige feiten over de toekomst zijn wél voorspelbaar. Als jij morgen een lot koopt, weten we niet welke ingewikkelde theorieën je zult toepassen bij het kiezen van je lotnummer, of misschien kies je juist wel op gevoel; we weten ook niet welke getallen getrokken gaan worden, maar al die onzekerheid bij elkaar opgeteld leidt tot de zeer betrouwbare voorspelling dat je niet de loterij zult winnen. Als je een ijsblokje in een

glas warm water laat vallen, valt ook onmogelijk te voorspellen waar elke molecuul zich tien minuten later zal bevinden, maar al die onzekerheid bij elkaar leidt wel tot een aan zekerheid grenzende voorspelling dat het ijsblokje zal smelten. Een groot deel van de natuurkunde werkt net zo: we kunnen niet berekenen welke route precies wordt afgelegd, maar we weten wel waar vrijwel alle wegen naartoe leiden.

Sommige aspecten van de toekomst zijn met de juiste kennis en inspanning voorspelbaar, andere zijn haast onmogelijk te voorspellen. Futurisme is gebouwd op het besef van het verschil tussen die twee.

De geschiedenis leert dat er één soort vrij eenvoudige toekomstvoorspelling is: als iets volgens de natuurkundige wetten in theorie mogelijk is, kun je voorspellen dat íémand het uiteindelijk zal doen. Vliegen met objecten die zwaarder dan lucht zijn, wapens die kernenergie uitstoten, raketten die naar de maan vliegen met een mens aan boord: die gebeurtenissen werden van tevoren voorspeld, om de juiste redenen, ondanks sceptici die met al hun wijsheid opmerkten dat die dingen nog nooit waren gebeurd en daarom waarschijnlijk ook nooit zouden gebeuren. Mensen die vleugels aan hun armen bonden en van heuvels afsprongen, zagen er op allerlei manieren idioot uit, werden door hun tijdgenoten bespot, verwondden zichzelf en hun pogingen mislukten. Maar dat weerhield de gebroeders Wright er niet van om te bedenken hoe ze konden vliegen.

Andersom is het in de loop van de geschiedenis veel moeilijker gebleken om exact te voorspellen wannéér een bepaalde technologie wordt ontwikkeld. Mensen zeggen over een technologie dat het nog twee jaar duurt, terwijl het in werkelijkheid nog vijftig jaar duurt; of zeggen dat het nog vijftig jaar zal duren terwijl het in werkelijkheid twee jaar is en ze die technologie zelf zullen bouwen. 'Het zal nog duizend jaar duren eer de mens vliegt,' zei Wilbur Wright in 1901 tegen Orville Wright, uit ergernis over het zweefvliegtuig zonder aandrijving dat ze op dat moment aan het testen waren. Twee jaar later, in 1903, vlogen de broers.

Het draait er bij juiste toekomstvoorspellingen niet om dat je de details voorspelt die doorgaans onvoorspelbaar zijn. Het gaat niet om het

bedenken van het volledige verhaal over wat er zal gebeuren en dan op wonderbaarlijke wijze je gelijk halen. Het gaat eerder om het zoeken naar bepaalde toekomstige gebeurtenissen die eenvoudig te voorspellen zijn, als je ze maar vanuit de goede invalshoek bekijkt.

We weten niet wanneer de wereld zal eindigen als mensen en landen niets veranderen aan de wijze waarop ze met artificiële intelligentie omgaan. We weten niet wat de kranten over twee of tien jaar over AI zullen schrijven, en niet eens of we überhaupt tien jaar te gaan hebben. We zullen niet beweren dat we zo slim zijn dat we dingen kunnen voorspellen die moeilijk voorspelbaar zijn. In plaats daarvan denken we dat één specifiek aspect van de toekomst – 'wat gebeurt er met alles en iedereen waar we van houden als in de nabije toekomst superintelligentie wordt gebouwd?' – met voldoende achtergrondkennis en zorgvuldig redeneren makkelijk kan worden voorspeld.

De uitroeiing van de mensheid door bovenmenselijke AI lijkt op het eerste gezicht misschien geen makkelijke voorspelling. Maar daar is de rest van het boek voor. Net zoals je iets van wiskunde moet weten om je kans te berekenen dat je de loterij wint, net zoals je een paar begrippen uit de thermodynamica moet kennen om te kunnen zeggen waarom een ijsklontje gegarandeerd zal smelten, heb je wat achtergrondkennis nodig om te begrijpen waarom artificiële intelligentie een naderend gevaar voor de uitroeiing van de mensheid vormt. Maar als die fundamenten op hun plaats liggen, wordt het voorspellen van de uitkomst van de weg die we nu bewandelen akelig, gruwelijk eenvoudig.

