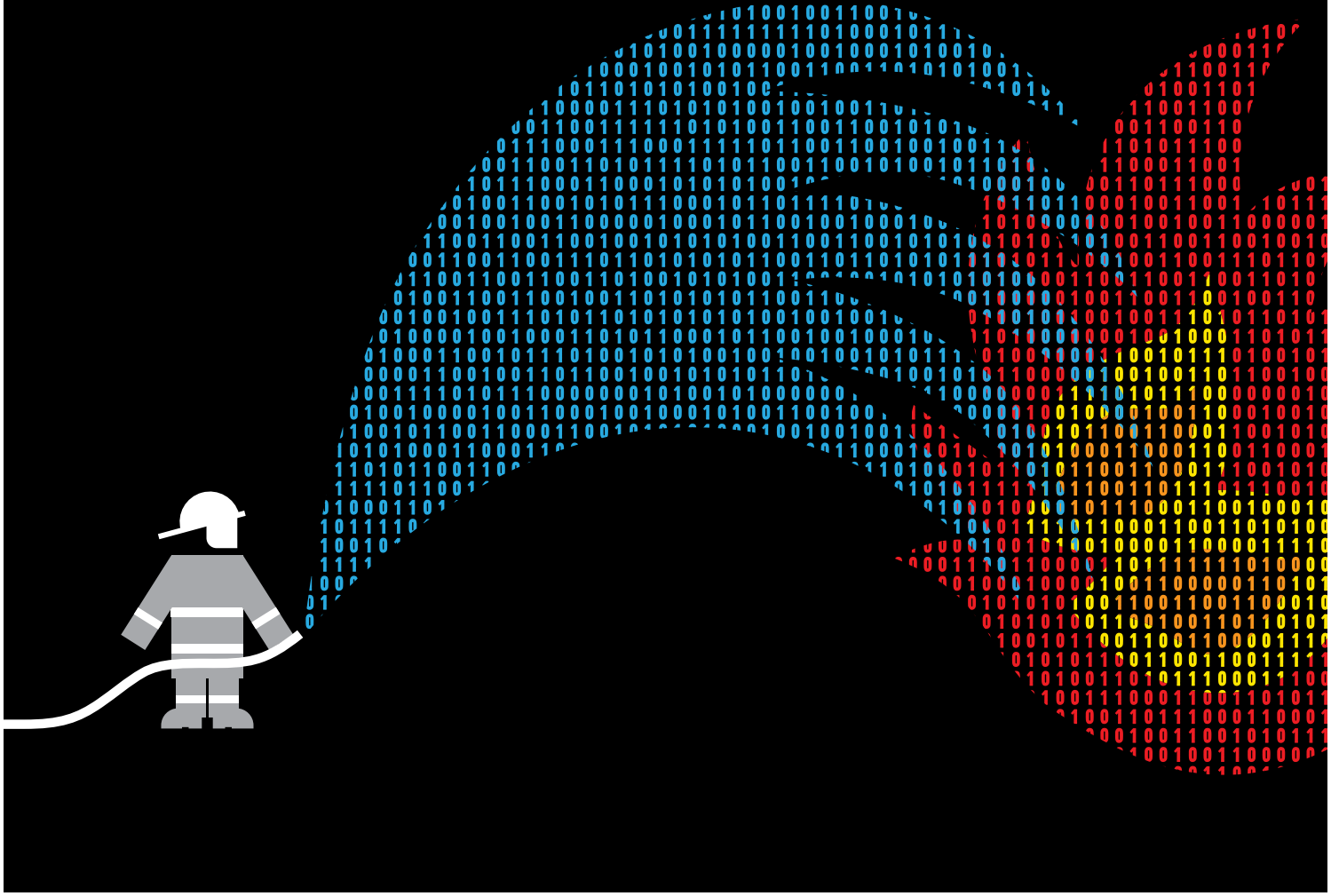
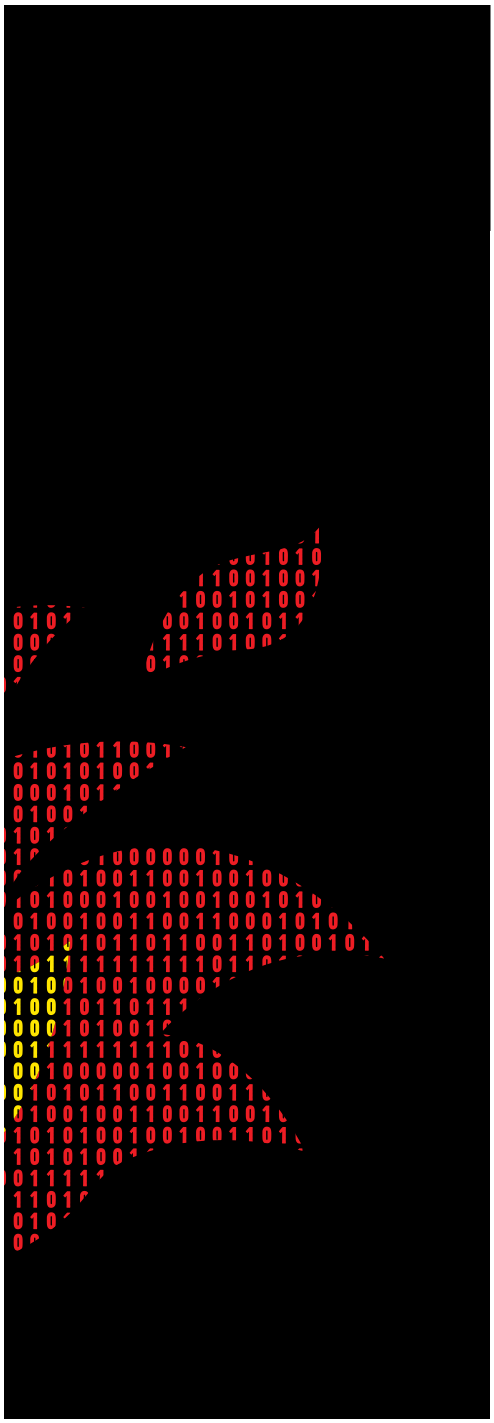
De opmars van de existential risk-beweging

