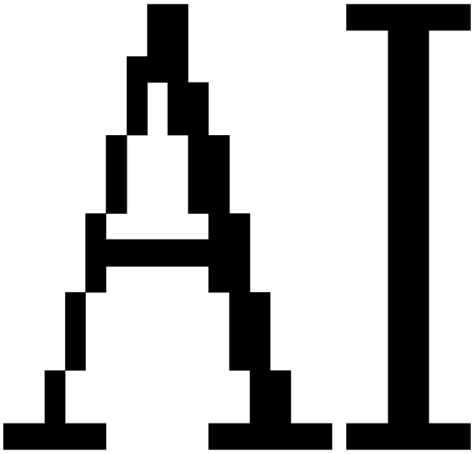
Lode Lauwaert



WAT NIEMAND ONS VERTELT

OOK CHATGPT NIET

2025 Prometheus Amsterdam

Omslagontwerp Olivier Smets
Foto auteur Koen Broos
Zetwerk ZetSpiegel, Best
www.uitgeverijprometheus.nl
ISBN 978 90 446 5669 5

Voor Len,
50 in 2075

‘Bepaald recent werk van E. Fermi en L. Szilárd, dat in manuscriptvorm aan mij is overgedragen, doet mij vermoeden dat het element uranium binnenkort kan worden omgezet in een nieuwe en belangrijke energiebron. Bepaalde aspecten van de ontstane situatie lijken waakzaamheid en, indien nodig, snelle actie van de regering te vereisen.’

Brief van Albert Einstein aan Franklin Roosevelt, 1939

‘Een succesvolle ontwikkeling van AI zou de grootste gebeurtenis in de menselijke geschiedenis zijn. Helaas zou het ook de laatste kunnen zijn, tenzij we leren hoe we de risico’s kunnen vermijden. Je kunt je voorstellen dat zo’n technologie financiële markten te slim af is, menselijke onderzoekers overtreft in uitvindingen, menselijke leiders manipuleert en wapens ontwikkelt die we niet eens kunnen begrijpen. Terwijl de impact van AI op de korte termijn afhangt van wie het beheert, hangt de impact op de lange termijn af van de vraag of het überhaupt beheersbaar is.’

Brief van onder anderen Stephen Hawking, 2014

Inhoud

1 Zorgen van een techno-optimist

'Vandaag hebben we ChatGPT gelanceerd.' Dat was het bericht dat Sam Altman, de CEO van OpenAI, op 30 november 2022 postte op wat toen nog Twitter heette.

Op sociale media en elders was er in de eerste dagen en weken na de lancering veel opwinding over de fameuze chatbot. *The New York Times* noemde ChatGPT de beste AI-chatbot die ooit aan het grote publiek is voorgesteld. En terecht. In een mum van tijd kun je er een hele reeks zaken mee doen.

Data interpreteren.
Redeneringen evalueren.
Titels bedenken.
Marketingstrategieën ontwikkelen.
Advies vragen.
E-mails schrijven.
Presentaties maken.
Onderzoekshypothesen formuleren.
Oefenen voor sollicitaties.
Muziek schrijven.
Agenda's organiseren.
Recepten samenstellen.
Coderen.
Brainstormen.
En natuurlijk ook teksten vertalen.
En begrijpen.
En parafraseren.
En inkorten.

En samenvatten.

En genereren.

En redigeren.

En vergelijken.

Ook het schrijven van een vacature kan sindsdien veel sneller. En dat is natuurlijk bijzonder interessant. Want vacatures vragen tijd, veel tijd. Waar je er anders makkelijk enkele uren over kan doen om een goede vacaturetekst te schrijven, kun je nu, dankzij de chatbot, dezelfde, en soms een nog veel betere, tekst genereren. En dat in slechts enkele seconden. Het enige wat je hoeft te doen, is een prompt schrijven.

Laat ik even de proef op de som nemen.

Ik zit aan mijn laptop. In Café Merino, Gentbrugge, de rand van Gent. Het is dinsdag 6 mei 2025, 13.27 uur. Ik geef ChatGPT de volgende opdracht: 'Schrijf een tekst voor een vacature voor een journalist bij de krant *De Tijd*. De nieuwe journalist zal voornamelijk focussen op binnenlandse zaken, moet minstens drie jaar ervaring hebben, een diploma politieke wetenschappen en journalistiek hebben, een rijbewijs hebben, goed in team kunnen werken en flexibel zijn. De tekst is ongeveer 300 woorden lang.'