Zelfs bij de confrontatie met bovenmenselijke artificiële intelligentie kan het verleidelijk zijn om te denken dat de wereld hetzelfde zal blijven als in de laatste paar decennia van je betrekkelijk korte leven. Het is waar, maar voor ons moeilijk om te realiseren, dat er een tijd is geweest – net zo echt als de onze, en nog maar een paar eeuwen geleden – waarin de menselijke beschaving radicaal anders was. Of dat er millennia geleden geen noemenswaardige beschaving bestond. Of dat er een miljoen jaar geleden nog geen mensen waren. Of dat een miljard jaar geleden meercellige kolonies nog geen gespecialiseerde cellen hadden.

Historisch perspectief kan ons helpen om iets te beseffen wat vanuit de optiek van onze korte levensspanne erg moeilijk te zien is: de natuur laat ontwrichting toe. De natuur laat crisissen toe. De natuur laat toe dat de wereld voorgoed verandert.

Lang geleden, tweeënhalf miljard jaar geleden, vond er een gebeurtenis plaats die biologen de zuurstofcrisis noemen: een nieuwe levensvorm leerde om de energie van zonlicht te gebruiken en waardevolle koolstof aan de lucht te onttrekken. Die levensvorm stootte een gevaarlijke en sterk reagerende afvalstof uit, die voor het merendeel van de bestaande levensvormen giftig was: een stof die wij tegenwoordig 'zuurstof' noemen. Die begon zich op te bouwen in de atmosfeer. De meeste levensvormen, inclusief de meeste bacteriën die die zuurstof uitstootten, konden die reactiviteit niet aan en gingen dood. Een paar fortuinlijke cellijnen pasten zich aan en ontwikkelden zich uiteindelijk tot organismen die zuurstof als brandstof gebruiken. Maar het werd nooit meer zoals het ooit was. De wereld was voorgoed veranderd.

Lang geleden bestonden de continenten uit kale rotsen. Tot ze in een evolutionaire oogwenk bedekt waren met vegetatie. Kort daarop wemelde het in bossen van het leven. De wereld was voorgoed veranderd.

Lang geleden domesticeerden een paar mensen tarwe en gerst. In een fractie van een evolutionaire oogwenk begonnen ze beschavingen te bouwen. De wereld was voorgoed veranderd.

Lang geleden, in de jaren 1930, waren er signalen dat bepaalde gezinnen niet meer veilig waren in Duitsland. Sommige vertrokken snel, de meeste bleven. Tot het naziregime hen van het burgerschap beroofde, hun paspoorten innam en ontsnappen veel moeilijker maakte. Een paar jaar later werden Duitse Joden, Roma en anderen opgepakt en naar vernietigingskampen getransporteerd. Volgens de getuigenissen van overlevenden waren veel van die families niet gevlucht omdat ze geloofden dat het leven weer normaal zou worden voor het allemaal al te erg uit de hand liep, niet omdat de ze de voortekenen hadden genegeerd.[1]

Lang geleden stond de mensheid op het punt om artificiële superintelligentie te scheppen...

Aan een normale toestand komt altijd een eind. Dat wil niet zeggen dat er altijd iets slechters voor in de plaats komt. Soms gebeurt dat wel, soms niet, en soms hangt het af van wat we doen. Maar wat in elk geval niet helpt, is om je vast te klampen aan de hoop dat er niet iets heel vreselijks zal gebeuren.

Mensen hebben het vermogen om met hun intellect richting te geven aan de toekomst. Maar dat vermogen werkt alleen als we het daadwerkelijk gebruiken: als we doen wat we moeten doen, op het moment dat het nodig is. Alleen dan heeft ons intellect een functie. Ze functioneert als het ons handelen verandert, anders niet.

De maanden en jaren die voor ons liggen zullen een beproeving op leven en dood voor de hele mensheid worden. Met dit boek hopen we mensen en landen te inspireren om te laten zien wat ze waard zijn.

In de volgende hoofdstukken zullen we laten zien wat de wetenschap achter onze bezorgdheid is, de perverse prikkels bespreken die een rol spelen in de huidige AI-industrie en uitleggen waarom de situatie zelfs nog onheilspellender is dan die lijkt. We zullen in begrijpelijke taal moderne *machine learning* bespreken en we zullen beschrijven hoe en waarom de huidige methoden volstrekt ontoereikend zijn om AI's te maken die de wereld beter maken in plaats van die te vernietigen.

