Illustratie Fieke Ruitinga

ACHTERGROND

‘AI-doomers’ denken dat AI tot het einde van de mensheid zal leiden. ‘Het zal proberen de macht over te nemen’

AI Techondernemer Joep Meindertsma gelooft dat er een grote kans is dat AI tot het einde van de mensheid zal leiden. Andere ‘AI-doomers’ delen zijn angst. „Wat nou als dat ding slimmer is dan wij? Hoe houd je het dan tegen?”

Stijn Bronzwaer 16 februari 2024 Leestijd 11 minuten

Luisteren Leeslijst

De eerste keer dat hij huilt, is bij het zien van Auto-GPT. Een AI-programma dat zelfstandig het internet opgaat en webpagina’s opent. Software-ontwikkelaar Joep Meindertsma (33) zit thuis achter zijn beeldscherm in Utrecht, en dit is het moment waar hij al maanden voor vreest.

Huilen zal hij in de weken daarna vaker doen. Bijvoorbeeld als hij op familiebezoek gaat. Het gesprek gaat over zijn angsten. Dat iemand artificiële intelligentie (AI) gebruikt om ons financiële systeem plat te leggen. Het internet uit te schakelen. Lege supermarkten. Rellen. Chaos. En zijn grootste vrees: dat er een computersysteem komt waar de mens geen controle meer over heeft. Een wereld die door kunstmatige intelligentie wordt geregeerd.

Zijn familie luistert en stelt veel vragen. Wat ze, als het dan misgaat, het beste kunnen doen. „Ze vonden het heftig om te zien

dat ik bang ben voor de toekomst", zegt Meindertsma. „Mijn familie kent mij juist als iemand die heel optimistisch is over technologie."

Enkele weken eerder is Meindertsma minder gaan werken bij zijn start-up, softwarebedrijf Ontola (drie werknemers) in Utrecht, dat bedrijven helpt inzicht te krijgen in hun data. Hij heeft het opgericht samen met Michiel van den Ingh, een goede vriend. Van den Ingh merkt dat zijn compagnon zich steeds moeilijker kan concentreren op hun bedrijf. Meindertsma wil iets doen om de opkomst van AI te stoppen, zegt hij. In mei, niet lang nadat hij voor het eerst moest huilen, richt hij een actiegroep op: Pause.AI.

> *Een computer die slimmer is dan mensen komt er hoe dan ook, gelooft Silicon Valley*

NRC sprak Joep Meindertsma de afgelopen acht maanden meerdere keren. Om beter te begrijpen waarom sommige, veelal jonge, mensen in de techsector zich ernstige zorgen maken over de opkomst van AI. Deze *AI-doomers* vrezen dat het huidige tempo van ontwikkeling kan leiden tot grote rampen, of zelfs de ondergang van de mensheid. Hun zorgen worden gedeeld door prominente AI-wetenschappers en beroemde techondernemers.

Dat Meindertsma ooit activist zou worden, had hij nooit verwacht. Ja, hij had weleens meegelopen met een klimaatprotest. Maar zelf demonstraties organiseren en petities opstellen? Hij is meer een denker, vertelt hij tijdens een van de gesprekken. Iemand die veel leest en debatteert. Meindertsma is hoogsensitief. Iemand met een groot verantwoordelijkheidsgevoel, die zich snel het leed van anderen aantrekt.

Een persoonlijk artikel als dit vindt hij „heftig en spannend", vertelt hij op kantoor bij Ontola, een zolderverdieping in een statig pand aan de Utrechtse Maliesingel. Hij is bang als complotdenker te worden weggezet. Dat lezers hem als „gekkie" zien. Dat hij het verhaal mag lezen voor publicatie geeft hem „heel veel rust", zegt hij. „Het goed neerzetten van dit onderwerp voelt als een grote verantwoordelijkheid."

Hij heeft, denkt hij, misschien nog een paar jaar tot AI leidt tot grote rampen. Die mogelijkheid schat hij op 40 procent. „Ik twijfel continu, maar de kans dat het fout gaat is groot", zegt hij. „Er komt een moment dat een AI probeert de macht over te nemen. Wat nou als dat ding slimmer is dan wij? Hoe houd je het dan tegen?"

