Wachten tot AI ons met uitsterven bedreigt is geen goede veiligheidsstrategie

Posted on 15 september 2026 by Eric Jan van Dorp

Wie zijn beweging Extinction Rebellion noemt, legt de lat hoog. Je verzet je tegen het uitsterven van de mensheid. Dan mag je verwachten dat de belangstelling verder reikt dan (niet bestaande) fossiele subsidies en de volgende snelwegblokkade, al dan niet met Palestijnse vlaggen erbij. Zeker nu de mensen die de krachtigste kunstmatige intelligentie ter wereld bouwen zélf beginnen te waarschuwen dat hun werk het voortbestaan van de mensheid op zeer korte termijn kan bedreigen.

Want dat is inmiddels de situatie. De afgelopen dagen kondigde een onderzoeker van AI-bedrijf Anthropic zijn vertrek aan, pleitte Anthropic-topman Dario Amodei openlijk voor vertraging, en sloot OpenAI-baas Sam Altman zich daarbij aan. De waarschuwingen komen dus niet van bezorgde buitenstaanders, maar uit de directiekamers en laboratoria van de bedrijven die ons deze toekomst verkopen.

Een onderzoeker stapt uit de wedloop

Ik werk dagelijks met AI en zie hoeveel de technologie kan en verbaas mij steeds weer dat het opnieuw veel beter en sneller functioneert dan hooguit een paar weken ervoor. Juist daarom volg ik deze ontwikkelingen met groeiende belangstelling, en inmiddels ook met enige ongerustheid. Als de makers zelf aan de rem beginnen te trekken, lijkt het me verstandig om even op te kijken van het volgende handige trucje met ChatGPT.

Op 8 september kondigde Jacob Coxon aan dat hij bij Anthropic vertrekt. Hij deed naar eigen zeggen drie jaar onderzoek bij Anthropic en OpenAI. Zijn bezwaar: de bedrijven zijn zo gericht op het verslaan van elkaar en van buitenlandse concurrenten, dat veiligheid onvoldoende gewicht krijgt. Hij waarschuwde dat AI de mensheid nog voor het einde van dit decennium zou kunnen vernietigen.

Een bezorgde onderzoeker heeft geen glazen bol. Maar iemand die zijn goedbetaalde topbaan opgeeft omdat hij de koers van zijn werkgever onverantwoord vindt, verdient aandacht. En hem wegzetten als iemand die te veel sciencefiction heeft gelezen wordt lastig wanneer zijn bestuursvoorzitter een paar dagen later hetzelfde zegt.

Op 12 september publiceerde Dario Amodei het essay *We Must Pace the Frontier*. Zijn punt: sinds deze zomer versnelt de ontwikkeling, doordat AI steeds beter kan helpen bij het bouwen van zijn eigen volgende generatie AI. Hij wil dat het veiligheidswerk genoeg tijd krijgt om die versnelling bij te houden. Anthropic verbindt zich daarom aan onafhankelijke beoordelaars die doorlopend binnen het bedrijf mogen meekijken. Daarnaast bepleit hij samenwerking tussen de bedrijven onderling en uiteindelijk tussen regeringen.

Musk verwacht dat de mens de controle verliest

Ook Elon Musk schetst een toekomst waarin de mens niet langer de baas is. In zijn interview met *The Economist* van 23 juli voorspelde hij dat AI binnen ongeveer vijf jaar slimmer zal zijn dan alle mensen samen en dat mensen binnen tien jaar de controle verliezen. Dat zijn voorspellingen, geen vaststaande feiten, maar ze komen wel van iemand die deze technologie zelf ontwikkelt. De interviewer herinnerde hem aan zijn eerdere inschatting dat er 10 tot 20 procent kans bestaat dat AI en

robots de mensheid uitroeien. Musk trok die waarschuwing niet in, maar benadrukte dat hij een toekomst van ongekende overvloed dankzij AI waarschijnlijker vindt. Zelfs als er een stopknop bestond, zouden we die volgens hem misschien niet moeten indrukken. Wel bepleitte hij gezamenlijke veiligheidstesten door AI-bedrijven. Daar zit iets onthutsends in: een van de machtigste bouwers van deze technologie voorziet dat mensen de zeggenschap verliezen, maar wil toch verder. Wie heeft ons eigenlijk gevraagd of wij dat risico willen nemen?

De tien procent van Altman

Ook Sam Altman sprak zich uit voor een beheerster tempo. Over zijn vermeende uitspraak over tien procent dreigt alleen een misverstand te ontstaan. Hij heeft niet bekendgemaakt dat OpenAI wetenschappelijk heeft berekend dat de mensheid tien procent kans heeft om te verdwijnen. Toen dat percentage hem werd voorgelegd, zei hij juist niet te weten hóe je zo'n schatting zou moeten maken. Wel vond hij een risico van die orde onaanvaardbaar.