In **Deel I** van dit boek zetten we het probleem uiteen en geven we antwoord op vragen als: Wat is intelligentie? Hoe worden moderne AI's gemaakt en waarom zijn ze zo moeilijk te begrijpen? Kunnen AI's dingen willen? Zullen ze die ook willen? En zo ja, wat zullen ze willen en waarom zouden ze ons willen vermoorden? Hoe zouden ze ons vermoorden? Ten slotte voorspellen we AI's die ons niet zullen haten, maar die bizarre, vreemde, buitenmenselijke voorkeuren zullen hebben die ze nastreven tot het punt dat de mens uitsterft.

In **Deel II** verbinden we al deze onderwerpen om een verhaal te vertellen over een AI die een eind maakt aan een wereld die sterk op de onze lijkt. Dit verhaal is geen voorspelling, want het is moeilijk te voorspellen hoe de toekomst zich precies zal ontwikkelen. Het enige deel van het verhaal dat wel te voorspellen is, is het laatste deel – en die voorspelling

zal alleen uitkomen als we toestaan dat een dergelijk verhaal überhaupt begint.

In **Deel III** evalueren we de moeilijke uitdaging waar de mensheid voor staat, en bespreken we hoe we er tot nu toe op hebben gereageerd. Hoe goed gaan AI-bedrijven met het probleem om? Waarom schenkt de wereld er niet meer aandacht aan? Hoe kan de samenleving het anders doen, als genoeg mensen besluiten dat ze niet dood willen? Wat is ervoor nodig om te verhinderen dat de aarde een artificiële superintelligentie zal kennen?

Op de website IfAnyoneBuildsIt.com is een uitbreiding van dit boek beschikbaar, in het Engels. Aan het eind van elk hoofdstuk zie je een url en een QR-code met een link naar dat hoofdstuk. Dat ziet er telkens zo uit:

IfAnyoneBuildsIt.com/intro

Mensen hebben allerlei tegenstrijdige gevoelens over artificiële intelligentie en door de jaren heen hebben we allerlei vragen en bezwaren gehoord, die berustten op een breed spectrum van vooronderstellingen en invalshoeken. In de online uitbreiding bespreken we meer bezwaren, nuances en vaak gestelde vragen, en ook een paar principiële theoretische onderbouwingen en uitvoeriger argumenten die dit boek vele malen dikker en veel minder toegankelijk zouden hebben gemaakt. Als je aan het eind van een hoofdstuk bezwaren voelt opkomen, raden we je aan om online verder te lezen.

Veel hoofdstukken beginnen met parabels: verhalen waarvan we hopen dat die sommige dingen duidelijker overbrengen. Misschien voegen ze ook wat lichtvoetigheid toe aan dit verder zo zware onderwerp. Dat sluit aan bij een oeroude traditie, die misschien wel ouder is dan de menselijke soort in zijn huidige vorm: om de dood in zijn gezicht uit te lachen.

Toegegeven, dit boek bevat geen geweldig nieuws. Maar we gaan je ook niet vertellen dat we ten dode zijn opgeschreven. Artificiële superintelligentie bestaat nog niet. De mensheid kan nog steeds besluiten om haar niet te bouwen.

In de jaren vijftig verwachtten veel mensen dat er tussen de grote wereldmachten een kernoorlog zou uitbreken. Gezien de geschiedenis van het menselijk conflict tot dan toe was er alle reden om pessimistisch te zijn. Toch heeft er tot nu toe geen kernoorlog plaatsgevonden. Niet omdat kernbommen sciencefiction bleken te zijn die nooit gemaakt zouden worden: de reden is dat mensen hard hebben gewerkt om robuuste systemen te bouwen om het beginnen van kernoorlogen te voorkomen. Allemaal omdat wereldleiders inzagen dat een kernoorlog voor zowel henzelf als voor hun burgers nogal een slechte dag zou zijn.

Als iemand ergens op aarde een artificiële superintelligentie creëert, zou dat ook een slechte dag zijn. Het is echt in *niemands* belang om met al zijn familieleden en vrienden, zijn land en de kinderen van zijn land te sterven.

