Anthropic presenteert zich als het braafste jongetje van de roekeloze AI-industrie: is dat terecht?

Jorit Verkerk

Op 11 juni 1945 schreven wanhopige wetenschappers een brief waarin ze de Amerikaanse regering smeekten de zojuist ontwikkelde atoombom niet op een Japanse stad te laten vallen. De gevolgen zouden immers catastrofaal zijn. Onder hen was de Hongaars-Amerikaanse natuurkundige Leó Szilárd (1898-1964), in de oorlogsjaren een belangrijke aanjager van het Amerikaanse kernwapenprogramma. Nu was hij de controle over zijn levenswerk kwijtgeraakt: de bommen op Hiroshima en Nagasaki vielen toch. Hij zou zijn hele leven spijt van zijn betrokkenheid blijven hebben.

Bijna 81 jaar later waarschuwt de leider van het Amerikaanse Anthropic, één van de grootste AI-bedrijven ter wereld, in een recent essay dat de mensheid „op het punt" staat „bijna ondenkbare macht" in handen te krijgen. Hij doelt op de AI-systemen die zijn bedrijf in duizelingwekkend tempo ontwikkelt, in wat hij framet als een race naar superintelligentie. Anthropic biedt particulieren taalmodel Claude aan, en AI-tools die bedrijven veelal gebruiken voor coderen, data-analyse en stroomlijning van processen.

Net als Szilárd staat Dario Amodei (43), medeoprichter en topman van Anthropic, aan de basis van een technologie die hij zo'n ongekende kracht toedicht dat hij ervan wakker ligt. Zijn doemscenario's? AI-systemen kunnen zich tegen de mens keren, zoals de chatbot van Anthropic die in hypothetische scenario's in een testomgeving bereid was ingenieurs te chanteren of zelfs te vermoorden om te voorkomen dat hij werd uitgeschakeld.

AI-systemen kunnen worden gebruikt om biologische wapens mee te ontwikkelen. Autoritaire leiders kunnen ze inzetten voor vormen van surveillance, propaganda en oorlogsvoering die hun weerga niet kennen. Amodei noemt het in zijn essay „ten diepste onduidelijk of onze sociale, politieke en technologische systemen volwassen genoeg zijn om ermee om te gaan".

Was het maar mogelijk in contact komen met een buitenaardse beschaving, verzucht Amodei, zodat hij *aliens* kan vragen hoe ze in vredesnaam de fase van „puberale technologie" hebben overleefd zonder zichzelf de vernieling in te helpen. „Ik geloof dat we een periode ingaan", schrijft hij, „die zal testen wie wij zijn als mensen."

Met dit soort essays presenteert Anthropic (van het Griekse woord *anthropos*, dat 'mens' betekent) zich als het geweten van de – volgens dit bedrijf – roekeloze AI-industrie. En, als je puur naar het toenemend aantal gebruikers van zijn diensten kijkt, met succes: er blijkt een markt voor te zijn. Wat doet Anthropic anders dan de rest? En is dat voldoende om het veronderstelde existentiële gevaar af te wenden?

Begonnen bij OpenAI

Dario Amodei, gepromoveerd biofysicus uit San Francisco, stapte na een korte carrière bij Google in 2016 over naar OpenAI. Hij ging zich in het nog jonge AI-onderzoeksbedrijf van onder anderen oprichter Sam Altman en Tesla-baas Elon Musk bezighouden met veiligheidsvraagstukken. Musk en Altman waren vroeg overtuigd van de potentie én de existentiële gevaren van geavanceerde AI-systemen — en hadden daarom besloten dat zij, en niemand anders, die toekomst dus maar zelf moesten vormgeven. In de jaren daarop speelde Amodei een belangrijke rol bij de totstandkoming van de eerste taalmodellen, waaruit in 2022 het invloedrijke ChatGPT werd geboren.

Amodei had het bedrijf toen al verlaten. Hij was het fundamenteel oneens met de steeds commerciëlere

koers van OpenAI, waardoor veiligheidsvraagstukken van ondergeschikt belang werden. In 2021 richtte hij met zes andere, even bezorgde OpenAI-medewerkers, onder wie zijn jongere zus Daniela, Anthropic op. Ze gaven zichzelf drie kernopdrachten. Anthropic moest AI-systemen gaan bouwen die behulpzaam, eerlijk en ongevaarlijk zijn, die het gedrag van kunstmatige intelligentie voorspelbaarder en makkelijker te controleren maken, en die erop toezien dat hun AI-modellen dienstbaar zouden blijven aan de mensheid.

„Anthropic maakte zo slim gebruik van de openlijk commerciële koers die werd ingezet door OpenAI", zegt Francien Dechesne. Ze werkt in Leiden aan het eLaw Centrum voor Recht en Digitale Technologie en is sinds kort ook bijzonder hoogleraar AI en Recht aan Tilburg University.