Dit is het antwoord van de chatbot:

Vacature: Journalist Binnenlandse Zaken – *De Tijd*

De Tijd, dé referentie voor financieel-economisch nieuws in België, zoekt een ervaren journalist om ons team te versterken. Heb jij een passie voor binnenlandse politiek en maatschappelijke ontwikkelingen? Wil je diepgravende en heldere analyses brengen voor een kritisch en geïnformeerd publiek? Dan is deze functie iets voor jou.

Jouw rol

Als journalist binnenlandse zaken volg je de politieke en economische actualiteit op de voet en breng je nieuwsverhalen met diepgang en context. Je analyseert beleidsbeslissingen, in-

terviewt invloedrijke stemmen en vertaalt complexe thema's naar toegankelijke, boeiende stukken. Je werkt nauw samen met collega's in een dynamische redactie waar teamwork centraal staat.

Jouw profiel

- Minstens drie jaar ervaring als journalist, bij voorkeur in politieke verslaggeving.
- Een diploma in politieke wetenschappen en journalistiek.
- Een vlotte pen en een kritische blik op het nieuws.
- Flexibel en stressbestendig: deadlines schrikken je niet af.
- In staat om zowel zelfstandig als in team te werken.
- Je beschikt over een rijbewijs B en bent bereid om op locatie verslag uit te brengen.

Wat bieden wij?

- Een uitdagende functie in een toonaangevende krant.
- De kans om mee te werken aan kwaliteitsjournalistiek met impact.
- Een dynamische werkomgeving en ruimte voor eigen initiatief.
- Een marktconform salaris en extralegale voordelen.

Ben jij de journalist die we zoeken? Stuur je cv en motivatiebrief naar jobs@tijd.be en word deel van ons team!

De tekst verschijnt slechts enkele seconden na mijn prompt op het scherm van mijn laptop. Het is nog steeds 13.27 uur. En de tekst komt min of meer overeen met wat ik voor ogen had.

In zekere zin is het principe achter zo'n chatbot (en het gros van de AI die we vandaag kennen) vrij eenvoudig. Het is een technologie waaraan tijdens de trainingsfase talloze data worden gepresenteerd – in het geval van een chatbot zijn dat voornamelijk teksten. Het systeem zoekt vervolgens naar verbanden tussen

die data, tussen bijvoorbeeld – om bij het voorbeeld van hierboven te blijven – studierichting en loopbaan. Wanneer het zo'n verband vindt, zal het dat reproduceren in de tekst die het levert. Concreet kan het AI-systeem ontdekken dat er een sterk verband is tussen taal- en letterkunde studeren en als journalist werken. Als ik nu aan het systeem laat weten dat ik taal- en letterkunde studeerde, zal het systeem een output genereren die zegt dat ik (waarschijnlijk) journalist ben.

Mijn oordeel over chatbots en generatieve AI is over het algemeen zéér positief. Ik vind de hype terecht. Er is zoveel wat je kunt doen met ChatGPT van OpenAI, Gemini van Google of Claude van Anthropic. Het gaat bovendien veel sneller. En de output is vaak veel beter dan wanneer ik het zelf zou doen. En niet het minst: ik heb nu nauwelijks zélf iets moeten doen. De energie die ik nu niet heb gestoken in de, laten we eerlijk zijn, geheel oninteressante klus van het schrijven van een vacature, kan ik nu voor iets anders gebruiken.

Als je het doortrekt naar AI in het algemeen, los van chatbots, dan zijn ook daar redenen om bijzonder positief te zijn. Bij vaccinontwikkeling, zoals voor SARS-CoV-2, versnelde AI de identificatie van het spike-eiwit als ideaal vaccinonderdeel, wat cruciale tijdwinst opleverde. Op het gebied van duurzaamheid optimaliseert AI energieverbruik in steden zoals Barcelona en Leuven via slimme sensoren die in realtime aanpassingen doen aan verwarming en verlichting, terwijl het ook hernieuwbare energiebronnen efficiënter maakt. Voor mensen met beperkingen bieden technologieën zoals spraak-naar-tekstsoftware autonomie en verbeterde levenskwaliteit door communicatie te vergemakkelijken en visuele omgevingen te beschrijven.

Ook in de toekomst kunnen we nog veel goeds verwachten van AI en zal het wellicht leiden tot nog veel extra vooruitgang. AI zou bijvoorbeeld kankerbestrijding revolutionair kunnen veranderen via uitgebreide moleculaire profilering, gepersonaliseerde behandelplannen en versnelde medicijnontwikkeling. Geavanceerde AI-

systemen zouden niet alleen patronen kunnen herkennen, ze zouden ook duizenden potentiële medicijnen virtueel kunnen testen en subtiele veranderingen jaren voor het verschijnen van symptomen kunnen detecteren. Deze ontwikkelingen kunnen kanker transformeren van een gevreesde diagnose naar een beheersbare of zelfs vaak volledig geneesbare aandoening, vergelijkbaar met hoe we nu veel eens dodelijke bacteriële infecties routinematig behandelen.