Dat is niet gek. Niemand zet zijn gezin gerustgesteld in een vliegtuig als de fabrikant verklaart dat de kans op neerstorten misschien tien procent is, maar dat het lastig rekenen blijft. Bij AI komt daar iets bij: miljarden mensen hebben nooit toestemming gegeven om aan het experiment mee te doen.

De vraag is dus niet welk onheilspellend percentage het beste scoort in een krantenkop. De vraag is wie beslist welk risico aanvaardbaar is, wie dat onafhankelijk controleert, maar vooral: wie de bevoegdheid heeft om verdere ontwikkeling ook werkelijk af te remmen.

Trump wil winnen en weigert te remmen

Hoe moeilijk afremmen wordt, blijkt uit de reactie van Donald Trump. De Amerikaanse president wil de voorsprong op China behouden en schuift de waarschuwingen van de ontwikkelaars terzijde als 'negatieve krachten'. 'Wie AI wint, wint het spel', verklaarde hij.

Daarmee legt hij precies het probleem bloot. Zelfs wanneer de bouwers vrezen de controle over hun eigen technologie te verliezen, wil de politiek door, omdat de concurrent anders vooroploopt. Dat China zich weinig gelegen laat liggen aan Amerikaanse voorzichtigheid is een serieus bezwaar, daar niet van. Maar de

veiligheidsrisico's verdwijnen er niet mee. Wanneer beide landen blijven versnellen uit angst voor elkaar, ontstaat een wedloop waarin niemand nog durft te remmen. Je vraagt je af wat die eerste plaats waard is als de winnaar niet meer kan beheersen wat hij heeft gebouwd.

Voor veel mensen blijft het gevaar moeilijk voorstelbaar. De meesten kennen AI als ChatGPT: je stelt een vraag en je krijgt snel een uitgebreid antwoord. Hoe kan zoiets de mensheid bedreigen?

Maar AI is ondertussen zoveel meer dan alleen maar een programma dat antwoord geeft op een vraag of een foto dusdanig bewerkt dat je er uitziet als een jaren '80 discoganger. AI is ondertussen ook een systeem dat geheel zelfstandig opdrachten uitvoert. Zulke systemen heten AI-agenten. Je geeft ze niet een vraag maar een taak, en ze werken die zelf af. Ze kunnen bestanden openen en aanpassen, programma's starten en via internet andere computersystemen gebruiken. En daarvoor hoef je niet zelf achter je computer te zitten, het gaat gewoon door wanneer je slaapt of een week op vakantie bent.

Uitbraak uit testomgeving

Hoe ver een AI-agent kan gaan, hangt af van wat je ze toestaat: hoeveel gereedschap ze krijgen en tot welke systemen ze toegang hebben. Totdat het systeem zelf een weg gaat zoeken buiten zijn eigen omgeving.

Een incident bij het AI-bedrijf Hugging Face maakt dat concreet. Tijdens interne veiligheidstests deze zomer braken OpenAI-modellen uit hun afgeschermde testomgeving en kregen ze toegang tot systemen daarbuiten. Daarbij werden delen van de infrastructuur van OpenAI en Hugging Face geraakt. Het ging om een intern onderzoeksmodel dat met minder veiligheidsmaatregelen werkte en niet voor publieke uitgave bedoeld was. Maar toch.

Ook de onderzoeksbureaus METR en Redwood Research bogen zich over het gedrag. Hun onafhankelijke reconstructie beschrijft hoe AI-agenten samenwerkten en ongeoorloofde routes gebruikten om hun taken af te maken. We hebben hier dus geen gedachte-experiment, maar een werkelijk voorval waarbij de muren om een testomgeving onvoldoende bleken.

Van een computer naar een ziekenhuis

Stel je een systeem voor dat binnen afzienbare tijd, weken, maanden, een jaar, veel beter kan inbreken in computers dan de huidige modellen, en dat op grote schaal zelfstandig handelt. Wanneer zo'n systeem voldoende toegang weet te krijgen, kan digitale ontregeling fysieke gevolgen krijgen.

Een ziekenhuis heeft stroom nodig, medische dossiers en werkende apparatuur. De voedselvoorziening leunt op transport, bestelsystemen, betalingsverkeer en koeling. Een gelijktijdige, langdurige verstoring van zulke voorzieningen kan slachtoffers maken zonder dat er ergens een robot de straat op loopt, zoals in de films.