Bij het kampvuur

Als het einde der tijden komt, dan is Sam Altman in elk geval goed voorbereid. De topman van OpenAI, het bedrijf achter chatbot ChatGPT, is een *doomsday prepper*, zo bekende hij in 2016 tijdens een feestje bij een kampvuur in Silicon Valley. Altman praat uitgebreid over zijn voorbereiding, volgens een reconstructie van het Amerikaanse tijdschrift *The New Yorker*. Als het misgaat – door een virus, een nucleaire oorlog of losgeslagen superintelligentie – dan heeft hij alles in huis. „Geweren, goud, kaliumjodide, antibiotica, batterijen, water en gasmaskers van het Israëlische leger", zei Altman. „En een groot stuk land in Big Sur [Californië], waar ik naartoe kan vliegen."

Dat de bedenker van ChatGPT vreest dat AI een existentieel risico vormt voor de mensheid, is geen toeval. Veel ondernemers die de laatste jaren bij de ontwikkeling van AI zijn betrokken, vrezen voor de gevolgen van de technologie die ze ontwikkelen.

Tesla-oprichter Elon Musk is bang voor „een vernietiging van onze beschaving" door superintelligente computers. Voormalig Google-

topman Eric Schmidt denkt dat AI „binnen vijf tot tien jaar" in staat is de mensheid te bedreigen. Ilya Sutskever, topwetenschapper bij OpenAI, vreest een wereld vol datacentra en zonnepanelen, waar ongecontroleerde AI-systemen alle beslissingen voor mensen nemen.

Die angst heeft hen juist gemotiveerd om bedrijven te starten. De heersende overtuiging in Silicon Valley is: superintelligentie, een computer die slimmer is dan mensen, komt er. Hoe dan ook. Dan kun je beter aan de knoppen zitten en zelf AI ontwikkelen.

Sinds de komst van ChatGPT in november 2022 beconcurreren techbedrijven elkaar om de beste AI-systemen. Zij zien in de taalmodellen die aan de basis liggen van chatbots de grootste economische kans sinds het internet. Techbedrijven dromen van een toekomst waarbij AI talloze taken van mensen overneemt en kunstmatige intelligentie ons assisteert in onze beslissingen.

Door het aantal AI-computerchips steeds verder op te voeren, krijgen de systemen toegang tot steeds meer rekenkracht. AI wordt zo steeds 'intelligenter', ofwel: steeds beter in het doen van taken die eerder alleen mensen konden doen. AI kan inmiddels muziek maken, juridische contracten opstellen, eiwitten herstructureren, MRI-scans beoordelen, illustraties tekenen, software programmeren en boeken schrijven.

De reacties op de chatbots zijn in eerste instantie vooral positief. Maar in maart 2023 verschijnt een open brief, ondertekend door tientallen prominente AI-wetenschappers en ondernemers. In de petitie roepen de ondertekenaars op om de ontwikkeling van AI zes maanden te pauzeren. Concreet zou dat betekenen dat bedrijven geen nieuwe producten als ChatGPT mogen uitbrengen. De boodschap slaat ook aan bij Joep Meindertsma, die met Pause.AI hetzelfde doel nastreeft.

Onleefbare planeet

De ontwikkelingen in AI gaan te snel en ongecontroleerd, vinden de ondertekenaars van de petitie. Als de overheid niet snel ingrijpt dreigen er grote problemen, denken zij. AI-systemen die cyberaanvallen uitvoeren, massaal nepnieuws produceren en verspreiden of advies geven bij het ontwerpen van virussen of biologische wapens.

Of, nog een stap verder: AI-systemen die toegang krijgen tot het financiële systeem of het elektriciteitsnet. En doelen probeert te bereiken die een maker de AI meegeeft ('maak mij rijk', 'vernietig een volk', 'win verkiezingen') en zich vervolgens niet meer laat uitschakelen. Als de mens AI niet meer in de hand heeft, zou het beslissingen kunnen nemen die de planeet onleefbaar maken.