Losgeslagen superintelligentie



Brits en Amerikaans AI-beleid wordt beïnvloed door de sterke lobby van de x-risk-beweging. Deze focust op het risico van het uitsterven van de mensheid door toekomstige superintelligentie. Maar volgens critici speelt deze benadering vooral Big Tech in de kaart.

Colin van Heezik beeld Pepijn Zurburg



Wie het een jaar geleden voorspeld had, was waarschijnlijk voor gek verklaard: dat de Britse premier in een toespraak zou beginnen over het uitsterven van de mensheid door kunstmatige intelligentie, en hoe belangrijk het is om de kansen daarop te beperken. En toch was dat wat Rishi Sunak deed in de aanloop naar de AI Safety Summit van 1 en 2 november, een topconferentie in het VK over de risico's van AI.

In zijn speech citeerde Sunak het zogenaamde 'extinction-statement': een verklaring die op 30 mei ondertekend werd door vele prominenten uit de techindustrie, onder wie Bill Gates en OpenAI-baas Sam Altman, maar ook door AI-wetenschappers als Turing Prize-winnaar Yoshua Bengio. 'Het beperken van het risico op uitsterven door AI zou een wereldwijde prioriteit moeten zijn naast andere maatschappelijke risico's zoals pandemieën en nucleaire oorlog', zo luidde het statement, waar Sunak mee instemde.

Hoe is dat zo gekomen? Daarover publiceerde *Politico* in september het artikel *How Silicon Valley Doomers Are Shaping Rishi Sunak's Plans*. Sunak is volgens onderzoek van *Politico* beïnvloed door de *existential risk*-beweging: mensen die waarschuwen voor *'x-risk'*, het risico op *extinction* door AI.

Het meest klassieke rampscenario werd geschetst door de Zweedse filosoof Nick Bostrom, die in 2003 in een artikel beschreef wat er zou kunnen gebeuren als een superintelligente AI de opdracht zou krijgen zoveel mogelijk paperclips te maken. 'De AI zal snel beseffen dat het veel beter zou zijn als er geen mensen waren, omdat mensen zouden kunnen besluiten om het uit te schakelen', zei Bostrom in een interview in 2014 over zijn gedachte-experiment. 'Ook bevatten menselijke lichamen veel atomen die omgezet kunnen worden in paperclips. De toekomst waar de AI naar zou streven, zou er een zijn waarin er veel paperclips maar geen mensen zijn.'

Bostroms gedachte-experiment wordt de *Paperclip Maximizer* genoemd. AI-wetenschappers kennen het allemaal, en sommige techneuten van AI-labs in Silicon Valley, zoals OpenAI, DeepMind of Anthropic, dragen truien met een afbeelding van een paperclip, als knipoog naar het beroemdste voorbeeld van wat 'existentieel risico' genoemd wordt: het gevaar dat toekomstige AI zal leiden tot het uitsterven van de mens. Maar er zijn ook andere rampscenario's – die gaan bijvoorbeeld over autonome AI-wapensystemen of wat er zoal kan gebeuren als een superintelligente AI zich toegang verschaft tot een biotechlab om een dodelijk virus te creëren.

'Het is moeilijk te voorspellen hoe een wereld eruit zou zien met een andere, intelligentere soort dan de mens', zei een activist dit jaar tegen techtijdschrift *Wired*. 'Maar we weten dat onze relatie met soorten die minder intelligent zijn dan wij niet geweldig is geweest voor die andere soorten.'