De voortdurende escalatie van AI-technologie een halt toeroepen, de hardware beteugelen die wordt gebruikt om steeds krachtiger AI-modellen te maken, dat is niet iets wat in de wereld van nu *makkelijk* te doen zou zijn. Maar het zou veel minder moeite kosten om de verdere escalatie van AI-capaciteiten tegen te houden dan het, bijvoorbeeld, was om de hele Tweede Wereldoorlog uit te vechten. Om een appèl te doen op de wil om te leven hoeven bepaalde landen en leiders en stemgerechtigden enkel te beseffen dat ze op de rand van een moeilijk in te schatten, maar mogelijk vrij korte afstand van de rand van de dood staan.

Het zal geen makkelijke opgave zijn, maar we zijn nog niet dood. De menselijke waardigheid en de waardigheid van de mensheid eisen dat we ons verdedigen.

Zolang er leven is, is er hoop.

DEEL I

Niet-menselijke geesten

1

Het bijzondere vermogen van de mensheid

Stel je voor, als je wilt – hoewel zoiets natuurlijk nooit is gebeurd en dit slechts een parabel is –, dat het biologische leven op aarde de uitkomst was van een spel dat goden met elkaar speelden. Dat er een tijgergod was die de tijgers had geschapen en een sequoiagod die de sequoia's had geschapen. Stel je voor dat er goden voor vissoorten en bacteriënsoorten waren. Stel je voor dat deze spelers streden om de heerschappij te veroveren voor de soort die zij steunden, de levensvormen die beneden op de planeet rondzwierven.

Stel je voor dat pakweg twee miljoen jaar geleden een obscure mensapengod uitkeek over hun enorme planeetgrote spelbord.

'Het zal me een paar extra zetten kosten,' zei de mensachtigengod, 'maar ik geloof dat ik dit spel ga winnen.'

Er heerste een verward zwijgen, terwijl vele goden op het spelbord zochten naar wat ze over het hoofd hadden gezien.

De schorpioenengod zei: 'Hoe dan? Jouw "mensachtigen"-familie heeft geen pantser, geen klauwen en geen gif.'

'Hun hersenen,' zei de mensachtigen-god.

'Ik besmet ze en dan gaan ze dood,' zei de pokkengod.[1]

'Heel even,' zei de mensachtigen-god. 'Jouw eind zal snel komen, Pokken, wanneer hun hersenen eenmaal leren hoe ze jou moeten bestrijden.'

'Ze hebben niet eens de grootste hersenen die er zijn!' zei de walvissengod.

'Het gaat niet altijd om de grootte,' zei de mensachtigengod. 'Het *ontwerp* van hun hersenen speelt ook een rol. Geef die hersenen twee miljoen jaar en dan lopen ze op de maan van hun planeet.'

'Ik zie écht niet waar de raketbrandstof in de spijsvertering van dit schepsel wordt geproduceerd,' zei de sequoiagod. 'Je kunt jezelf niet in een baan om de aarde *denken*. Op een bepaald moment moet die soort van jou spijsverteringsprocessen ontwikkelen die raketbrandstof raffineren – ze moeten trouwens ook behoorlijk groot worden, het liefst lang en smal – en met een harde buitenschil, zodat hij zichzelf niet opblaast en sterft in het luchtledige van de ruimte. Hoe hard jouw mensaap ook denkt, hij zal gewoon op de grond blijven, diep in gedachten verzonken.'

'Sommigen van ons spelen dit spel al miljarden jaren,' zei een bacteriëngod met een schuine blik op de mensachtigengod. 'Tot nu toe zijn hersenen geen bijzonder groot voordeel geweest.'

'En toch,' zei de mensachtigengod.

Dit is hoe de geschiedenis van onze soort echt is gegaan. Mensen kregen hersenen die abnormaal groot waren voor een dier met hun formaat. Ze bedwongen het vuur, bouwden boerderijen en smolten ijzer. Verbazingwekkend snel volgens de maatstaven van de biologische evolutie landden er mensen op de maan, ook al kan onze spijsvertering geen raketbrandstof raffineren en is onze huid niet bestand tegen het luchtledige.

Andere soorten op aarde worden met specialistische vaardigheden geboren: bijen bouwen bijenkorven, bevers bouwen dammen. Een mens kijkt naar de dam van de bever en achterhaalt hoe je dat doet; wij hebben geleerd om dammen te bouwen zonder dat die kennis is onze genen ingebakken hoefde te zijn. Nu dammen we rivieren af, net als de

bevers, we bouwen huizen voor onszelf, net als de bijen en we weven draden tot een net, net als de spinnen. En we bouwen energiecentrales en ruimteraketten die geen enkele andere soort bouwt.