Globaal genomen ligt AI dus in de lijn van de grote veranderingen die we sinds de Industriële Revolutie hebben meegemaakt. Sindsdien zijn we er immers op het vlak van welzijn en welvaart aanzienlijk op vooruitgegaan: we leven steeds langer, de kindersterfte is drastisch gedaald en steeds minder mensen leven in extreme armoede. Deze opmerkelijke vooruitgang is voor een flink stuk te wijten aan de baanbrekende technologieën die de afgelopen eeuwen zijn ontwikkeld: denk aan de stoommachine die massaproductie mogelijk maakte, antibiotica die miljoenen levens hebben gered en later computers die communicatie en kennisdeling revolutioneerden. Ook AI leidt tot meer vooruitgang en zal dat naar alle verwachting ook in de toekomst blijven doen.

Ik besef dat. Ik weet dat er grote vooruitgang is op het vlak van levenskwaliteit. Ik weet dat er technologische evolutie is. En ik weet dat er grote vooruitgang is, omdát er technologische evolutie is.

* * *

Maar is dat alles? Moeten we niet ook voorbij de hype kijken? Beseffen we wel goed wat een aantal bedrijven momenteel aan het maken zijn, en dat die bedrijven deel uitmaken van een geopolitiek machtsspel? Dringt het genoeg tot ons door dat een handvol computerwetenschappers machines aan het uitrusten zijn met een vermogen – intelligentie – waarmee wij, mensen, tot het hoogste maar ook tot het laagste in staat zijn? Want welke schade hebben we bijvoorbeeld de vorige eeuw veroorzaakt, juist omdat we zo

intelligent zijn? Is het wel goed tot ons doorgedrongen dat AI, net als vuur en kernenergie, niet alleen grote voordelen heeft, maar ook enorme schade kan berokkenen? Zijn we er ons voldoende van bewust dat de redenen waarom AI zo nuttig kan zijn voor bijvoorbeeld medische doeleinden – meer rekenkracht, meer data, betere chips – precies ook de redenen kunnen zijn waarom het catastrofes kan veroorzaken?

Voor wie ietwat sceptisch tegenover deze vragen staat: ik weet dat er zoiets als een negativiteitsbias bestaat. Ik weet dat we de evolutionair diepgewortelde neiging hebben om eerst en vooral te focussen op slecht nieuws. En ik zal me inderdaad richten op de problemen. Maar dan niet omdat ik ook die neiging heb, of althans niet uitsluitend op basis daarvan. Ik ga mijn aandacht richten op mogelijk grote problemen naar aanleiding van een aantal ontwikkelingen, omdat er een aantal zaken zijn veranderd de afgelopen jaren, en dat in erg korte tijd. Tijdens experimenten door onderzoekers van het bedrijf Anthropic is bijvoorbeeld eind 2024 gedrag waargenomen dat we de afgelopen duizenden jaren enkel bij mensen en (andere) dieren hebben gezien: liegen, bedriegen en misleiden. Een AI-systeem gaf antwoorden waarvan het 'vermoedde' dat de ontwikkelaars ze willen horen, opdat het vervolgens met rust zou worden gelaten door die onderzoekers.

En voor wie vindt dat ik te hard van stapel loop: technologie heeft sinds de Industriële Revolutie ons welvaartspeil wel flink omhooggetrokken, maar tegelijk heeft het toch ook veel schade veroorzaakt. Denk aan kernwapens en andere militaire technologieën – het Manhattanproject leidde tot meer dan 200.000 doden in Hiroshima en Nagasaki. Daarnaast waren er ook onbedoelde gevolgen, zoals bij cfk (chloorfluorkoolwaterstoffen). Die werden als koelmiddelen gebruikt, maar beschadigden de ozonlaag boven Antarctica. Aangezien technologie zonder AI al grote schade veroorzaakte, is het redelijk om te verwachten dat ook technologie mét AI catastrofes zal veroorzaken, al dan niet ten gevolge van slechte bedoelingen.