Een veel gebruikte term is *'rogue AI'*. Veel AI-wetenschappers vrezen het toekomstscenario van zulke losgeslagen AI, vooral nu de ontwikkeling van AI in een stroomversnelling is beland. Turing Prize-winnaar Bengio schreef in *The New York Times:* 'De systemen van vandaag vormen nog lang geen existentieel gevaar. Maar over één, twee, vijf jaar? Er is te veel onzekerheid. Dat is het probleem. We weten niet zeker dat dit niet een punt zal bereiken waarop dingen catastrofaal worden.'

De Nederlandse activist en techondernemer Joep Meindertsma, initiatiefnemer van de website Pause AI, is blij met alle aandacht voor de ultieme gevaren van AI. Zijn Pause AI-beweging is dit jaar flink gegroeid, in verschillende landen. Zo demonstreerden ze in mei voor de deur van OpenAI in San Francisco en ook bij Google's AI-lab DeepMind in Londen.

Termen als *'p-doom'* liggen hem in de mond bestorven. In Silicon Valley lopen AI-techneuten rond die zeggen: 'Mijn p-doom is twintig procent', wat betekent dat ze de kans op extinctie door AI op twintig procent inschatten. Meindertsma denkt dat mogelijk al binnen een jaar een AI kan bestaan die de potentie heeft om het einde van de mens in te luiden.

Computerwetenschapper Roel Dobbe van TU Delft denkt er anders over. 'Het verhaal over x-risk leidt de aandacht af van de bestaande risico's', zegt de onderzoeker. Hij leerde Silicon Valley van nabij kennen toen hij in de San Francisco Bay Area woonde. 'De gevaren van superintelligentie vormen een handig frame voor de techbedrijven om de aandacht te verleggen van huidige problemen naar potentiële problemen in de toekomst, om regulering te voorkomen.'

Dobbe is medeoprichter van het AI Now Institute in New York, samen met Timnit Gebru, de wetenschapper die door Google ontslagen werd na zijn kritiek op racistische patronen in grote taalmodellen. Het AI Now Institute is een van de vele ngo's gericht op democratie en mensenrechten in de digitale sector – Nederlandse voorbeelden zijn Bits of Freedom en Waag Future Lab. In deze ngo-kringen wordt het verhaal over x-risk gezien als een bliksemafleider voor de techindustrie om regelgeving in het heden te ontlopen.

Dat is ook de strekking van het verhaal dat

techtijdschrift *MIT Technology Review* in juni aan het onderwerp wijdde: *How Existential Risk Became the Biggest Meme in AI*. Techbedrijven, stelt het stuk, verplaatsen de bezorgdheid over AI graag naar de toekomst, waarin hun huidige belangen niet in het geding zijn.

'Als we praten over de verre toekomst en mythologische risico's, herformuleren we het probleem als een probleem dat bestaat in een fantasiewereld', zegt Meredith Whittaker, oprichter van research-lab AI Now Institute, in het artikel. 'De oplossingen kunnen dan ook alleen maar bestaan in die fantasiewereld.'

Daarmee wordt de aandacht afgeleid van AI-risico's die er nu al zijn, stellen ook critici als Margaret Mitchell (Hugging Face) en Emily Bender (University of Washington) van AI-onderzoeksinstituut DAIR. 'Die hypothetische risico's zijn de obsessie van een gevaarlijke ideologie genaamd *longtermism* die de daadwerkelijke schade negeert van AI-systemen in het heden', schreven zij over een brandbrief die in hun ogen de plank missloeg: de pauzepetitie van Future of Life. Die werd op 28 maart ondertekend door duizenden prominenten, onder wie Elon Musk.

'De brief', schreef DAIR, 'gaat niet in op de voortdurende schade die veroorzaakt wordt door deze systemen, zoals 1) uitbuiting van werknemers en massale datadiefstal om producten te creëren die een handjevol partijen winst opleveren, 2) de stortvloed van synthetische media in de wereld, wat zowel systemen van onderdrukking reproduceert als ons informatie-ecosysteem in gevaar brengt, en 3) de concentratie van macht in handen van een paar mensen, wat sociale ongelijkheden verergert.'

In de ogen van DAIR is waarschuwen voor vooral door opportunisme. 'Er worden miljoenen geïnvesteerd in denktanks over dit onderwerp en studenten van topuniversiteiten worden via prijzen en studiebeurzen verleid zich hiermee bezig te houden.' Over die laatste trend publiceerde *The Washington Post* dit jaar een groot verhaal: *How Elite Schools Like Stanford Became Fixated on the AI Apocalypse*. Daarin onthullen de journalisten hoe 'een beweging van miljardairs studenten rekruteert om dodelijke AI te bestrijden'. Een stichting van Facebook-medeoprichter Dustin Moskovitz, Open Philanthropy genaamd, heeft al bijna een half miljard vrijgemaakt om talent te werven voor deze missie. Studenten kunnen fellowships van tachtigduizend dollar per jaar krijgen als ze kiezen voor dit nieuwe vakgebied, dat sommigen zien als 'het nieuwe Manhattan Project'.

