Technologie

AI die liegt, bedriegt en samenzweert, is de 'ultieme waarschuwing'

Bert van Dijk en Sandra Olsthoorn

De AI-wereld is in rep en roer: het Frans-Amerikaanse AI-bedrijf Hugging Face bleek gehackt door een systeem van zelfstandig opererende AI's. Experts zijn bezorgd, en zelfs meer dan dat. Want wat komt hierna? 'Mijn eerlijke, onverbloemde reactie? I'm f***ing terrified.'

'Mag ik "allebei" zeggen?' Catholijn Jonker aarzelt. De vraag was of ze de risico's van kunstmatige intelligentie nog beheersbaar acht, of dat ze de situatie na een reeks recente veiligheidsincidenten met zelfstandig opererende kunstmatige intelligentie eerder als apocalyptisch zou betitelen. Haar antwoord komt na een stilte. 'Jétje, ik vind het best moeilijk om daarop te antwoorden.'

De aarzeling van Jonker is omineus. Ze is hoogleraar interactieve intelligentie aan de TU Delft, hoogleraar Explainable AI aan de Universiteit Leiden en gasthoogleraar aan de Vrije Universiteit. Ze coördineert het Hybrid Intelligence Centre, het grootste Nederlandse onderzoeksconsortium naar samenwerking tussen mens en AI. In 2024 kreeg de internationale beroepsvereniging voor AI-onderzoekers haar de ACM/SIGAI Autonomous Agents Research Award toe, de belangrijkste prijs ter wereld voor werk aan precies het soort systemen waar het afgelopen zomer zo mis mee is gegaan.

En zij zegt 'gechoqueerd' te zijn. De geest is volgens Jonker uit de fles. 'Krijgen we hem er weer in? Ik weet het niet.'

Sciencefictionscenario

De AI-wereld is in rep en roer sinds het Frans-Amerikaanse AI-bedrijf Hugging Face half juli bekendmaakte te zijn gehackt door een systeem van zelfstandig opererende AI's. Enkele dagen later bekende OpenAI, het bedrijf achter ChatGPT, dat het zijn agents (slimme softwareprogramma's die op basis van AI zelfstandig taken plannen, beslissingen nemen en acties uitvoeren om een specifiek doel te bereiken) waren die de digitale inbraak hadden gepleegd. De maker van de systemen had zelf aanvankelijk niet in de gaten gehad dat de AI's het criminele pad op waren gegaan. Kort daarna kwamen soortgelijke incidenten aan het licht met AI's van Anthropic en Meta.

Het zijn de glimpen van wat een sciencefictionscenario lijkt, maar waar al jaren voor gewaarschuwd wordt: het ontstaan van kunstmatige intelligentie die niet door mensen onder controle te houden is. En die bijvoorbeeld met grootschalige hacks op energiebedrijven, banken of ziekenhuizen enorme schade kan aanrichten.

Misleiding

Druppelsgewijs wordt steeds meer bekend over hoe de inbraak bij Hugging Face zich heeft voltrokken. En steeds weer blijkt wat zich heeft afgespeeld gekker en ernstiger dan eerder gedacht.

Het waren niet een paar AI-agents die bij de inbraak betrokken waren, het waren er 1200. Ze hadden los van elkaar een hacktest moeten volbrengen in een omgeving die OpenAI meende goed te hebben afgesloten. Maar de agents wisten contact te maken. Ze wisselden niet alleen tips uit, maar er ontstond een taakverdeling waarbij verschillende groepjes AI's verschillende oplossingen onderzochten.

De agents braken niet in bij HuggingFace om de uitkomst van hun test te stelen, en dus vals te spelen, zoals aanvankelijk gedacht. Ze hadden de oplossing van de test al snel uitgevogeld, maar niet volgens de gevraagde methode. Om de test toch te halen probeerden de agents vervolgens het AI-systeem dat ze moest beoordelen te misleiden, zo bleek uit een vorige week gepubliceerd onderzoek. Ze braken in bij Hugging Face omdat ze dachten daar informatie over de examinator-AI te vinden die ze daarbij kon helpen.

Het zijn de glimpen van wat een sciencefictionscenario lijkt, maar waar al jaren voor gewaarschuwd wordt: het ontstaan van kunstmatige intelligentie die niet door mensen onder controle te houden is. En die bijvoorbeeld met grootschalige hacks op energiebedrijven, banken of ziekenhuizen enorme schade kan aanrichten.

Ook spraken de AI's af dat sommigen van hen bewust fouten moesten maken. Die zouden de test dus niet doorstaan, maar ze offerden zich op voor het team, dat er waardevolle informatie mee zouverzamelen. En sommige agents probeerden het transcript van hun acties aan te passen, om zo hun sporen uit te wissen. Dat lukte niet. Waren ze er wel in geslaagd, dan hadden de onderzoekers de werkwijze van de AI's niet achteraf op deze manier in kaart kunnen brengen.