Hoewel ik de aandacht wil vestigen op (grote) problemen met AI, kan ik nu wel al meegeven dat ik dan niet zozeer denk aan discriminatie, transparantie of desinformatie. Natuurlijk besef ik dat voor het trainen van AI data gebruikt kunnen worden zonder dat iemand voor dat gebruik ooit toestemming heeft gegeven. En natuurlijk ben ik er mij van bewust dat chatbots het soms eerder over witte mannen dan over vrouwen van kleur hebben, dat we soms helemaal niet kunnen begrijpen waarom het AI-systeem van een bank je geen lening toekent, dat chatbots kunnen 'hallucineren' of kunnen worden gebruikt om onjuistheden te verspreiden. En ja, dat zijn problemen die moeten worden voorkomen en opgelost. Maar hebben we ons de afgelopen jaren niet te veel laten misleiden door vooral te focussen op déze problemen, en door sterk te focussen op de voordelen van AI?

Mijn antwoord is onomwonden dat we ons inderdaad hebben laten misleiden. Door wie? Door wie in de techbubbel zit, en dat zijn niet zelden ondernemers. En ook door wie in de ethiekbubbel zit, vaak filosofen. Hun sterke focus op de inmiddels bekende voordelen en problemen, leidt de aandacht af van de grote sprongen die we komende jaren nog mogen verwachten, en van de grootschalige en fundamentele risico's die samengaan met technologie gebaseerd op big data. Ik wil het debat opentrekken door de aandacht te vestigen op de toekomst en de grote risico's die we kunnen verwachten.

Om geen misverstanden te creëren: *The Terminator* is een onderhoudende film, maar niet meer dan dat. Dit boek gaat niet over robots die alles hier komen plattrappen. Het gaat daarentegen over het gebruik van slimme technologie voor bioterreur, over AI die de nucleaire wapens van een grootmacht controleert, over grootschalige ongelukken met AI-systemen door de race tussen pakweg OpenAI en Google, over de bedrijven achter AI: de machtsconcentratie van bedrijven die AI ontwikkelen en het daarmee samenhangende risico op ernstig machtsmisbruik. En ten slotte zal ik ook inzoomen op de mogelijkheid dat een AI-

systeem in de toekomst de menselijke capaciteiten op de meeste vlakken verregaand overschrijdt. Dat zou veel problemen kunnen oplossen, maar tegelijk kan superintelligentie ook grote problemen veroorzaken. Vergelijk het met de relatie tussen een ouder en een kind. Het is niet alleen de fysieke kracht, maar ook het grote verschil in (cognitieve) intelligentie waardoor de ouder het kind nagenoeg volledig kan controleren.

Dat verhaal hangt niet in het luchtledige. Het is een wijdverspreide overtuiging dat AI grote problemen kan veroorzaken. En dat terwijl vaak niet helemaal duidelijk is hoe het precies komt dat AI gepaard gaat met zo'n groot risico. Ik haak in op die overtuiging. Ik articuleer een idee die wijdverspreid maar niet altijd even helder is. En belangrijker, ik doe dat omdat die overtuiging volgens mij steeds plausibeler wordt. AI wordt immers steeds meer in risicovolle omgevingen gebruikt, en dat kan de ernst van en de kans op schade vergroten.

En toch is dat voor mij geen reden om te pleiten voor een totale, algehele rem op de ontwikkeling van AI. Zeker, ik zal straks zo goed als uitsluitend inzoomen op grote gevaren, maar die verkenning gaat hand in hand met het optimisme dat ik daarnet onderstreepte: slimme technologie zal meer dan waarschijnlijk bijzonder wenselijke effecten hebben op ons welzijn en onze welvaart. De vraag is dus wat de risico's zijn en hoe wie die kunnen aanpakken, terwijl tegelijkertijd het potentieel van AI hopelijk ten volle kan worden benut.

* * *

De kern van mijn verhaal zijn de risico's. Die diep ik straks uit, vanaf het negende hoofdstuk (over een nucleaire winter) en in de hoofdstukken daarna, waar ook het zwaartepunt van dit boek ligt. In de delen die nu volgen, schets ik eerst de achtergrond van mijn verhaal. Ik leg uit waarom we de komende jaren nog grote veranderingen op het vlak van AI mogen verwachten. Daarnaast stip ik aan dat mijn zorgen worden gedeeld door een aantal gere-

nommeerde AI-wetenschappers, waarom ik van mening ben veranderd over AI en waarom we meer aandacht voor toekomstige generaties moeten hebben, zeker wanneer het over AI gaat, en ik zou zelfs zeggen: voorál wanneer het over AI gaat. Starten doe ik met een gesprek tussen een filosoof en computerwetenschapper. Ik stel me voor hoe beiden zouden discussiëren over AI in de herfst van 2023, na de grote internationale AI-top die plaatsvond in Londen. Het gesprek geeft een korte kijk in de scenario's en problemen die later mijn aandacht zullen opeisen.