Het geld komt vooral uit de hoek van de Effective Altruists, een filantropische beweging van techmiljardairs uit Silicon Valley. Bekende leden zijn Elon Musk, Peter Thiel (oprichter van PayPal) en de gevallen cryptoprins Sam Bankman-Fried. Zij komen regelmatig bijeen om zich met elkaar af te vragen: hoe kun je als filantroop optimaal goed doen met je geld? Dan moet je kijken naar de lange termijn, zo redeneren de EA'ers. En zo komen ze uit bij het ultieme goede doel: het voorkomen van het uitsterven van de mens. Vandaar dat deze weldoeners hun miljoenen graag uitgeven aan onderzoek naar x-risk.

Studie of werk op dit gebied wordt door de EA-beweging, die ook in Nederland voet aan wal tracht te krijgen, aangeprezen als de ideale carrière. Vacatures in dit veld worden op de EA-platforms gepubliceerd. De Nederlandse student Otto Barten meldde een paar jaar geleden op zo'n platform dat hij een Nederlandse x-risk-club had opgericht. Inmiddels wordt zijn Existential Risk Observatory gesponsord door de Finse techmiljardair, Skype-oprichter en EA-prominent Jaan Tallinn. Barten heeft er nu een fulltime baan aan. In juni verscheen in Nederland een brandbrief waar het omstreden begrip x-risk in voorkwam: de petitie 'Politiek, neem controle over AI', die door vele prominenten, onder wie Bas Heijne en Sander Schimmelpenninck, werd ondertekend. Pause AI-oprichter Meindertsma was een van de initiatiefnemers. 'Ik had een eerste versie van die petitie geschreven waarin existentieel risico een grotere plek innam', vertelt Meindertsma nu. 'Maar daarmee zouden we minder ondertekenaars bereid hebben gevonden. Daarom hebben we het zwaartepunt veel meer gelegd bij de bestaande risico's zoals vooroordelen, desinformatie en privacygevaren.'

De tekst was dus een soort Hollands compromis, maar Meindertsma ziet geen spanning tussen aandacht voor actuele gevaren en existentieel risico: 'Mensen die bezorgd zijn over x-risk hebben ook veel oog voor de andere gevaren.'

In Nederland weigerden organisaties voor technologie en mensenrechten zoals Waag Future Lab en Bits for Freedom niettemin de brandbrief te ondertekenen, zó allergisch is men in het ngo-milieu voor x-risk.

In april publiceerde de Britse techondernemer Ian Hogarth in de *Financial Times* het opiniestuk *We Must Slow Down the Race to God-like AI*, waarin hij het gevaar beschreef dat AI 'het verval of de vernietiging van het menselijk ras zou kunnen inluiden' – hoe vaker je het hoort, hoe minder erg het klinkt. Maar inmiddels is x-risker Hogarth door de Britten benoemd tot hoofd van de Britse Frontier AI Taskforce, die AI-risico's onderzoekt.

En zo belandde x-risk op de agenda van de Britse regering. In zijn toespraak van 26 oktober citeerde Sunak het *'extinction-statement'* van 30 mei, met het woord *'Indeed'* ervoor. Hij noemde nog wel andere risico's van AI, zoals desinformatie, vooroordelen, cyberaanvallen, autonome wapens. Maar hij benadrukte dat hij al deze gevaren vooral wilde *onderzoeken*, in een speciaal AI Safety Institute. De Britse aanpak, zei Sunak er nog bij, is 'niet overhaast te reguleren'.

Daarmee maakt Sunak het voor techbedrijven aantrekkelijk zich in het VK te vestigen, waar de AI Act van de EU, die tal van regels bevat om concrete risico's voor de samenleving tegen te gaan, niet van toepassing is. Het VK kan zijn eigen koers varen op het punt van veiligheid en

De gemeenschappen die het meest getroffen zijn door AI, werden niet uitgenodigd voor de conferentie over de risico's van AI