Mensen kunnen dingen doen die hun voorouders nooit deden, en die andere dieren niet kunnen, vanwege een eigenschap die soms 'intelligentie' wordt genoemd. Onze genen hebben ons niet uitgerust met de meeste van onze vermogens: in plaats daarvan hebben we geobserveerd, uitgeprobeerd, onthouden, conclusies getrokken en daarmee dingen bereikt.

Mensen zijn niet de enigen met het vermogen om te leren. De hersenen van een muis kunnen leren om de weg in een doolhof te vinden. Maar wij beschikken over een veel krachtiger versie van het muizenbrein. Wij kunnen onze weg vinden in de scheikunde, en kunnen daardoor goedkopere mest produceren. Wij kunnen gecompliceerde experimenten opzetten, natuurkunde begrijpen en satellieten uitvinden. Wij kunnen zelfs muizen in doolhoven zetten en bestuderen hoe zíj dingen leren.

Onze hersenen stellen ons in staat om meer van de wereld te begrijpen en ons in meer verschillende situaties te redden dan welk ander dier dan ook. Dát is ons bijzondere vermogen.

Hoe werkt dat bijzondere vermogen? Wat doet het, en hoe?

Volgens ons* gaat het bij intelligentie om twee fundamenteel verschillende functies: de functie om de wereld te voorspellen, en de functie om die aan te sturen.

'Voorspellen' is raden wat je zult zien (of horen, of aanraken) voordat je het zintuiglijk ervaart. Als je met de auto naar het vliegveld rijdt, voert je brein de taak van het voorspellen telkens met succes uit wanneer je erop anticipeert dat een stoplicht op oranje springt, of dat de chauffeur die voor je rijdt op de rem trapt.

* Deze mening berust op enkele theorieën die we in de online extra's bespreken. Uiteindelijk moeten we ons niet al te druk maken over definities. Als een bos waarin je je bevindt door blikseminslag in brand vliegt, kun je jezelf niet redden door 'vuur' zodanig te definiëren dat het alleen door mensen aangestoken branden omvat. Je moet gewoon wegrennen.

'Aansturen' betekent: dingen doen die een bepaalde uitkomst opleveren. Wanneer je naar het vliegveld rijdt, stuurt je brein je succesvol aan wanneer het een zodanig patroon van afslagen vindt dat je bij het vliegveld uitkomt, of de juiste zenuwsignalen vindt om je spieren zodanig te spannen dat je aan het stuur draait.

De meeste patronen die je brein via zenuwimpulsen naar je vingers zou kunnen sturen, zouden er niet voor zorgen dat je op de juiste manier aan het stuur draaide. De meeste van de patronen in je brein zouden tot wilde spiertrekkingen en spasmen leiden: het zou eruitzien als een epileptische aanval. Toch weet je brein elke dag zenuwimpulsen te vinden waarmee je een auto bestuurt of die ervoor zorgen dat je rechtop staat. Het kiest de juiste mogelijkheden in plaats van allerlei verkeerde. Wanneer je naar het vliegveld rijdt, berekent je brein niet alleen een precieze reeks afslagen waarmee je je bestemming bereikt, maar selecteert het ook een van de ontelbare patronen van zenuwimpulsen die je spieren op de juiste manier samentrekken om aan het stuur te draaien.

Voorspellen en aansturen zijn onderling met elkaar verbonden. Als je met een auto naar het vliegveld rijdt, moet je voorspellen via welke straten je op Airport Boulevard belandt. Bij het voorspellen welke straten naar Airport Boulevard leiden, moet je je vingers aansturen om een kaart op je telefoon te gebruiken.

Volgens ons is er nog steeds een fundamenteel verschil tussen voorspellen en aansturen, een verschil dat heel belangrijk zal blijken te zijn.

Goed kunnen voorspellen is eenvoudig meetbaar. Als iemand verwacht Airport Boulevard voor zich te zien, maar in plaats daarvan Second Street ziet, was de voorspelling onjuist.

Om het succes van aansturen te bepalen, moet je echter enig idee hebben *waar hij naartoe probeerde te gaan.*

Als iemands auto bij de supermarkt belandt is dat geweldig als hij probeerde boodschappen te doen. Als hij de Spoedeisende Hulp probeerde te bereiken is hij daarin niet geslaagd.

Als twee inwoners van dezelfde stad slimmer worden, zou je verwachten dat ze het steeds vaker eens zijn over vragen over een voorspel-

ling, bijvoorbeeld of het op werkdagen om vijf uur 's middags meestal druk is op Second Street. Maar je zou niet verwachten dat ze gaan aansturen op dezelfde plekken; zo gaat de een liever naar het park en de ander naar het theater.