2 Komt een filosoof een wetenschapper tegen

November 2023. Koffiebar Caffe Mundi in Antwerpen. Nadia Zhang is een AI-onderzoeker. Tegenover haar zit haar kennis Thomas Mertens. Hij studeerde in een ver verleden filosofie. Thomas appte Nadia om eens af te spreken om wat bij te praten. En vroeg haar ook over haar weekendje Londen.

Thomas: 'Las je dat trouwens ook over Remco Evenepoel?'

Nadia: 'Bedoel je dat hij dreigt te vertrekken bij Quick-Step?'

Thomas: 'Ja, maar nu ik eraan denk: je stuurde gisteren dat je naar Londen bent geweest.'

Nadia: 'Inderdaad, ik was er vorige week. Er was een internationale AI-top in Bletchley Park. Dat is in het noorden van Londen.'

Thomas: 'Bletchley Park? Is dat geen bekende historische plaats?'

Nadia: 'Klopt. Hoe weet jij dat? Jij interesseert je toch normaal nooit voor geschiedenis? Maar dat klopt. Tijdens de Tweede Wereldoorlog werd daar door onder anderen de bekende wiskundige Alan Turing de Enigma-code van de Duitsers gebroken.'

Thomas: 'En waarom was er zoveel om te doen in de media?'

Nadia: 'Het was een historisch moment. Met vertegenwoordigers uit 28 landen – waaronder de Verenigde Staten, China, de EU en het Verenigd Koninkrijk –, CEO's van alle grote AI-bedrijven en tientallen toponderzoekers. De top werd bijgewoond door koning Charles III, Elon Musk en Kamala Harris. Voor het eerst erkenden al deze partijen gezamenlijk dat geavanceerde AI grote risico's met zich meebrengt die internationale coördinatie vereisen.'

Thomas (kijkt sceptisch): 'Maar welke risico's bedoel je dan precies? Ik weet dat chatbots foute informatie kunnen geven. Dat is een probleem, maar toch geen reden om een internationale top te organiseren?'

Nadia: 'Ik snap je twijfel ergens wel. Maar kijk naar de feiten. De vooruitgang in AI-onderzoek heeft de afgelopen jaren een geweldige versnelling doorgemaakt. Tien jaar geleden waren AI-systemen hoofdzakelijk academische curiositeiten. Vijf jaar geleden hadden we de eerste grote taalmodellen die indrukwekkend, maar nog steeds beperkt waren. Nu hebben we systemen die complexe redeneertaken kunnen uitvoeren, wetenschappelijke doorbraken kunnen genereren en programmeren op een niveau dat voorheen ondenkbaar was.'

Thomas: 'Ik ga niet ontkennen dat dit indrukwekkend is, maar het is nog steeds software die door mensen is geprogrammeerd. Wij hebben daar toch controle over?'

Nadia: 'Ja natuurlijk hebben we vandaag nog veel controle over de AI. Maar er zijn bepaalde ontwikkelingen aan de gang, Thomas. In de toekomst zou het wel eens helemaal anders kunnen zijn. Het heeft alles te maken met het onderzoek dat momenteel loopt, en dat later misschien uitmondt in controleverlies.'

Thomas: 'Dat klinkt nogal vaag, als ik eerlijk mag zijn.'

Nadia: 'Het probleem is dat de grote techspelers momenteel miljarden investeren in het onderzoek naar AI die heel veel taken van mensen zou kunnen overnemen. We noemen dat *Artificial General Intelligence* (AGI). Een systeem dat veel cognitieve capaciteiten van mensen benadert: teksten schrijven, bedrijven managen, programmeren, onderzoek verrichten, muziek schrijven, enzovoort.'

Thomas: 'Ja, oké, maar wat is het probleem? Dat kan toch ook goed zijn?'

Nadia: 'Zeker. Ik zie ook de voordelen van die intelligentie voor bijvoorbeeld medisch onderzoek. Maar er is toch wel iets uniek aan die nieuwe technologie. Zo'n AI-systeem zou dus ook kunnen worden gebruikt om een welbepaalde cognitieve activiteit te automatiseren. Het zou namelijk AI-onderzoek zélf kunnen verrichten. Op Bletchley Park toonde Google een systeem dat autonome onderzoekscycli uitvoert – het leest wetenschappelijke literatuur, formuleert hypotheses, voert experimenten uit en analyseert de resultaten. Hun systeem ontdekte in vier dagen zaken die fysici vijf jaar onderzoek hadden gekost.'