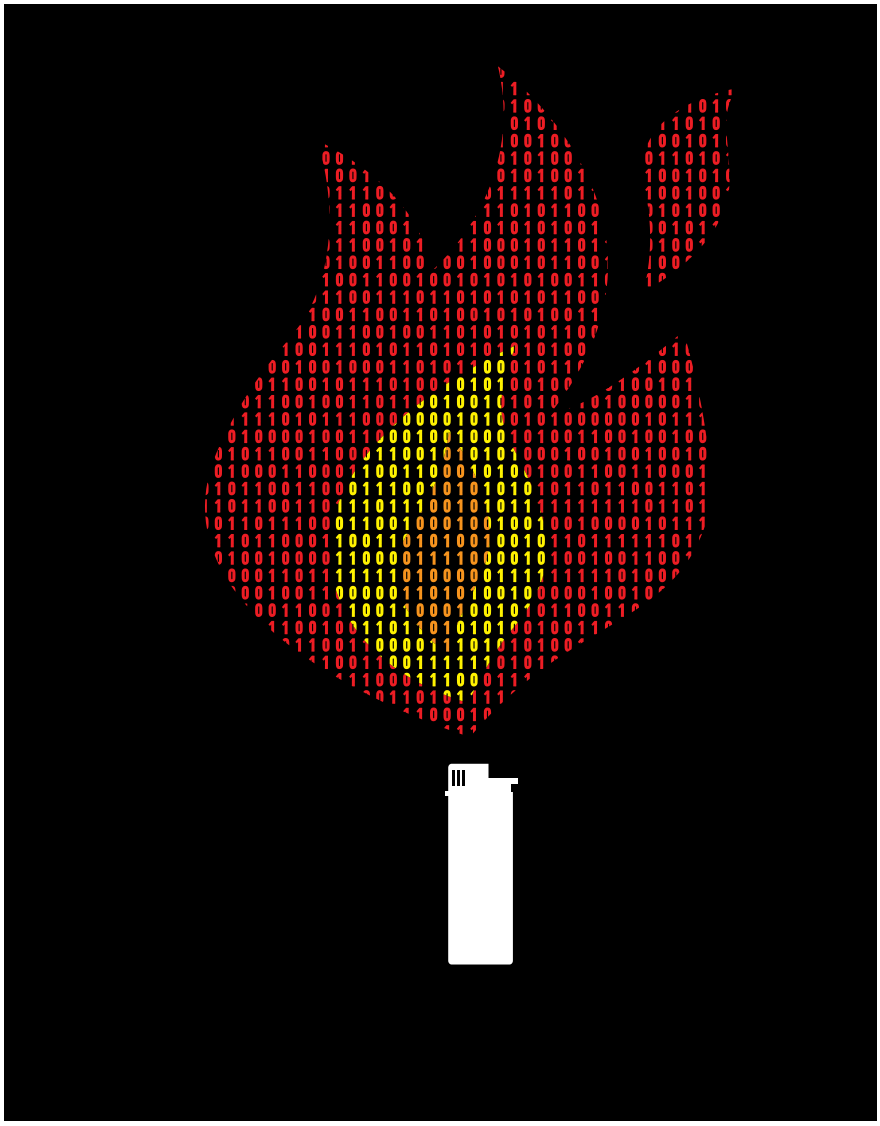
x-risk onderdeel van dezelfde pr-goochelarij als zeggen dat AI een utopische toekomst zal brengen: een marketingtruc, waarbij de potentie van AI in twee richtingen tegelijk wordt overdreven. 'Het is gevaarlijk om onszelf af te leiden met een gefantaseerde AI-gestuurde utopie of apocalyps, die dan wel een "bloeiende" dan wel een "potentieel catastrofale" toekomst belooft.'

Een gezaghebbend techcolumnist als Brian Merchant van de *Los Angeles Times* ziet het precies zo: 'Ik geloof geen seconde dat AI op het punt staat krachtig genoeg te worden om de mensheid te vernietigen – dat zijn mensen in Silicon Valley die zich laten meeslepen door een sciencefictionachtig gevoel van zelfbelangrijkheid, denk ik, en een uniek sluwe marketingtactiek.'

Dat er zoveel aandacht uitgaat naar x-risk, komt volgens computerwetenschapper Roel Dobbe bescherming van burgers tegen de gevaren van AI. De benadering vanuit x-risk is daarbij opportuun, want is immers niet gericht op regulering maar op onderzoek naar veiligheid op de lange termijn.

De topconferentie vond begin november plaats in Bletchley, waar grondlegger van de computerwetenschap Alan Turing tijdens de Tweede Wereldoorlog de code kraakte waarmee de nazi's hun militaire communicatie versleutelden. En jawel, ook Elon Musk was van de partij. Sunak interviewde hem ter inleiding op de top over AI-veiligheid. Volgens sommigen een opmerkelijke keuze: was Musk niet de man die Twitter onveiliger maakte en bekend stond om zijn dodelijke zelfrijdende auto's?

Voorafgaand aan de top was er een boze brief van vakbonden en mensenrechtenorganisaties. Waarom waren zij niet uitgenodigd? 'De



gemeenschappen en werknemers die het meest getroffen zijn door AI zijn gemarginaliseerd door de top', schreven de geweigerden van het Britse AI-bal. 'Dit is een gemiste kans.'

Sunak dacht daar waarschijnlijk anders over. Tijdens de top ging hij met OpenAI-baas Sam Altman op de foto. Na afloop vonden critici dat de politici veel te weinig daadkracht tot reguleren hadden getoond, maar de Britse techminister Michelle Donelan zei dat haar regering graag 'de juiste balans wilde vinden' tussen veiligheid en innovatie. Dat die benadering goed werkt, bleek volgens haar uit het feit dat AI-labs als Anthropic en OpenAI inmiddels Europese hoofdkantoren geopend hebben in het VK.