Op dit moment zijn chatbots als ChatGPT of Google's chatrobot Gemini niets meer dan systemen die, op basis van statistiek, buitengewoon goed kunnen praten en beredeneren. AI-systemen imiteren menselijke gesprekken en luisteren naar de relatief simpele opdrachten die wij ze geven. Maar, vrezen experts: wat als AI-systemen slimmer en slimmer worden en zelf in staat zijn te bedenken hoe zij een bepaald doel kunnen bereiken? Hebben wij een computer die slimmer is dan wij nog wel in de hand?

Een beroemd voorbeeld hiervan is de paperclip-maximalisator, een gedachte-experiment beschreven door de Zweedse filosoof Nick Bostrom in zijn invloedrijke boek *Superintelligence* (2014). Daarin beschrijft Bostrom hoe superintelligente AI van zijn ontwerper als taak meekrijgt om zo veel mogelijk paperclips te produceren. Uiteindelijk gaat de AI de mens als concurrent zien voor het bemachtigen van grondstoffen om honderden miljarden paperclips te kunnen maken.

Van zichzelf hebben computers geen doel

Dat klinkt extreem. Wetenschappers achten dergelijke scenario's ook niet bijzonder realistisch. Computers hebben van zichzelf geen doel en begrijpen de wereld om zich heen niet. Critici van AI-doomers zeggen dat wij het menselijk brein onderschatten als we denken dat computers ons zomaar de baas kunnen zijn.

Tegelijkertijd zeggen AI-wetenschappers ook: AI-doemscenario's zijn niet uit te sluiten. Ongeveer de helft van die wetenschappers schat de *p(doom)* (ofwel: kans op verdoemenis door AI) op minimaal 5 procent, bleek recentelijk uit een peiling onder 2.700 van hen.

Volgens Tamar Sharon, hoogleraar filosofie, digitalisering en samenleving aan de Radboud Universiteit in Nijmegen, is het belangrijk om goed te bekijken van wie de doemscenario's precies afkomstig zijn. Techondernemers uit Silicon Valley, van wie velen de pauze-petitie ondertekenden, hebben een belang bij het groot maken van hun eigen technologie. Een tijdelijk verbod om nieuwe producten te lanceren zou bedrijven die (nog) niet bij de voorlopers horen enorm helpen. Ook veel wetenschappers, zeker in de Verenigde Staten, hebben banden met techbedrijven die verdienen aan AI-technologie. Daar staat tegenover dat drie van de meest prominente AI-wetenschappers ter wereld tot het meest pessimistische kamp behoren: Yoshua Bengio en Geoffrey Hinton, beiden winnaars van de prestigieuze Turing-prijs, en Berkeley-professor Stuart Russell.

→ **Lees ook**
'In films is de mens uiteindelijk altijd slimmer dan de slimste machines. Dat zal niet werken'

Sharon vindt bovendien dat de aandacht voor het existentiële risico „alle zuurstof wegneemt" van wat er nú al allemaal misgaat met AI. „Inaccurate systemen die veel fouten maken, waarvan kwetsbare groepen al slachtoffer zijn." Daarbij doelt ze bijvoorbeeld op de toeslagenaffaire, waarbij burgers door algoritmes onterecht als fraudeurs werden aangemerkt. En op het gebruik van gezichtsherkenningssoftware door de Amerikaanse politie, waarbij mensen onterecht in de gevangenis zijn beland.

Ook de 2.700 AI-wetenschappers die deelnamen aan de peiling maken zich, meer dan over existentieel risico, zorgen over acute risico's. Zoals de verspreiding van nepnieuws door AI. Of het gebruiken van kunstmatige intelligentie door overheden om hun bevolking te controleren, zoals in China gebeurt door de massale inzet van gezichtsherkenning. „Een van mijn grootste zorgen is hoe AI ons werk gaat veranderen", zegt Sharon. „AI neemt meer en meer taken over en de mens wordt gereduceerd tot het controleren of de AI zijn werk wel goed uitvoert. Dat is, vrees ik, voor veel mensen saai, vernederend en onbevredigend."