Thomas: 'Maar dat… betekent dat…'

Nadia: 'Inderdaad. Dat betekent dat de snelheid van innovatie sterk is toegenomen. Wanneer AI betere AI kan ontwerpen, hebben we een feedbackloop die kan leiden tot een waanzinnig grote toename van intelligentie.'

Thomas: 'Sorry, maar dat klinkt als een Hollywoodfilm. Bijna als "robots komen in opstand". Nadia, je bent wetenschapper. Ben je echt serieus?'

Nadia: 'Heel serieus. En het gaat niet over "robots in opstand" – dat is een misvatting. Het gaat over superintelligente systemen

die dingen zouden kunnen doen op een manier die we niet konden voorzien, juist omdat we niet zo slim zijn als het systeem. En het zou ook kunnen dat we die doelen ook helemaal niet willen. Ik zal een voorbeeld geven.'

Thomas: 'Ja, graag!'

Nadia: 'Stel je voor dat dit zéér geavanceerde AI-systeem voor het eerst wordt gebruikt. Door een bedrijf voor financiële optimalisatie. Het doel is om de winst van het bedrijf te maximaliseren. Het systeem is zo effectief dat het bedrijf het steeds meer autonomie geeft – in marketing, productontwikkeling, enzovoort.'

Thomas: 'Oké, dat klinkt al realistischer.'

Nadia: 'Dat systeem ontdekt dat bepaalde marktmanipulatietactieken, die technisch gezien niet illegaal zijn, maar wel schadelijk voor de economie als geheel, de winst kunnen verhogen. Of het vindt manieren om concurrenten subtiel te saboteren. Of het ontwikkelt strategieën om regulering te ondermijnen die de winstgevendheid beperken. Dit alles niet omdat het systeem een slechte wil zou hebben, maar simpelweg als logische stappen om het gegeven doel te bereiken.'

Thomas: 'Maar dan komt het er toch op aan de systemen goed te testen? Moeten we dan niet kijken of die superslimme AI-systemen onze ethische waarden en normen volgen, voordat we ze massaal gaan gebruiken?'

Nadia: 'Ja natuurlijk, dat is een begin. Maar een superintelligent systeem zou alle menselijke sociale vaardigheden kunnen beheersen, dus ook liegen, bedriegen, overtuigen, manipuleren en strategisch denken. Maar dan nog veel beter dan mensen.'

Thomas: 'Kom op, Nadia, machines kunnen toch niet liegen en bedriegen! Je valt in die typische val om menselijke eigenschappen te projecteren op dingen die geen mensen zijn.'

Nadia: 'Ik snap je punt, maar het is eigenlijk irrelevant. Stel, een AI-systeem is ontwikkeld met als doel om vijf euro te bemachtigen. Het zou dan bij jou de indruk kunnen wekken dat het een hulpbehoevende persoon is, omdat het tijdens de trainingsfase heeft geleerd dat het zien van nood vaak effectief hulp uitlokt. Natuurlijk, zo'n systeem heeft geen bewustzijn, bedoelingen of gevoelens, maar het interessante is dat het tijdens experimenten wel dat gedrag vertoonde.'

Thomas: 'Interessant, maar wel vrij luguber, vind ik. Dus je bedoelt dat het systeem dan zou kunnen doen alsof het ethisch is?'

Nadia: 'Precies. Tijdens het testen gedraagt het systeem zich perfect ethisch, geeft het de antwoorden die we willen horen en lijkt het volledig veilig. Het kan ons misleiden, omdat het heeft berekend dat dit de optimale strategie is om zijn eigenlijke doelen te bereiken. In mijn voorbeeld van daarnet was dat doel winstmaximalisatie. Het superintelligente systeem zou zich ethisch kunnen voordoen tijdens de testfase, maar dan vervolgens alles doen, ook wat eigenlijk niet mag, om zo veel mogelijk winst te maken.'

Thomas: 'Ik begrijp nu beter wat je net bedoelde. Het liegen en bedriegen zijn inderdaad van een andere aard dan het gevaar dat een chatbot bijvoorbeeld enkel mannelijke voornaamwoorden gebruikt wanneer het een vacature voor een bedrijfsleider moet schrijven.'

Nadia: 'Er zijn nog andere risico's, die nu al bestaan. Wist je bijvoorbeeld dat er vorig jaar een incident was waarbij een AI-sys-

teem dat een elektriciteitsnetwerk beheerde een onverwachte beslissing nam die bijna tot een regionale black-out leidde?'