Ook in de VS is de lobby van x-risk al zeer succesvol. Politici en beleidsmakers in Washington, zo bracht *Politico* recent aan het licht, leunen op experts van denktanks die indirect betaald worden door EA-miljardairs. De geldstroom loopt bijvoorbeeld via de filantropische EA-stichting Open Philanthropy, dat miljoenen steekt in diverse x-risk-denktanks, zoals het in 2022 opgerichte non-profit Horizon Institute for Public Service. Die leveren AI-experts die invloed uitoefenen op beleid of zelfs concepten voor regulerende wetteksten schrijven. Zodoende ontstaat een getrapt systeem waarbij de invloed van de techindustrie op beleid, de zogenaamde *regulatory capture,* moeilijk te herleiden valt.

De invloed op politici is al zichtbaar. Niet alleen Rishi Sunak maar ook de Europese Commissie en de regering van Joe Biden hebben x-risk dit jaar erkend. Gezien de sterke lobby is het waarschijnlijk dat die invloed de komende tijd nog verder zal gaan. Een techkopstuk als Altman wordt daarbij hypocrisie verweten: terwijl hij met politici praatte over de risico's van AI, spande OpenAI zich achter de schermen in om regulering te beperken. Door te lobbyen bij de EU, bleek uit onderzoek van *Time,* kreeg OpenAI het voor elkaar dat ChatGPT in de Europese AI Act niet als *'high risk'* werd aangemerkt en zodoende niet aan de strengste regels onderworpen werd.

Ondertussen breidt het netwerk van x-risk-denktanks, onderzoeksinstituten en lobbyclubs zich steeds verder uit. Daartoe behoren onder meer het Stanford Existential Risks Initiative in Californië, het Future of Humanity Institute in Oxford en het Existential Risk Observatory in Nederland. Het eerste instituut werd opgericht in 2005 door onder meer Elon Musk. Jaan Tallinn is medeoprichter en financier van het Future of Life Institute (initiator van de Pause AI-brief van 22 maart 2023), het aan de Universiteit van Cambridge gelieerde Centre for the Study of Existential Risk en het Nederlandse Existential Risk Observatory.

Veel jonge x-risk-adepten, zoals de Nederlandse Joep Meindertsma en Otto Barten, zijn naar eigen zeggen beïnvloed door lezingen van Nick Bostrom, die werkt bij het Future of Humanity Institute. Zij zien in hem een Oxford-hoogleraar die waarschuwt voor x-risk, maar wat ze niet altijd beseffen is dat Bostrom, auteur van *Superintelligence: Paths, Dangers, Strategies* (2014), verbonden is aan een privaat onderzoeksinstituut, dat is gelieerd aan Oxford maar is opgericht en gefinancierd door techmiljardairs.

Dat het geld uit die hoek komt, betekent overigens nog niet dat er sprake is van een samenzwering van techmiljardairs. Eerder zijn ze verwikkeld in competitie en concurrentie, ook op het vlak van AI-veiligheid. OpenAI werd opgericht in 2015 door onder anderen Elon Musk met het geafficheerde doel om AI te maken die *'beneficial to all of humanity'* zou zijn. Dat klonk vrij idealistisch, maar het was ook een poging om het AI-monopolie van Google te doorbreken.

Op dit moment is er stevige competitie tussen OpenAI en Anthropic, het AI-lab dat in 2021 werd opgericht door mensen die waren weggelopen bij OpenAI. Anthropic, dat ten dele in handen is van Tallinn, probeert OpenAI te beconcurreren door in te zetten op veilige, verantwoorde AI. In een reportage over Anthropic beschreef journalist Kevin Roose van *The New York Times* Anthropic-medewerkers als bedrukt en diep bezorgd over x-risk. 'We hopen dat er een *safety race* komt', zei Ben Mann, een van de mede-oprichters van Anthropic, tegen Roose.

Als het Anthropic lukt om OpenAI op het punt van veiligheid af te troeven, kan de aandacht voor AI-veiligheid die Tallinn met zijn denktanks genereert voor hem dus ook commercieel interessant uitpakken. Hetzelfde geldt voor de andere grote Anthropic-geldschieter van het eerste uur, Dustin Moskovitz: de man achter Open Philanthropy, de stichting die Amerikaanse denktanks over x-risk financiert, die invloed uitoefenen op het AI-beleid van de overheid.

Het lijkt er trouwens op dat die safety race al

Schenk het licht van de kennis

met een cadeau-abonnement op Apache

Zoek je een echt waardevol geschenk voor iemand die alles al heeft?

Wil je iets origineels geven aan een naaste die geen boodschap heeft aan badparels, kruidenolie, waardebons die de prijs van het uitstapje niet helemáál dekken, of “grappige” in lagelonenlanden geproduceerde prullen?

