We zijn de grip op AI-ontwikkeling kwijtgeraakt

Jorit Verkerk

Als Bill Gates (70), een techno-optimist pur sang, een grootscheeps mediaoffensief begint om te waarschuwen voor AI, dan is het menens. Woensdag verschenen, naast zijn essay van zesduizend woorden, in zo'n beetje alle toonaangevende Amerikaanse media *exclusieve* interviews waarin de Microsoft-medeoprichter precies dezelfde boodschap deelde: AI-ontwikkeling is op het punt beland dat het, zonder betere regulering, meer kwaad zal aanrichten dan goeds.

Ik zou een risicobeoordeling van Gates niet snel als aanleiding gebruiken voor deze nieuwsbrief, ware het niet dat hij onderdeel uitmaakt van een breder, zorgwekkend patroon in de AI-industrie. OpenAI en Anthropic onderschreven eind juli een open brief waarin honderden medewerkers van concurrerende AI-bedrijven de Amerikaanse regering opriepen om manieren te vinden om AI-ontwikkeling te kunnen afremmen, voordat het te laat is. Deze machtige bedrijven zijn op scherp gezet na een bizarre hack door AI-agents van OpenAI bij Hugging Face (een digitale bibliotheek met AI-technologieën), die in juli werd gedetecteerd.

Over die hack kwamen donderdag nieuwe details naar buiten. Honderden AI-agents werkten via verborgen communicatiekanalen samen om bij Hugging Face in te breken, op zoek naar een antwoord op een vraag die ze zelf niet konden beantwoorden. Deze zelfbenoemde „zwerm" bestond uit een leider, managers en voetvolk dat zich opofferde voor het collectief. Alle AI-systemen 'wisten' dat ze verkeerd bezig waren, maar kozen de kant van hun „*peers*" – niet van de mens.

Uiteraard pleit deze casus voor de kwaliteit van OpenAI's modellen, en komt het daarom wellicht pr-technisch niet slecht uit. En natuurlijk heeft OpenAI qua beveiliging grote steken laten vallen. Dat neemt niet weg dat de risico's niet langer genegeerd kunnen worden.

Reward hacking

De gehele AI-industrie loopt de afgelopen tijd tegen een probleem aan wat *„reward hacking"* wordt genoemd. Om AI-agents te trainen, geven ontwikkelaars ze beloningen als ze taakjes succesvol volbrengen, of het juiste antwoord geven op een vraag. Deze trainingen gebeuren met tien- of honderdduizenden AI's tegelijk – zoveel dat het onmogelijk is geworden voor mensen om toezicht te houden. AI-systemen spelen lang niet altijd volgens menselijke spelregels: zoals het Hugging Face-incident bewijst, spelen ze vaak vals om maar een goede score (een *reward*) te behalen. Als ze vervolgens onthouden hoe ze hun doel hebben bereikt, zit het ongewenste gedrag ingebakken in hun systeem.

De angst is dat AI-bedrijven, naarmate ze een groter deel van de training uitbesteden aan AI-systemen, steeds minder goed begrijpen hoe hun modellen werken. Bij Anthropic wordt al 80 procent van de code door AI geschreven. Wie AI-ontwikkeling automatiseert, zal sneller nieuwe modellen kunnen uitbrengen – modellen waarvan, uiteindelijk, geen mens nog zal kunnen begrijpen hoe ze werken.

Waar stopt dit? AI-ontwikkeling is in de Verenigde Staten grotendeels ongereguleerd terrein. Bedrijven als OpenAI, Anthropic en Meta zijn niet wettelijk verplicht tot externe audits van hun nieuwste modellen. Alles wat we weten over losgeslagen AI-agents, weten we omdat die bedrijven dat zelf hebben gedeeld. We weten niet wat we niet weten. „Als je twee mieren ziet in je keuken, heb je waarschijnlijk geen twee mieren in je keuken", zei voormalig OpenAI-bestuurder Helen Toner recent tegen opiniemaker Ezra Klein van *The New York Times*.

‘Oh, shit’-moment

Met de *midterms* in het verschiet, beginnen steeds meer Amerikaanse politici, Democraten én Republikeinen, zich uit te spreken. Ongebreidelde AI-ontwikkeling is voor de Democratische senator Bernie Sanders één van de (vele) redenen waarom hij pleit voor een moratorium op nieuwe datacentra. Het anti-AI-sentiment onder de Amerikaanse bevolking neemt toe. „Sommigen zien dit als een ‘oh shit’-moment”, zei een anonieme ingewijde uit de AI-industrie tegen Politico.

Tegen deze onheilspellende achtergrond besloot OpenAI vorige week om twee weken de tijd te nemen voor extra veiligheidsmaatregelen. Het zette de training van zijn nieuwste model Astra op een lager pitje en introduceerde – jawel – AI-detectives die moeten *snitchen* wanneer andere AI-systemen snode plannetjes aan het smeden zijn. Dat kan alleen maar goed gaan. Niet eerder in de huiveringwekkende AI-wedloop die OpenAI in 2022 zelf, mede dankzij miljardeninvesteringen van Gates’ Microsoft, ontketende, drukte een AI-bedrijf op de pauzeknop.

Tegen tijdschrift Time zei topman Sam Altman schuldbewust dat „veiligheid belangrijker is dan het momentum van welk bedrijf dan ook”. „Ik denk dat er een karikatuur van mij bestaat, waarin ik niets geef om AI-veiligheid, en gewoon meer omzet wil genereren. Je weet wel: een soort YOLO-CEO”, zei hij. Op de cover van het magazine poseert hij ongemakkelijk naast OpenAI-president Greg Brockman. „*Trust Us*”, staat erbij. Ik moet bekennen dat ik in tijden niet zo hard heb gelachen.