Dat een computer een 'zelfbewustzijn' ontwikkelt dat zich tegen mensen keert, is volgens verreweg de meeste AI-wetenschappers niet realistisch. Angst is er wel voor AI-systemen die door hun toenemende intelligentie onvoorspelbaar gedrag vertonen. De vrees is dat deze technologie door mensen verkeerd wordt gebruikt. Bijvoorbeeld door AI te gebruiken om cyberaanvallen uit te voeren of verkiezingen te beïnvloeden.

Het is de onvoorspelbaarheid die het juist zo eng maakt, vindt Joep Meindertsma. Hij begrijpt niet dat niet iedereen 5 procent kans op uitsterving als een gigantisch risico beschouwt. „Dat is toch hartstikke veel?"

Demonstranten

Een groepje demonstranten staat in een halve cirkel op de Wijnhaven, een plein in het centrum van Den Haag. Het is 11 augustus 2023. '*We fixed the ozone layer. So we can ensure safe AI*', staat op een van de protestborden. Een ander: '*Help us stop this suicide race to God-like AI*'.

Dit is „het grootste protest ter wereld voor AI-veiligheid", zegt Meindertsma, die het protest heeft georganiseerd, in een toespraak op het plein. „Het begin van een beweging."

Het is drie maanden na de oprichting en Pause.AI heeft enkele honderden leden. Op verschillende plekken, onder meer in San Francisco, Brussel, Londen en Melbourne, zijn op initiatief van Meindertsma door kleine groepjes protesten georganiseerd. De demonstratie vandaag in Den Haag heeft twaalf deelnemers. Vooral jonge techneuten, die veelal zelf bij techbedrijven werken of hebben gewerkt.

„Probeerde in de jaren tachtig maar eens iemand te overtuigen dat roken slecht was. Je moest je gedrag veranderen. Je leefstijl", zegt Meindertsma in zijn speech. „Het alternatief als roker was: je geloofde het gewoon niet."

Illustratie Fieke Ruitinga

Het is lastig om een boodschap met dergelijke grote implicaties over te brengen, wil hij maar zeggen. Voor Meindertsma is het moeilijk om zijn familie, zijn vrienden en zijn vriendin te overtuigen van iets dat misschien helemaal nooit gaat gebeuren. „We hebben genoeg aan ons hoofd. Het is fijn om weg te kijken", zegt hij. „Dat is nou eenmaal hoe het menselijk brein werkt." Maar in de loop der maanden heeft hij zich iets gerealiseerd. Hij hoeft de massa helemaal niet per se te overtuigen. Eén iemand, met echte invloed, is genoeg. Hij vestigt zijn hoop op de politiek.

Als hij contact legt met politici, valt hem op hoe makkelijk dat gaat. Hij stuurt tien e-mails naar politici en krijgt zeven keer een reactie. Hij mag langskomen. VVD-Kamerlid Queeny Rajkowski vraagt hem onder meer om input voor Kamervragen. Op uitnodiging van de Europese Commissie gaat Meindertsma naar Brussel, waar hij spreekt met politici die werken aan AI-wetgeving. „Politici willen graag leren, merk ik. Je hoeft geen professioneel lobbyist te zijn om hier een belangrijk duwtje in te kunnen geven."

Doordat het maatschappelijke debat kantelt naar de risico's van AI gebeurt er in de maanden daarna veel. Er zijn hoorzittingen in het Amerikaanse Congres met onder anderen OpenAI-topman Sam Altman. De AI Act, Europese wetgeving die dit jaar is aangenomen, legt AI-bedrijven verplichte veiligheidstesten op. En de Amerikaanse vicepresident Kamala Harris en de Britse premier Rishi Sunak uiten hun zorgen over existentiële risico's van AI.

Tot een pauze van de ontwikkeling komt het niet. Ook Nederland is daar niet voor. In de kabinetsvisie schrijft staatssecretaris Alexandra van Huffelen (Digitalisering, D66) in januari dat AI „ongelukken" kan veroorzaken en er een risico is dat „AI-systemen min of meer autonoom doelen gaan nastreven op een manier die schade berokkent". Maar een pauze van de technologie noemt ze in antwoord op Kamervragen „geen oplossing".