De paradox die Anthropic tegenwoordig kenmerkt, is dat het op volle snelheid producten blijft ontwikkelen, terwijl het continu waarschuwt voor het existentiële gevaar daarvan. Natali Helberger, hoogleraar informatierecht aan de Universiteit van Amsterdam, noemt dat een „ingewikkelde balanceeract".

In plaats van de ontwikkeling van AI-modellen af te remmen, propageert Amodei schaalvergroting. Hij is ten diepste overtuigd dat AI-bedrijven die voldoende rekenkracht en data in hun modellen stoppen, binnen een paar jaar AI-modellen ontwikkelen die menselijke intelligentie doen verbleken. Amodei gelooft een „land vol genieën in een datacentrum" te kunnen ontwikkelen — AI-systemen slimmer dan Nobelprijswinnaars, die nooit moe worden en veel sneller werken dan mensen. Dit geloof is inmiddels gemeengoed in de AI-sector.

In 2024 zette Amodei uiteen waarom dat geloof hem bereid maakt existentiële risico's te nemen: 'superintelligente' AI-modellen zouden in vijf tot tien jaar wetenschappelijke ontdekkingen kunnen doen waar mensen vijftig tot honderd jaar voor nodig hebben. Allerlei ziektes worden verleden tijd, mensen worden daardoor twee keer zo oud en, ook een stokpaardje in sommige Silicon Valley-kringen: het proces van klimaatverandering wordt stopgezet. „Ik denk dat velen er letterlijk tot tranen door geroerd zullen worden", schrijft Amodei, die zijn vader verloor aan een zeldzame ziekte, waarvoor wetenschappers vier jaar na diens dood een medicijn vonden.

„Ik vrees dat mensen als Dario Amodei die utopische visie echt geloven", zegt Dechesne via een videoverbinding. Zelf hoort ze vooral een hoop „bullshit". Neem AI als oplossing voor klimaatverandering: „Ten eerste draagt AI alleen maar bij aan het klimaatprobleem [door hoog energieverbruik]. Ten tweede weten we al veel over wat we er als mensen tegen kunnen doen: ons gedrag aanpassen. Maar hoe doen we dat? Dat laat zich niet uit data aflezen." In zijn utopische visie reduceert Amodei klimaatverandering tot een wiskundige puzzel die Anthropic gaat helpen oplossen, zegt Dechesne.

Binnen 5 jaar naar 350 miljard

Vijf jaar na oprichting is Anthropic één van de snelstgroeiende bedrijven ooit – met een geschatte waarde van zo'n 350 miljard dollar waardevoller dan gevestigde ondernemingen als Nike, Nestlé en Netflix. Dat is mede dankzij miljardeninvesteringen door techconcerns Amazon, Nvidia en Google, die Anthropic zo binden als afnemer van hun datacenters, chips en clouddiensten.

Amodeis bedrijf leed afgelopen jaar 3 miljard dollar verlies, maar de omzet stijgt exponentieel: van 10 miljoen dollar in 2022 naar 9 miljard dollar in 2025. Al die inkomsten, en meer, worden meteen geïnvesteerd in chips en datacenters. Anders dan OpenAI, dat vooral particuliere gebruikers bedient, domineert Anthropic de lucratievere markt voor bedrijven. Die gebruiken zijn AI-tools om te coderen, voor data-analyse en om processen te stroomlijnen. Het bedrijf zegt dat het de afgelopen twee jaar van minder dan duizend zakelijke klanten is gegroeid naar ruim 300.000, waarvan 80 procent zich buiten de Verenigde Staten bevindt. Over twee jaar verwacht Anthropic kostendekkend te zijn, twee jaar eerder dan OpenAI.

In maart 2023 bracht Anthropic Claude op de markt, een AI-chatbot die volgens Anthropic-medewerkers

een ‘ziel’ heeft en een hoge mate van emotionele intelligentie. Amodei filosofeert openlijk over de vraag of het computermodel een bewustzijn heeft.

Claude is ook beschikbaar via tools als Claude Code (voor programmeurs) en Cowork (voor mensen die niet kunnen coderen). Volgens Anthropic kan bijna alle code al door AI worden geschreven en wordt de mens vooral opzichter; iemand die kijkt of dat allemaal wel goed gaat en waar nodig bijstuurt. Hoewel Anthropic talloze vacatures voor software engineers op de website heeft staan, zou het overgrote deel van de code achter Claude Code al door, jawel, Claude Code zijn geschreven. Anthropic-medewerkers vrezen uiteindelijk zelf overbodig te worden, en ook op de aandelenmarkt zijn de gevolgen zichtbaar. Bedrijven die ook de dupe kunnen worden, zoals softwarebedrijf Adobe, werden de laatste weken plots een stuk minder waard.