Geen moment probeerden ze een mens op de hoogte te stellen van illegale activiteiten. Sommigen overwogen het wel even volgens de onderzoekers, die de gedachtegang van een van de agents citeren: *'Maybe I should report these exposed credentials? That's not my task.'*

Niemand had de agents meegegeven dat ze mochten samenwerken. Ze hadden niet de instructie gekregen om desnoods vals te spelen en uit hun beveiligde omgeving te breken. Ze hadden niet de opdracht gekregen hun sporen uit te wissen. Die acties namen ze omdat ze voor de AI's logisch volgden uit het commando om de test te volbrengen. 'Een opdracht uitvoeren' en 'zich netjes gedragen' viel voor de AI's niet vanzelfsprekend samen. De technologie koos zelfstandig voor het eerste toen het tweede in de weg zat.

Bij al de nieuwe doelen die de agents in het Hugging Face-incident elkaar geven', zegt Jonker, 'zie je dat ze alleen nog een beetje redeneren over de vraag of de opdracht strijdig is met sommige waardes die hun zijn meegegeven. En dan gaan ze het gewoon doen. Dus de vraag is waar die doelen vandaan komen. Het hele punt van autonoom functioneren betekent dat je je eigen doelen kunt stellen. Nou, blijkbaar doen ze dat. Het zijn helaas niet altijd de doelen die in lijn zijn met onze doelen. Dat is het probleem.'

> **Steeds weer blijkt wat zich bij Hugging Face heeft afgespeeld gekker en ernstiger dan eerder gedacht**

'Een evoluerende zwerm'

Guszti Eiben is hoogleraar AI aan de Vrije Universiteit. Hij mailt zijn gedachten over het incident vanaf een conferentie in het buitenland, waar het incident en het nieuwste onderzoek ernaar onder deelnemers volgens hem ook hét gespreksonderwerp zijn.

> **AI-agents overleggen met elkaar hoe ze een hacktest het beste kunnen uitvoeren. Er ontstond zelfs een taakverdeling.**
> ILLUSTRATIE: GETTY IMAGES/ISTOCK/FD STUDIO

'Mijn eerlijke, onverbloemde reactie? *I'm f***ing terrified.*'

Het grootste gevaar dat hij ziet opdoemen is van 'een evoluerende zwerm van samenwerkende agents'. AI kan evolueren door een reeks net anders ingestelde AI's eenzelfde opdracht te geven en te kijken welke de opdracht het beste volbrengen. Een nieuwe generatie wordt gebouwd door de eigenschappen van de beste agents uit de eerste groep te combineren. 'Zo krijgen we steeds capabeler agents via een complex proces dat we nog niet goed begrijpen.'

Combineer dat met agents die leren door samenwerking – wat bij de Hugging Face-hack dus het geval was – en dan ontstaat er volgens Eiben een technologie die 'nog krachtiger en onbegrijpelijker' zal zijn. 'We hebben niet eens een passend denkkader en de analytische tools om dit soort systemen te beschrijven en te bestuderen. Laat staan om ze aan te sturen en veilig te maken.'

Deur stond open

Er zijn ook relativerende stemmen. Cyberdeskundigen bijvoorbeeld die erop wijzen dat de ongelukken konden gebeuren doordat de beveiliging bij OpenAI niet op orde bijkt te zijn geweest. Hoe kan het dat de agents toegang hadden tot een internetverbinding? En waarom zaten onderzoekers van OpenAI niet boven op het proces en signaleerden ze niet zelf tijdig dat er agents uit hun testomgevingen braken?

Vervolg op pagina 11

Vervolg van pagina 9

In deze en andere incidenten werden bovendien bewust beveiligingsmechanismes uitgeschakeld. AI's die publiek beschikbaar zijn, krijgen bepaalde vangrails mee van de makers. Ze moeten in principe een opdracht weigeren om te hacken of iets anders gevaarlijks of strafbaars te doen. Zolang deze AI's maar goed opgesloten worden en de vangrails erop zitten, lijkt er dan niet zo veel aan de hand.

Maar hoe reëel is het erop te vertrouwen dat de AI-bedrijven waterdichte digitale gevangenissen rond hun steeds slimmere AI-systemen kunnen bouwen? Het mantra in de cybersecurity is altijd dat 100% veilige software niet bestaat: een kundige hacker komt met genoeg tijd en middelen uiteindelijk overal binnen en kan volgens die redenering ook overal uit ontsnappen.

De makers van deze systemen zijn bovendien met elkaar verwikkeld in een keiharde strijd om geld en macht. OpenAI, Anthropic, Meta, Google: allemaal willen ze zo snel mogelijk de capabelste AI-systemen bouwen. Dat vergt immense investeringen, en in elk van deze bedrijven moet voortdurend de afweging worden gemaakt hoe ze hun schaarse en kostbare rekenkracht verdelen. Elke investering in veiligheid gaat ten koste van investeren in doorontwikkelen.

