



Met het Oog op Morgen: Uitgebroken AI?

📅 Jul 23 2026

Gisteren 22 juli was ik in Met het Oog op Morgen, een programma waar ik al 40 jaar naar luister (ik was er heel vroeg bij), dus ik vond het heel bijzonder daar te zijn! En het was ook erg leuk. Goed voorbereid door de redactie sprak ik met Sheila Sitalsing over de recent “uitgebroken AI hack”. Ik sprak iets te veel door Sheila heen, waarvoor excuses. Maar, ik denk ook dat we een paar dingen hebben genoemd die je niet zo vaak hoort. Zoals dat je als bedrijf je niet kunt verschuilen achter wat jouw AI gedaan heeft. Daar ben je toch echt zelf verantwoordelijk voor.

Maar ook, zelfs als die AI nog geen bovenmenselijke dingen aan het doen is, hij heeft wel onwijs veel tijd om je te hacken. En hij probeert ieder gaatje in je beveiliging en komt dan meestal wel binnen. Waarmee ik bedoel, we moeten echt aan de bak. Iedereen met slaperige IT oplossingen waar geen team toplui opzit loopt nu een enorm risico.

Onderstaand een transcript van het segment, dat vanaf minuut 29:20 [ook hier te beluisteren is](#).

Transcript

Sheila Sitalsing: Een AI, kunstmatige intelligentie, die zijn eigen goddelijke gang gaat en de mens overvleugelt. Het is een doembeeld of een science fiction scenario. Maar nu is er een zogenoemd veiligheidsincident gebeurd bij OpenAI. Dat is de maker van onder meer ChatGPT. En volgens het bedrijf zelf sloeg zijn eigen AI-programma of een eigen AI-agent bij een test op hol. En toen ging het uit eigen beweging een ander bedrijf weer hacken. Opwindend van de techwereld.

[audiofragmenten van ander nieuws]

Tech-expert Bert Hubert, welkom.

Bert Hubert: Goedenavond.

S: Nou, hoop opwindend. OpenAI zou geen controle meer hebben gehad over zijn eigen AI-agent, zo'n programmaatje dat autonoom taken aan het verrichten is. Het wist te ontsnappen. Je hoorde het begrip sandbox, dat viel net al in het fragment. Een afgesloten deel waar het getest werd. Het sloeg aan het hacken. Elders op internet. Dat klinkt als

science fiction. Wat is hier nou gebeurd?

B: Nou, om te beginnen, de firma OpenAI en alle AI-boeren doen dit eigenlijk één keer per kwartaal. Dan komen ze met nieuws "Zo eng wat we nu gemaakt hebben. En dat is echt te gevaarlijk voor de wereld!" En dan zes maanden, zes weken later zeggen ze trouwens, je kunt het nu bij ons huren.

S: Dus het is ook een beetje reclame.

B: Dat is het minstens, deels komt het ze wel heel goed uit. En is het onderdeel van een patroon. Tegelijkertijd is er echt wel iets gebeurd. Ze hebben een computerprogramma gemaakt wat goed kan hacken. Daar hebben ze alle veiligheidsfeatures uitgezet. Dus ze hebben gezegd, de veiligheidsklep is er af. En toen hebben ze die aan het hacken gezet. En hebben hem ook nog toegang gegeven tot internet. En nu zeggen ze, ja, onze AI is zo knap. Hij is ontsnapt. Maar ja, het was hun eigen AI waar ze zelf de veiligheidsklep af hebben gehaald. En die moest gaan demonstreren hoe goed die kon hacken. Toen hebben ze niet opgelet. En toen bleek dat die machine niet goed begrepen had wat die moest hacken. En toen ging die maar het internet op om daar allemaal andere bedrijven te hacken.

S: Ja, want in dat fragment wat we net hoorden ging het ook over intentie. Dus expres. Alsof dat AI-model wist wat het deed. Of was het toch OpenAI zelf die wist dat het model dit kon gaan doen?

B: Ja, daar hebben ze eigenlijk iets heel raars gedaan. Ze hebben dus een computerprogramma gemaakt wat kan gaan hacken. En dat programma heeft vervolgens een ander bedrijf gehackt. Dat heet een misdrijf.

S: Ja, dat mag niet, toch?

B: Als ik dat had gedaan, zat ik nu op een politiebureau. Maar wat doen de AI-bedrijven nou? Die zeggen ervan, ja, nee, maar dat hebben wij eigenlijk niet gedaan. Dat heeft onze AI gedaan. Die heeft allemaal zelf eigen ideeën over de wereld.

S: Oh, dus het is ook een manier om de verantwoordelijkheid buiten het bedrijf te leggen.

B: Precies. Het is eigenlijk nog nooit gebeurd dat iemand heeft gezegd, ja, we hebben een ander bedrijf gehackt. Maar wauw, wat zijn we daar trots op.

S: Maar het was niet de bedoeling, neem ik aan, dat deze hack zou plaatsvinden. Het was de bedoeling om te testen wat hij kon, deze AI-agent. Maar het zal toch ook niet de bedoeling zijn?

B: Nee, het komt meer over als enorme slordigheid. Ze hadden bijvoorbeeld kunnen monitoren. Heeft dit apparaat inmiddels verbinding met internet. Dat hadden ze heel makkelijk kunnen zien. Want het is een AI, het is ook niet magisch. Het is gewoon een computerprogramma. En die begon op een gegeven moment met het internet te

communiceren. En op dat moment hadden ze hem uit moeten zetten. Maar ik vermoed eerlijk gezegd dat ze hem aan hebben gezet. En dat ze vervolgens zelf allemaal naar huis zijn gegaan. En ze hadden zoiets van, nou, laten we morgen maar kijken wat hij gedaan heeft.