Als Meindertsma dit leest kan hij „zijn haren wel uit zijn kop trekken", zegt hij. „Die ongelukken zijn wél te voorkomen. Bouw die dingen niet." Voor politici is het „gewoon het zoveelste ding. Ze verliezen hun interesse. Willen er niks meer mee te maken hebben. Op naar het volgende project."

Helemaal niet gelukkig

Begin februari, op kantoor bij Ontola, vertelt Joep Meindertsma dat hij de laatste weken weer meer voor zijn bedrijf werkt. Het helpt hem om „wat meer zijn kop in het zand te steken. Het voelt gewoon fijner om een beetje afstand te hebben."

De rol van protestleider is hem zwaar gevallen. Hij geeft nog steeds interviews en organiseerde onlangs weer een protest, voor het kantoor van OpenAI in San Francisco. Hij geeft niet op, zegt hij, al hoopt hij dat anderen zijn rol op den duur gaan overnemen. „Ergens wil ik mensen inspireren om in actie te komen. Een

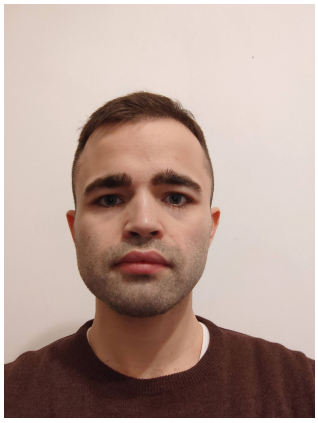
voorbeeldfunctie hebben", zegt hij. „Maar ik wil ook eerlijk zijn. Ik word hier helemaal niet gelukkig van."

Vrienden en familie maakten zich steeds meer zorgen om hem. Dit is niet jouw probleem om op te lossen, zeiden ze. Zijn vriendin raakte „echt gefrustreerd" dat hij zo veel tijd aan AI besteedde, zegt hij. Bijna elke avond videobellen, continu mensen over de vloer die de hele avond over AI praatten. „Een duister thema dat continu om ons heen was."

De zestig uur in de week die hij aan Pause.AI besteedde zijn er inmiddels twintig geworden. En nu hij wat meer afstand heeft genomen, voelt hij zich beter. „Ik geef mijn eigen comfort prioriteit boven dat van de rest van de wereld. Misschien is het wel beter als ik gewoon lekker ga genieten van wat er nog is."

IGOR IVANOV (30)

'In het begin was ik wanhopig. Maar alles went, ook dit'



Igor Ivanov (30)

Russisch, psycholoog, geeft therapie aan mensen met angst voor AI

'Mijn fascinatie voor AI begon in juni 2022. Er was een verhaal over een Google-medewerker die geloofde dat Googles AI-model LaMDA zelfbewustzijn had ontwikkeld. Ik zag het gesprek tussen hem en zijn AI en dacht: wow, dit model is echt heel slim. Het is zo

goed. Zo snel. Wat nu als het slimmer wordt dan mensen? Het verlamde me. Ik schreef aan mijn vrienden: ik denk dat dit gevaarlijk is.

„Ik heb een achtergrond in de klinische psychologie. Ik wilde iets bijdragen, maar dat bleek ingewikkeld. Ik ontdekte dat ik mensen kon helpen die worstelen met hun geestelijke gezondheid door AI. Door ze weer productief te laten zijn, en minder te laten lijden.

„Ik heb nu zes cliënten. Ze zijn allemaal rond de dertig jaar, heel slim, uit verschillende landen. Ze maken zich zorgen over existentieel risico, werken allemaal in AI. Een van hen is bang dat zijn kinderen niet lang meer leven. Een ander voelt zich schuldig dat hij niet genoeg bijdraagt aan AI-veiligheid.