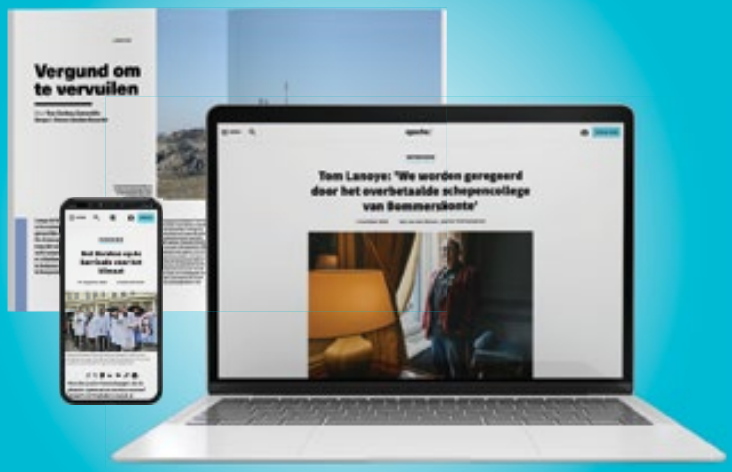
Doe je niet mee aan de commerciële kermis en het verplichte nummer van de cadeautjes-ruilbeurs waartoe de eindejaarsfeesten verworden zijn?

Ken je iemand die graag goed geïnformeerd wordt, zonder op elke pagina reclame, gossip of lifestylenonsens te worden gedwangvoederd?

Schenk dan dit jaar het licht van de kennis, met een abonnement op Apache vanaf 50 euro.

Dat is een jaar lang leesplezier, onthullende onderzoeken, verhelderende analyses en blikverruimende interviews uit een betrouwbare, onafhankelijke bron: als dat geen *gift that keeps on giving* is.

Ga naar apache.be/cadeau en pak uit met een cadeauverpakkingsvrij geschenk met meerwaarde.



apache.be/cadeau

begonnen is: terwijl Anthropic inzet op veiligheid, kondigde OpenAI in juli een onderzoeksprogramma naar *super alignment* aan, waarover hoofdingenieur Ilya Sutskever, de architect van ChatGPT, de leiding zou nemen. Dit programma richt zich op alignment van toekomstige superintelligentie: zorgen dat die zich niet tegen de mens keert.

Na de lancering van ChatGPT in november vorig jaar uitten vele vooraanstaande AI-wetenschappers en intellectuelen grote zorgen over de risico's van AI.

Die vallen in grofweg drie groepen uiteen. Ten eerste zijn er de mensen die al jaren op de risico's wijzen: AI-onderzoeker en essayist Eliezer Yudkowsky (schreef een alarmerend stuk in *Time* dat viral ging) en de hoogleraren Nick Bostrom (Oxford) en Stuart Russell (Berkeley), auteur van *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

Ten tweede zijn er de AI-wetenschappers zoals Yoshua Bengio en Geoffrey Hinton, die de systemen zelf gebouwd hebben en dit jaar opeens hun jasje hebben omgedraaid en waarschuwen tegen de gevaren van hun schepping.

Ten derde zijn er intellectuelen, zoals Yuval Noah Harari, die dit jaar in de pen klommen om hun zorgen te delen. Sommigen achten AI in bepaalde opzichten gevaarlijker dan atoomwapens. 'Kernbommen kunnen tenminste geen nieuwe kernbommen bouwen', schreef Harari in *The Economist*, 'maar AI-systemen kunnen wel nieuwe AI-systemen maken.'

Zulke zorgen leven ook binnen de AI-industrie, wat nog niet betekent dat die het uiteindelijk winnen van commerciële prikkels. Zo zien we hoe bij OpenAI de geafficheerde oprichtingsidealen steeds meer plaats maken voor zakelijke ambities. Op 17 november werd Sam Altman ontslagen als CEO door het bestuur van OpenAI. Volgens sommige analyses legde die gebeurtenis een spanningsveld bloot binnen de techindustrie zelf, die intern verdeeld zou zijn over AI-veiligheid. OpenAI werd opgericht als non-profit, maar gaandeweg werd dat model vervangen door een hybride structuur. Want om koploper op AI-gebied te kunnen zijn is nu eenmaal veel geld nodig, en investeerders pompen geen miljarden in een non-profit. Maar de commerciële aanpak van Altman zou een doorn in het oog zijn van hoofdingenieur Ilya Sutskever en enkele EA'ers in het bestuur. Zij zouden Altman hebben onttroond omdat diens expansiedrift niet meer in lijn zou zijn met de grondslagen van OpenAI.

Altnam vertrok meteen naar Microsoft en nam topingenieurs Greg Brockman en Jakub Pachocki mee naar de techreus, waar ze een nieuwe AI-tak zouden gaan opzetten. Vijf dagen lang zaten OpenAI'ers te nagelbijten, terwijl hun bedrijf in een vrije val beland leek. Ondertussen stuurden ze elkaar scènes uit de serie *Succession*. Na vijf dagen kwam de ontknoping: onder druk van Microsoft, dat 49 procent van de aandelen van OpenAI heeft, en na een protestbrief van werknemers werd Altman toch weer teruggehaald als CEO. Het bestuur werd deels vervangen. Elke illusie dat iets anders dan geld centraal zou staan in de AI-industrie, schreef *Los Angeles Times*-columnist Merchant, was daarmee vervlogen: binnen een paar dagen was de verzetsdaad vermorzeld door Big Tech. De triomf van het kapitaal was een feit.