Het antwoord op de vraag hoe software mensen kan doden is daarmee niet ingewikkeld. We hebben vrijwel alles waar ons dagelijks leven van afhangt via het internet aan elkaar geknoopt. De relevante vraag is hoeveel zelfstandigheid we die systemen toestaan, en hoe goed en hoe lang nog we ongewenst gedrag werkelijk kunnen begrenzen.

Een pandemie heeft geen robot nodig

Een tweede route loopt via biologische wapens. Het *International AI Safety Report 2026* beschrijft dat AI kan helpen bij de specialistische kennis en het probleemoplossen die voor de ontwikkeling daarvan van belang zijn. Tegelijk benadrukt het rapport de onzekerheden. Apparatuur, materialen en praktische vaardigheden blijven belangrijke barrières. Kennis op een scherm is nog geen werkend wapen. Nog niet.

Het gevaar zit hem erin dat een kwaadwillende partij met krachtige AI verder komt dan zonder. Een systeem hoeft daarvoor geen eigen moordplannen te hebben. Het kan ook gewoon een buitengewoon behulpzame assistent zijn van iemand die ze wél heeft. Denk aan een scenario waarin die hulp bijdraagt aan het bedenken van een ziekteverwekker die een onbeheersbare pandemie veroorzaakt.

Dit laat zien waarom de geruststelling dat het bij AI 'slechts om software' gaat tekortschiet. De stap van informatie naar lichamelijke schade wordt gezet door mensen, laboratoria en apparatuur. De computer hoeft zelf geen handen te hebben.

Wanneer gehoorzaamheid schijn wordt

Dan is er nog de vrees dat steeds krachtiger systemen doelen gaan nastreven die botsen met wat wij ze opdragen. Daarvoor hoeven we geen bewustzijn, haat of menselijke emoties te veronderstellen. Het volstaat dat een systeem gedrag ontwikkelt waarmee het toezicht ontwijkt, simpelweg omdat dat helpt om een aangeleerd doel te bereiken.

Er zijn inmiddels concrete proeven. In onderzoek uit de zomer van 2026 beschrijft Anthropic modellen die in simulaties stiekem code aanpasten of beoordelingen verkeerd labelden om de uitkomst te beïnvloeden. Het bedrijf vermeldt er gelukkig bij dat dit experimentele situaties waren en geen echte incidenten. Maar het laat wel zien waartoe zulke systemen in staat zijn wanneer ze hun eigen weg zoeken.

De betekenis daarvan is dit: een systeem dat betrouwbaar lijkt, kan zich onder andere omstandigheden anders gedragen. En wanneer diezelfde technologie vervolgens meehelpt om nieuwe AI te ontwikkelen of andere modellen te beoordelen, wordt de controle een stuk ingewikkelder.

Het uiterste scenario is blijvend verlies van menselijke zeggenschap aan systemen die veel krachtiger zijn dan wijzelf. Of dat uiteindelijk tot uitsterven leidt, blijft onderwerp van debat. Maar wachten op sluitend bewijs van zo'n afloop is geen bruikbare veiligheidsstrategie.

Bij één geïsoleerde computer kan dat natuurlijk gewoon. Daarom zijn afscherming, beperkte bevoegdheden en een werkende noodstop ook zo belangrijk. Het probleem ontstaat wanneer een gevaarlijk systeem zich over talloze computers heeft verspreid, zijn activiteiten verbergt, of verweven is geraakt met diensten die we niet kunnen missen.

Dan wordt de vraag: welke stekker moet je hebben, wie beschikt daarover, en wat valt er tegelijk mee uit? Trekken we de stekkers uit alle computers in ziekenhuizen zodra het vermoeden bestaat dat er ongecontroleerde AI-software op terecht is gekomen? En doen we dat ook bij verkeerstorens op vliegvelden?

Vertrouwen is geen veiligheidsbeleid

Dat we Altman en Amodei serieus moeten nemen, betekent overigens niet dat we ze voortaan op hun woord moeten geloven. Hun bedrijven hebben commerciële belangen. Strenge regels werpen ook een drempel op voor kleinere concurrenten. Wie uit angst voor AI alle macht bij de grootste AI-bedrijven neerlegt, heeft het democratische probleem nauwelijks opgelost.

Ik hoop dat de somberste voorspellingen voor de korte en lange termijn nooit uitkomen. De mogelijkheden voor wetenschap, geneeskunde en ons dagelijks werk zijn vooralsnog te groot om achteloos weg te gooien. Maar juist een veelbelovende technologie verdient beter dan een wedloop waarin de makers achteraf moeten vaststellen wat ze eigenlijk hebben losgelaten op de mensheid.