Anders gezegd: *intelligente geesten kunnen aansturen op verschillende einddoelen* zonder dat dat iets zegt over hun intelligentie.

Voorspellen en aansturen zijn geen functies die alleen biologische geesten hebben: machines kunnen dat ook. Maar op dit moment zijn mensen nog steeds de besten op deze planeet in...

Ja, waarin precies? Mensen zijn niet langer de wereldkampioen schaken. Mensen zijn niet langer de enige taalgebruikers op de planeet. Mensen zijn niet langer de enigen die een medische grafiek kunnen lezen of een tumor diagnosticeren.

Mensen zijn nog wel kampioen in iets wat veel dieper gaat, maar dat speciale 'iets' is tegenwoordig lastiger te beschrijven dan vroeger.

Volgens ons hebben mensen nog steeds een voorsprong op het gebied van iets wat je 'algemeenheid' zou kunnen noemen. Wat betekent dat precies? Wij zouden zeggen: intelligentie is *algemener* naarmate ze op meerdere gebieden kan voorspellen en aansturen. Mensen zijn niet in alle dingen het beste: zo zal een inktvissenbrein beter zijn in acht armen aansturen. Maar in bredere zin lijkt het overduidelijk dat mensen algemener kunnen denken dan inktvissen. Wij hebben grotere gebieden waarop we met succes kunnen voorspellen en aansturen.

Sommige AI's zijn op beperkte gebieden slimmer dan wij. In 1997 versloeg Deep Blue, de supercomputer van IBM, als eerste machine een menselijke wereldkampioen tijdens een schaakpartij. Op schaakgebied was Deep Blue heel vaardig in voorspellen en aansturen. Maar Deep Blue kon niet voorspellen hoe je bij de supermarkt komt en melk koopt, laat staan hoe je de auto bestuurt waarmee je daarnaartoe rijdt. Geesten kunnen meer of minder vaardig zijn op verschillende gebieden van voorspellen en aansturen.

Nieuwere AI's hebben veel algemenere capaciteiten. Aan het o1-model van OpenAI kun je vragen welke temperatuur de aarde zou

hebben als het zonlicht in infrarood veranderde, en vervolgens zal o1 het antwoord achterhalen door natuurkundige berekeningen te doen. Vervolgens kun je vragen of de mensheid in die nieuwe wereld voedsel zou kunnen verbouwen, en zal o1 antwoord geven vanuit zijn kennis van plantkunde. Hij switcht niet tussen twee verschillende databases onder de motorkap, hij weet gewoon van zowel natuurkunde als biologie.

OpenAI's o1 weet dat de buitenwereld bestaat en kan daarover redeneren. Deep Blue had geen idee. Het heeft decennia geduurd tot AI dat punt had bereikt.

Toch bevindt het algemene logische denkvermogen van o1 zich in zekere zin niet op menselijk niveau. Mensen staan in de technologie en wetenschap nog steeds bovenaan: de grote doorbraken zijn het werk van menselijke onderzoekers, (nog) niet van AI's. Bovendien voelt het nog steeds – in elk geval voor deze twee schrijvers – alsof o1 minder intelligent is dan zelfs mensen die geen grote wetenschappelijke doorbraken bewerkstelligen. Het wordt steeds moeilijker om de vinger te leggen op wat er ontbreekt, maar wij hebben het gevoel dat hoewel o1 meer weet en onthoudt dan welke individuele mens ook, hij in een of ander belangrijk opzicht nog steeds 'oppervlakkig' is vergeleken met een mens van twaalf jaar oud.

Dat zal niet eeuwig zo blijven. Het is moeilijk te voorspellen hoe snel AI zich zal verbeteren en welke route ze zal volgen, maar het eindpunt is wel makkelijk te voorspellen, omdat de begrenzingen van de technologie machines veel voordelen biedt ten opzichte van biologische hersenen. Om slechts enkele te noemen:

Pure snelheid. Transistors, een fundamentele bouwsteen van alle computers, kunnen miljarden keren per seconde aan- en uitgaan; zelfs ongewoon snelle neuronen pieken daarentegen maar honderd keer per seconde. Stel dat er duizend transistorhandelingen voor nodig waren om het werk van één neurale piek te doen, en stel dat artificiële intelligentie beperkt was tot moderne hardware, dan nog impliceert dat dat denken op menselijk niveau tienduizend keer zo snel kan