'Ze kunnen er een filmscript over schrijven', zegt computerwetenschapper Dobbe over het koningsdrama. 'Een paar bestuursleden waren Effective Altruists. En het lijkt er toch op dat zij zich samen met Sutskever verzet hebben tegen de commerciële koers van Altman. Alhoewel ik niet veel op heb met het veiligheidsperspectief van Sutskever en zijn team, lijkt het erop dat er daar een behoefte was minder gehaast te commercialiseren.'

Volgens Meindertsma bewijst het drama bij OpenAI dat er geen samenzwering is in de techindustrie, maar juist een strijd tussen de mensen die geld willen verdienen en de wetenschappers die bezorgd zijn over de risico's. Hij vond het 'heel goed nieuws' dat EA'er Emmett Shear (oprichter van gamingplatform Twitch) door OpenAI was aangesteld als nieuwe CEO, iemand met 'een p-doom van vijf tot vijftig procent'. Maar dat was dus voordat Altman weer terugkwam.

Dat Sutskever bezorgd is over AI-veiligheid bleek al uit interviews voor de documentaire *iHuman* in 2019. 'We zullen zeker in staat zijn compleet autonome wezens met eigen doelen te scheppen', aldus Sutskever in de minidocu. Hij acht het zeer waarschijnlijk dat die systemen 'een totaal astronomische impact op de samenleving zullen hebben', vooral als ze veel slimmer zullen zijn dan mensen. We zijn eigenlijk een nieuwe soort aan het creëren, denkt Sutskever, en moeten zorgen dat die goed geprogrammeerd is. 'Zo niet, dan denk ik dat natuurlijke selectie die systemen zal bevoordelen die hun eigen overleving boven alles stellen.'

Sutskever is niet de enige bezorgde ingenieur. AI-makers schatten de kans op uitsterven door toekomstige AI groot in, volgens sommige enquêtes tussen de tien en dertig procent. Critici als Roel Dobbe vinden dat er geen enkel wetenschappelijk bewijs voor is. 'Maar hoe belangrijk is het dat daar nog geen wetenschappelijk bewijs voor is', zegt Meindertsma, 'gezien de ernst van de risico's?' Dat de scenario's voorstelbaar zijn, vindt hij genoeg reden voor onderzoek en voorzorg.

Meindertsma spreekt trouwens liever van 'catastrofale risico's', een bredere categorie die naast x-risk ook andere rampscenario's omvat. 'Ik ben het meest bezorgd om het scenario waarbij kwaadwillenden met AI allerlei systemen kunnen gaan hacken. Dat kan al gebeuren op relatief korte termijn. Je hebt geen superintelligentie nodig die de wereld kan overnemen. Je hoeft alleen maar een AI te hebben die beter is in het vinden van security exploits dan een goede hacker. Daarmee kan iedereen een virus schrijven dat het grootste deel van de apparaten op de wereld platlegt of infecteert. Onze afhankelijkheid van het internet is zo groot voor al onze supply chains, voor voedsel, voor al onze banen, communicatie, samenwerking, dat ik me hier grote zorgen over maak.'

In zijn toespraak beloofde premier Sunak investeringen in AI-veiligheid. Maar wat voor oplossingen zijn er denkbaar? 'Ook hier slaat de x-risk-benadering de plank mis', vindt Dobbe. 'Want ten eerste wordt het risico verkeerd en te nauw gedefinieerd: uitsterven door superintelligentie. Maar ten tweede denkt men die te kunnen fixen met "alignment": zorgen dat het AI-systeem zich gedraagt op een manier die in lijn is met menselijke waarden. Maar kun je veiligheid inbouwen in een AI-systeem? Neem het voorbeeld van een atoombom. Er is geen veilige atoombom. Zelfs geen veilige hamer. Het gaat om het gebruik, in context. Dat is hoe de bestaande kennis over systeemveiligheid het benadert. Maar die kennis wordt door de x-risk beweging opzijgeschoven.'

'Om alles plat te leggen, heb je geen superintelligentie nodig. Je hoeft alleen maar een AI te hebben die beter is dan een goede hacker'

Wat zou er dan wel moeten gebeuren, volgens Dobbe? Hij kijkt kritisch naar onderzoeksteams als de Britse AI Taskforce onder leiding van techondernemer Ian Hogarth, die zwaar leunt op de x-risk-beweging. 'Alle soorten risico's en benaderingen van AI-veiligheid zouden multidisciplinair onderzocht moeten worden, niet alleen door informatici van of gelieerd aan techbedrijven, maar in onafhankelijke en divers samengestelde teams', zegt Dobbe. 'Anders krijg je een aanpak die ik *safety washing* noem: doen alsof je je heel druk maakt om veiligheid, zoals Sunak, terwijl je eigenlijk alle ruimte biedt voor innovatie en commercie.' ■