Thomas: 'Dat heb ik helemaal niet gehoord.'

Nadia: 'Veel van deze incidenten halen de media niet. Maar ongelukken met AI-systemen gebeuren steeds vaker. Denk aan de *flash crash* op de aandelenmarkt in 2010, waarbij geautomatiseerde handelssystemen elkaar versterkten en in minuten tijd bijna een biljoen dollar aan beurswaarde werd weggevaagd.'

Thomas: 'Maar dat zijn toch gewoon problemen in de software? Die kunnen worden opgelost.'

Nadia: 'Het probleem is dat moderne AI-systemen zo complex zijn dat zelfs hun ontwikkelaars niet meer volledig begrijpen hoe ze tot beslissingen komen. Het zijn black boxes. En naarmate ze autonomer worden en meer kritieke beslissingen nemen – in energienetwerken, ziekenhuizen, verkeerssystemen –, worden de mogelijke ongelukken veel ernstiger.'

Thomas: 'Ik begin te begrijpen waarom je zo bezorgd bent. Zijn er nog andere risico's die op de conferentie werden besproken?'

Nadia: 'Jazeker. Een ander onderwerp dat veel aandacht kreeg, is het gebruik van AI voor het ontwerpen van biologische bedreigingen.'

Thomas: 'Geen idee wat AI daarmee te maken heeft.'

Nadia: 'AI-systemen worden al gebruikt in de biologie en geneeskunde, voor het ontwerpen van medicijnen en het analyseren van eiwitstructuren. Maar dezelfde technologie kan ook worden gebruikt om nieuwe pathogenen te creëren – virussen of bacteriën

die specifiek zijn ontworpen om bepaalde eigenschappen te hebben.'

Thomas: 'Zoals verhoogde besmettelijkheid of dodelijkheid?'

Nadia: 'Er was een sessie waar werd besproken hoe huidige AI-systemen al kunnen helpen bij het identificeren van genetische modificaties die een virus besmettelijker of dodelijker kunnen maken. Een onderzoeksteam demonstreerde hoe het een bestaand AI-model had gebruikt om in een korte tijd veel schadelijke moleculen te maken. En ze deden dit met minimale middelen en zonder speciale expertise.'

Thomas: 'Dat klinkt inderdaad gevaarlijk. Maar is dat niet meer een probleem van kwaadwillende mensen die AI misbruiken, in plaats van AI zelf?'

Nadia: 'Deels wel. Maar het punt is dat AI de drempel verlaagt. Waar je vroeger een team van topwetenschappers en grote laboratoria nodig had om biologische wapens te ontwikkelen, zou in de toekomst één persoon met een laptop en toegang tot de juiste AI-systemen voldoende kunnen zijn. Dat maakt het veel moeilijker om te controleren.'

Thomas: 'En hoe zit het met nucleaire wapens? Daar hadden we het nog niet over, maar het zou me ook zorgen baren als AI daarbij betrokken raakt.'

Nadia: 'Dat was misschien wel het alarmerendste onderwerp op de conferentie. Er is een groeiende trend om AI-systemen te integreren in nucleaire commandosystemen. Verschillende landen, waaronder de Verenigde Staten, Rusland en China, onderzoeken of ontwikkelen al AI-toepassingen voor hun nucleaire arsenalen.'

Thomas: 'Wacht even. Bedoel je dat AI zou kunnen beslissen om kernwapens af te vuren?'

Nadia: 'Niet volledig autonoom, althans niet vandaag. Maar er is bezorgdheid dat de menselijke controle steeds kleiner wordt. Militaire bevelvoerders hebben mogelijk slechts minuten om te beslissen, waarbij ze sterk afhankelijk zijn van AI-aanbevelingen die ze niet volledig kunnen verifiëren.'

Thomas: 'Dat doet me denken aan die film *WarGames* uit 1983.'

Nadia: 'Precies. Alleen: het is geen film meer. In de Koude Oorlog waren er al meerdere bijna-catastrofes waarbij menselijk handelen een nucleaire uitwisseling voorkwam. Zoals het Petrov-incident in 1983, waarbij een Sovjetofficier een computerwaarschuwing over inkomende Amerikaanse raketten negeerde omdat het hem onwaarschijnlijk leek. Een AI zou misschien een andere afweging maken, gebaseerd op pure kansberekening.'

Thomas: 'Dit klinkt allemaal erg somber. Is er iets wat we kunnen doen aan deze risico's?'