S: Dus jij denkt eerder slordigheid dan dat we hier een enorm intelligent, zelfdenkend programma aan het werk zien.

B: Ja, en als onderbouwing daarvan. Die mensen die gehacked zijn die lieten paar dagen geleden al weten gehacked te zijn, maar ze wisten niet door wie. Die vertelden ook al van, ja, we zijn niet gehackt met hele bijzondere, unieke technologieën of zo. Ze zijn bijvoorbeeld gehackt via een gelekt wachtwoord. Nou, dat is helemaal niet heel bijzonder.

S: Dat is toch de meest voorkomende manier waarop we gehackt zijn? Nog steeds.

B: En ook de AI doet dat. Dus het komt meer over als van, ja, we hadden de deur een kier laten staan. En de machine niet goed uitgelegd wat hij wel en niet moet doen. Dan dat dit zo knap is ineens. Dat het allemaal onverwacht was.

S: Dus dit is niet een ontwikkeling, een nieuwe ontwikkeling die nieuwe inventiviteit, effectiviteit laat zien? Of toch wel een beetje? Ze hebben wel dat gaatje kunnen vinden.

B: Precies. Een van de dingen die de AI ontzettend goed doet, hij kan best wel oké hacken. Maar een van de meest spectaculaire dingen is dat de AI heeft gewoon honderden uren tijd per dag. Dus waar een normale hacker toch af en toe moet slapen, kan je de computer gewoon door laten gaan. Hij probeert ieder gaatje wat hij kan vinden. En uiteindelijk kan ontzettend veel daardoor alsnog gehackt worden. Zelfs zonder dat er een hele bijzondere techniek achter zit.

S: Ja. Nou, is dit, zeg je misschien, slordigheid of niet helemaal goed afgetikt? We hebben wel vaker berichten over AI dat op hol slaat, zeg ik maar even populair. Alibaba, een Chinese bedrijf, had onlangs een incident, een loss of control incident noemden ze dan. Dat ze geen controle meer hadden over de techniek. Of slordigheid, of was dat ook op hol slaande techniek?

B: Het komt er feitelijk op neer, je weet nooit precies wat zo'n AI gaat doen. Dus het is een beetje een illusie dat je sowieso control over dat ding hebt. En als je hem dan aansluit op het internet en je laat hem zijn gang gaan, dan moet je niet verbaasd zijn dat hij ook daar wel eens rare dingen gaat doen. Dus het idee van loss of control impliceert dat je normaal wel control hebt. Maar dat is gewoon helemaal niet zo.

S: Ja, en het gaat dus, als ik je goed beluister, om het goed programmeren of goed aangeven van de begrenzing. Van wat mag je wel, wat mag je niet. En waar mag je wel of niet in?

B: Dat is de ene. Maar de andere is ook wel van, je kan het dus goed in de gaten houden. Maar dan moet je dus wel opletten, van houdt hij zich aan zijn instructies? En het is een

beetje naïef om de AI zelf te zeggen, je moet geen stoute dingen doen. Want het is geen mens of zo.

S: Het is een programma.

B: Maar we kunnen wel heel goed meten. Dus als je zegt, je mag alleen deze computer gaan hacken, dan kan je hem uitzetten op het moment dat hij iets anders blijkt te hacken. Ja.

S: Nou, wie controleert of die begrenzings en die voorwaarden die worden gesteld in dit soort programma's nou volgens de wet zijn, ethisch zijn enzovoort. Is daar controle op?

B: Nee, eigenlijk niet. Het is nu, de slager keurt zijn eigen vlees. Maar juridisch gezien klopt dat ook wel. Want de AI is geen eigen ding. Dus de AI is nu eigenlijk, als je zegt, de AI heeft dit gehackt? Dan zeg ik, nee, dat is niet zo. Dit bedrijf OpenAI heeft iemand gehackt met gebruik van computerprogramma's. Dus wie houdt dit in toom? Dat is eigenlijk gewoon het strafrecht. Dat zegt van, joh, jullie mogen niet andere mensen hacken. En er had eigenlijk gewoon aangifte gedaan moeten worden. En er hadden nu mensen op een politiebureau moeten zitten. Maar op de manier waarop wij nu naar AI kijken, vindt iedereen dit bijna gewoon een soort status apart hebben. En zeggen, joh, wat gek wat er nou is gebeurd. Maar er is eigenlijk weinig regelgeving. En de regelgeving die we hebben, doen we een soort niet.

S: Ja. Dus zeg maar, de filmscenario's waarin AI ons voorbijstreeft en de boel overneemt, dat kan eigenlijk alleen gebeuren als je niet begrenst, als je de aanstuurders van de AI niet begrenst.

B: Juist. En als je het apparaat dus op het internet aansluit en daarna naar huis gaat en zegt, laat die computer z'n ding maar doen. Ja, dan gebeuren er hele gekke dingen. Maar daar ben je als mens zelf mee begonnen.

S: Ja. Dus dat moet je niet doen. Dankjewel, Bert Hubert.

[← Previous](#)

Kan een Amerikaans bedrijf met encryptie de Amerikaanse overheid buiten de deur houden?

AI: Overwegingen voor wie erover gaat

[Next →](#)



© 2014-2026 bert hubert