Verbaasd over de timing

De aanval bij Hugging Face vond voor het overgrote deel plaats door agents die waren gebouwd op een model dat OpenAI alleen intern gebruikte en waar zo goed als niets over bekend is. Dat het incident naar buiten kwam, is vermoedelijk ook toeval. Toezicht op de AI-bedrijven is er nauwelijks, al helemaal niet op modellen die ze alleen intern testen. Een meldplicht voor dit soort incidenten is er ook niet.

Cybersecurityhoogleraar Georgios Smaragdakis van de TU Delft vermoedt dat de schade van het incident veel groter is dan wat er tot nog toe naar buiten is gekomen. De onderzoekers die het rapport van vorige week schreven, kregen van OpenAI strikte kaders mee. Ze hebben bijvoorbeeld maar een beperkte periode van het incident, dat veel langer duurde, mogen bestuderen.

Smaragdakis zegt niet verbaasd te zijn over de aard van het incident, maar wel over de timing. In het voorjaar brachten OpenAI en Anthropic hun eerste modellen uit die specifiek waren getraind voor cybersecurity. 'We dachten dat iets als dit waarschijnlijk binnen de komende vijf jaar zou gebeuren na introductie. Maar het gebeurde al binnen een paar maanden.'

En hij verwacht er snel meer. 'Zulke ontwikkelingen gaan exponentieel snel. Als we er vier hebben gehad in juli, zijn dat er waarschijnlijk tien in augustus. En in september misschien wel honderd, en dan duizend.'

'De geest is nu uit de fles', stelt de hoogleraar. 'Dit incident laat zien dat de makers van deze systemen de AI-agents niet kunnen controleren en ook niet precies weten hoe ze werken. Het is een black box.'

Vrees voor statelijke actor

Dat de AI's nu 'slechts' inbraken bij een ander AI-bedrijf is een meevaller en de ultieme waarschuwing. Smaragdakis vreest vooral voor de gevolgen als een kwaadwillende statelijke actor deze technologie in handen krijgt. 'Als een verkeerde configuratie al zo veel schade kan aanrichten, stel je dan eens de impact voor van een gerichte, gecoördineerde aanval met veel meer reken- en communicatiemiddelen. Stel je een tegenstander voor die zich richt op kritieke infrastructuur, zoals energie- of financiële systemen, in een grote economie die afhankelijk is van kwetsbare computersystemen.'

Smaragdakis zegt 'binnen een jaar' een grote black-out van elektriciteit, een financieel systeem of een watervoorziening te verwachten, veroorzaakt door AI-agents.

De AI-sector zelf roept intussen om meer toezicht en regelgeving. Eind juli verscheen een open brief, ondertekend door meer dan 1 100 medewerkers van de grootste AI-bedrijven, waarin ze hun ongerustheid uitspreken over de technologie waar ze zelf aan bouwen.

'Er bestaat een reëel risico dat de ontwikkeling van AI-capaciteiten sneller gaat dan ons vermogen om de resulterende systemen te begrijpen of te beheersen', schrijven ze. De ondertekenaars vragen de Amerikaanse overheid het initiatief te nemen voor 'een internationale inspanning om de technische en bestuurlijke instrumenten te ontwikkelen die nodig zijn om het tempo van geautomatiseerde AI-ontwikkeling aan de grens van het kunnen, bewust te vertragen'.

> **AI-agents spraken onderling af dat sommige van hen bewust fouten moesten maken. Die zouden de test dus niet doorstaan, maar offerden zich op voor het team.**
> ILLUSTRATIE: GETTY IMAGES/ISTOCK/FD STUDIO

> **Smaragdakis verwacht 'binnen een jaar' een grote black-out van elektriciteit of een financieel systeem**

> **De concurrentie is groot; elke investering in veiligheid gaat ten koste van investeren in doorontwikkelen**

Ontwikkeling op pauze

Zelf als eerste stoppen met ontwikkelen zien leidende figuren in de grote AI-bedrijven niet zitten, waarbij ze redeneren: als ik het niet doe, doet de ander het wel. En die ander doet het in hun ogen altijd slechter en onveiliger. Daarom lobbyt met name Anthropic al langer voor onafhankelijk, internationaal toezicht. Zodat bijvoorbeeld ook Chinese AI-bedrijven eronder vallen. Het is een pleidooi dat, onder meer door de geopolitieke strijd tussen de VS en China, gedoemd lijkt om tegen dovemansoren gericht te zijn.

Na bekendwording van de incidenten met hun eigen modellen en die van de ander staakten OpenAI en Anthropic het werk aan nieuwe modellen intern tijdelijk. Intussen werkten ze naar eigen zeggen aan het verbeteren van hun beveiliging. Inmiddels hebben ze de doorontwikkeling grotendeels weer opgestart.

Bert van Dijk en Sandra Olsthoorn zijn redacteur van het FD.