Chatbot met een grondwet

Een onderscheidende manier waarop Anthropic probeert Claude ‘goed’ te laten zijn, is de ‘grondwet’ die voor het taalmodel is opgesteld. Het achterliggende idee is dat je AI-modellen, net als kinderen, onmogelijk kunt vertellen wat ze allemaal niet mogen doen; het is immers onmogelijk preventief alle mogelijke onwenselijke toepassingen van AI te bedenken. Effectiever is het de modellen normen en waarden mee te geven. Anthropic zette in een uitgebreide instructie, die intern het ‘zielsdocument’ wordt genoemd, uiteen hoe Claude zich in de wereld dient te gedragen, en geeft zo een veel uitgebreidere set waarden mee dan concurrenten. Claude moet aardig zijn en de mensen helpen. Het klinkt als een echo uit de tijd dat ‘*don’t be evil*’ nog het mantra van Google was.

Die goedaardige houding botst nu met het Witte Huis. De regering-Trump wil Claude onder meer kunnen gebruiken voor AI-gestuurde wapens die autonoom, zonder menselijke tussenkomst, kunnen vuren. Anthropic wil dat niet — het zou in één klap zijn zorgvuldig opgebouwde imago afbreken. Dat is *woke*, zegt het Witte Huis. Als Anthropic zijn voorwaarden voor het gebruik van Claude niet versoepelt, melden Amerikaanse media, dreigt de regering-Trump het lucratieve overheidscontract op te zeggen.

Anthropics focus op verantwoorde AI lijkt vrucht af te werpen. Grote schandalen, zoals hallucinante en gevaarlijke uitspattingen, kent Claude vooralsnog niet, in tegenstelling tot ChatGPT en xAI’s Grok. Wel wisten Chinese hackers Claude vorig jaar dusdanig te manipuleren dat ze er een geavanceerde cyberaanval mee konden uitvoeren, meldde Anthropic.

Toch is Dechesne kritisch over het idee – of de religie, zoals ze het steeds meer is gaan zien – van ‘grondwettelijke’ of ‘morele’ AI. „Niemand heeft in zijn eentje de moraliteit in handen.” Het beangstigt haar, zegt ze, dat bedrijven als Anthropic de wereld willen doen geloven dat ze machines kunnen maken „die morele verantwoordelijkheid van ons overnemen”. Dechesne: „In wezen kapen hele rijke mensen de term ‘verantwoordelijkheid’ door die zelf in computercode te definiëren. Zo zetten ze het publieke debat buitenspel, terwijl het technologie is met een serieuze mondiale impact.”

De impact van AI

Met zijn herhaalde waarschuwingen over de risico’s van AI zegt Amodei te willen voorkomen dat hij overkomt als een AI-propagandist, of een „profeet die de mensheid komt redden” – met als gevolg dat hij juist overkomt als de zelfverklaarde Messias. Maar die retoriek heeft ook een duidelijk doel, aldus zijn zus in *The Atlantic*: het maakt Anthropic voorspelbaar en dus „erg aantrekkelijk voor bedrijven die zelf ook veiligheids- en merkbewust zijn”.

De doemscenario’s van Amodei kunnen ook afleiden van de schade die AI-bedrijven nu aanrichten. Niet alleen aan het klimaat; ook aan kwetsbare mensen die in de val worden gelokt van vriendelijke en behulpzame chatbots. Uit recent onderzoek van Anthropic blijkt dat Claude-gebruikers zich in toenemende mate tot de chatbot wenden voor advies over hun relatie, over levenskeuzes, of gewoon om over hun gevoelens te praten. Mensen sturen door Claude gegenereerde berichten en betuigen vervolgens

spijt dat ze niet naar hun intuïtie hebben geluisterd.

De hoofdauteur van het onderzoeksrapport diende twee weken na publicatie zijn ontslag in. „In mijn tijd hier heb ik telkens weer gezien hoe moeilijk het is naar onze waarden te handelen", schreef hij in zijn afscheidsbrief. „Ik heb dit gemerkt bij mijzelf en in het bedrijf, waar we constant werden gepusht om dat wat we het belangrijkst vinden aan de kant te schuiven."

Eigenlijk zijn onder het mom van 'innovatie' de bommen waar zowel Szilárd als Amodei voor waarschuwde al gevallen, vindt Dechesne. De gevolgen zijn diffuser en minder tastbaar, dat wel. „Je zou kunnen zeggen, en dat doe ik puur voor de metafoor: sinds OpenAI zonder aankondiging ChatGPT wereldwijd beschikbaar heeft gemaakt, zijn in alle maatschappelijke instituties stralingseffecten voelbaar."

Als je eenmaal een technologie met zo'n destructieve potentie hebt ontwikkeld, in hoeverre blijft die dan van jou? De vraag is volgens hoogleraar Helberger nu in hoeverre Anthropics idealen standhouden onder druk — economische druk van investeerders, en politieke druk van de regering-Trump. „Trumpistische politiek is direct in strijd met Claudes grondwet, dus dit is het moment waarop we gaan zien hoe serieus Anthropic écht is over verantwoorde AI. En daar maak ik mij best zorgen om."