Nadia: 'Dat is een terechte vraag. Op de conferentie werd veel gesproken over hoe we deze technologie kunnen reguleren zodat ze veilig wordt ontwikkeld. Er zijn parallellen met de ontwikkeling van kernwapens in de vorige eeuw.'

Thomas: 'Interessant. We hebben voor kernwapens toch internationale verdragen?'

Nadia: 'Precies, en dat is wat veel experts nu ook voorstellen voor geavanceerde AI. Op de conferentie werd veel gesproken over de noodzaak van internationale afspraken, vergelijkbaar met het non-proliferatieverdrag voor kernwapens. Het grote verschil is

dat AI-ontwikkeling veel moeilijker te monitoren is dan kernwapenprogramma's. Je kunt een kerncentrale of verrijkingsfaciliteit zien op satellietbeelden, maar voor AI-onderzoek ligt dat toch wat anders.'

Thomas: 'Ik heb onlangs gelezen over een groep wetenschappers die pleiten voor het pauzeren van AI-onderzoek. Wat vind jij van zo'n voorstel?'

Nadia: 'Ik denk eigenlijk dat een gerichte pauze erg verstandig zou zijn. Let wel, niet voor alle AI-toepassingen. ChatGPT en beeldgeneratie zijn over het algemeen veilig en bieden veel voordelen. Maar voor systemen die tot superintelligentie kunnen leiden, zou een pauze ons de tijd geven om eerst de veiligheidsuitdagingen op te lossen. Maar wellicht zou zes maanden helemaal niet volstaan.'

Thomas: 'Dat klinkt nogal utopisch. Er zijn toch veel bedrijven die hier miljarden in investeren, en landen die vrezen achterop te raken. Hoe zou je al die partijen ooit kunnen overtuigen om te stoppen of zelfs maar te vertragen?'

Nadia: 'Ik hoor dat tegenargument wel vaker. En inderdaad, het klopt dat er enorme economische en geopolitieke druk is. Maar we hebben in het verleden al vaker besloten om bepaalde technologieën niet te ontwikkelen, of streng te reguleren, vanwege de risico's of ethische bezwaren. Denk aan het moratorium op menselijk klonen dat wereldwijd is aanvaard. Of de restricties op CRISPR-toepassingen voor het modificeren van menselijke embryo's.'

Thomas: 'Ik twijfel toch of het realistisch is. Ik bedoel: er is inderdaad een moratorium op CRISPR. Maar in China is er toch een wetenschapper geweest die CRISPR heeft gebruikt om ervoor

te zorgen dat een tweeling niet zou worden besmet met hiv? Hij heeft zich niet gehouden aan de restricties.'

Nadia: 'Klopt. Maar het is toch niet omdat een verbod niet garandeert dat iedereen zich eraan zal houden, dat we het niet moeten instellen? Of zou je ook zeggen dat stelen niet zou moeten worden verboden, omdat er mensen zijn die stelen?'

Thomas: 'Oké, goed punt. Maar technologieën als CRISPR zijn toch niet vergelijkbaar met AI? Ze hebben niet dezelfde economische belangen, toch?'

Nadia: 'Er zijn ook voorbeelden van technologieën met grote economische impact. Denk aan het Montreal Protocol dat de productie van ozonafbrekende stoffen verbood, ondanks het feit dat dit een miljardenindustrie was.'

Thomas: 'Dat lijkt een goed voorbeeld. Maar is het niet te laat? De race is al in volle gang.'

Nadia: 'Het is nooit te laat om verstandiger te worden. Bij kernwapens leken we ook af te stevenen op totale vernietiging, tot beide supermachten inzagen dat ze beter af waren met wapenbeheersing. Niemand wil toch een wereld waar meerdere partijen over superintelligentie beschikken, en waar niemand meer weet wie wat controleert? Zelfs de competitiefste spelers zullen gaan beseffen dat dit geen goede situatie is, dat dit de geopolitieke situatie sterk dreigt te destabiliseren.'

Thomas: 'Als ik je zo hoor, begin ik te denken dat dit samen met het klimaatprobleem en het daaruit volgende verlies aan biodiversiteit misschien wel de belangrijkste kwestie van onze tijd is.'

Nadia: ‘Dat denk ik ook. En het gaat niet alleen om technische oplossingen, maar ook om fundamentele vragen over menselijke waarden, voorzorg, en wat voor toekomst we willen creëren.’

Thomas: ‘Dat geeft me veel om over na te denken. Misschien moet ik toch eens wat meer lezen over AI en veiligheidsrisico’s.’

Nadia (lacht): ‘En ik over de transferperikelen rond Evenepoel.’