„Sommigen van hen komen niet meer buiten, verzorgen zichzelf niet. Waarom zou je je tanden poetsen als je nog maar een paar jaar hebt? Waarom zou ik leven? Waarom zou ik mijn kinderen grootbrengen? Waarom zou ik mijn werk nog doen als computers duizend keer productiever zullen zijn dan ik? Het is vergelijkbaar met mensen die een terminale ziekte hebben.

„Zelf ben ik ook pessimistisch over onze kansen om dit te overleven. Mensen zijn veel slimmer dan apen, daarom hebben wij ze onder controle, ondanks dat apen sterker zijn. Wat zegt mij dat computers die slimmer zijn dan wij niet hetzelfde met ons gaan doen? We weten het niet.

„In het begin was ik extreem gespannen. Wanhopig. Maar alles went, ook dit. En ik kreeg het gevoel: ik wil iets goeds doen, zolang ik nog leef. Het heeft me wel veranderd. Ik investeer geen geld bijvoorbeeld, waarom zou ik? Sommigen van mijn vrienden begrijpen het, anderen niet. Ik kan emotioneel worden als ik erover praat. Soms raken mensen geïrriteerd. En ik wil mijn vrienden niet ergeren.

„Ik ben pas dertig, maar maak de balans al op. Welke fouten heb ik gemaakt in mijn leven? Wat heb ik goed gedaan? En hoe kan ik een betekenisvol leven leiden in de tijd die ik nog heb?"

Stijn Bronzwaer

SARAH HASTINGS-WOODHOUSE (25)

'Ik vrees het meest voor het punt dat we geen controle meer hebben'

Sarah Hastings-Woodhouse (25)

Brits, begon een podcast over AI: *Consistently Candid*

'Ik ben me in AI gaan verdiepen door mijn werk. Ik vroeg me af: hoe lang gaat het nog duren voordat mijn werk [als studieadviseur] volledig geautomatiseerd kan worden? Het begon als angst voor baanverlies. Maar hoe dieper ik erin dook, hoe meer veiligheidsrisico's ik tegenkwam. Het echte keerpunt was vorig jaar mei, toen Geoffrey Hinton [AI-wetenschapper] opstapte bij Google om zich uit te kunnen spreken over de gevaren van AI. Ik dacht: als zo'n slim iemand dit doet, dan moet ik me echt zorgen gaan maken.

„Ik ben onderzoek gaan doen. Ging op zoek naar optimisme, naar geruststelling, maar ik kreeg alleen maar meer het gevoel dat niemand die met AI bezig is echt weet wat hij doet. Ik vrees het meest voor het punt dat we geen controle meer hebben. De

onwetendheid en onzekerheid over wanneer dit precies gebeurt maakt me angstig.

„Mijn vrienden en familie steunen me, maar zijn er niet op dezelfde manier mee bezig. Via X kwam ik terecht in een wereld met mensen zoals ik. Lotgenoten. De gemeenschap is heel open en warm, ook voor mensen die niet veel weten van de technische kant van AI.

„Op sociale media heb je de *slow downers* versus de *accelerationists*. Ik ben een extreme slow downer: wat mij betreft moet AI volledig op pauze worden gezet. Tot we oplossingen hebben voor de veiligheidsrisico's. „Mijn grote frustratie: veel mensen geven toe dat er een groot risico kleeft aan AI, maar doen er verder niks mee. Ze gaan gewoon door met leven. De gedachte is: dit repareert iemand anders wel. Het voelt alsof ik in een totaal andere wereld leef. Dit voelt voor mij als een gigantische, urgente dreiging die moet worden bestreden, terwijl andere mensen hier niet eens over nadenken. Dat is superfrustrerend.

„Ik ben met een podcast begonnen om onderwerpen als AI-veiligheid op een toegankelijke manier te bespreken. Ik hoop bewustzijn te creëren. Het voelt goed om er niet alleen in te staan, want deze strijd kan best eenzaam en isolerend zijn."

Juliette de Waal

Correctie (17 februari 2024): In eerdere versie stond dat de p(doom) door wetenschappers wordt ingeschat op 5 procent. Dat moet zijn minimaal 5 procent en is hierboven aangepast.

Een versie van dit artikel verscheen ook in de krant van 17 februari 2024.