



Het gevaar van ChatGPT

De hedendaagse sofist

Met ChatGPT lijkt de doorbraak van artificiële intelligentie een feit. Net als met de sofisten in het oude Griekenland roept deze nieuwe creatieve metgezel fundamentele vragen op. Zal de digitale schaduwwereld ons opslokken? **Wessel Reijers beeld Milo**

In de vijfde eeuw voor Christus werd het oude Griekenland opgeschud door de opkomst van een nieuwe groep 'wijze mannen', de sofisten. Zij boden het grote publiek toegang tot allerlei nieuwe inzichten, verhalen en wijsheden, vaak tegen betaling. Sommigen van hen beweerden zelfs dat ze de antwoorden kenden op alle vragen.

Hun belofte was niet zonder verdienste. Men werd inderdaad vaak wijzer, en kennis verspreidde zich onder de bevolking. Maar tegelijkertijd verspreidde zich een hele hoop onzin. Dit kwam doordat de sofisten voornamelijk waren getraind in de retorica, de taalkundige techniek ter vergroting van de overredingskracht. Het doel was om de luisteraar te overreden, niet primair om de waarheid te verkondigen.

Hetzelfde, maar dan op steroïden, lijkt nu te gebeuren met ChatGPT. Het AI-programma kan zowel briljante (menselijke) inzichten in een nieuw formaat presenteren als 'vloeiende bullshit' uitkramen.

'GPT' staat voor 'Generative Pre-trained Performer' en is ontwikkeld door het Amerikaanse bedrijf OpenAI. Het is in staat om een nieuwe tekst te genereren gebaseerd op een tekstopdracht als: 'Schrijf een opstel over de theorie van het nationalisme'. Dit biedt ongelooflijk veel mogelijkheden, want ChatGPT kan helpen software te schrijven, geloofwaardige academische essays te produceren en een screenplay voor de volgende sitcom op Netflix in elkaar te zetten. De ontwikkeling gaat razendsnel, en onlangs is een nieuwe, nog krachtigere versie al gelanceerd door OpenAI: GPT-4.

Net als tijdens een WK lijkt nu iedereen op het internet een AI-fan, een nieuwe AI-expert of geheelonthouder. Jesse Frederik schrijft over de reacties van zijn collega's bij *De Correspondent*: 'Sommigen blijven liever holbewoner, terwijl anderen juist het vuur herkennen. Word wakker mensen, het is tijd voor de bronstijd!'

We kunnen, zoals Frederik voorstelt, ChatGPT als een 'liefdevolle metgezel' zien. De kunst is om de bullshit te cureren, het goede over te nemen en het slechte te laten liggen. Zo zullen sommigen in het oude Griekenland de sofisten ook hebben beschouwd. 'Take it or leave it', je kunt zelf bepalen wat je aanneemt van die oude man die staat te oreren op het plein. En zo zullen er ook mensen zijn geweest – de holbewoners – die er schouderophalend langs waren gelopen.

Een groepje filosofen onder leiding van Socrates zag dit anders: zij namen de sofisten serieus, maar herkenden ze ook als een ernstig gevaar. Taal is geen neutraal gereedschap maar kan de luisteraar betoveren, meeslepen naar een wereld van schijn en bedrog.

Net als de socratische filosofen moeten we voorzichtig zijn met de taalmagie van AI. Als programma's als ChatGPT het internet overnemen, wat gezien de explosie in gebruikers niet onwaarschijnlijk is, komen we werkelijk te leven in een digitale schaduwwereld. Die zal ons ongetwijfeld veel moois brengen en veel dingen een stuk makkelijker maken, maar zal ons ook betoveren. >

Ondanks het nu nog beperkte gebruik van ChatGPT zijn er al zorgen over de rol van chatbots in het vergroten van het probleem van waarheidsvinding. Zo lijkt het alsof het programma academische referenties kan toevoegen aan een tekst, maar blijken die vaak compleet verzonnen. Op eenzelfde manier kan ChatGPT neppe nieuwswebsites en nieuwsberichten produceren, wat de verspreiding van desinformatie in de hand zal werken. Natuurlijk kunnen we eindeloos gaan factchecken, maar met de grote hoeveelheid nieuwe tekst die wordt geproduceerd lijkt dat dweilen met de kraan open.

Om het gevaar van de sofisten in te dammen kwamen de socratische filosofen op de propen met een nieuwe methode voor waarheidsvinding, waar de logica uit voortkwam. Plato, een leerling van Socrates, stelde de waarheidsvraag in zijn bekende allegorie van de grot. De sofisten, zo beweerde hij, projecteren slechts schaduwen op de muur, weerkaatsingen en vervormingen van een schijnwerkelijkheid. De echte waarheid vinden we enkel wanneer we de grot verlaten, en de wereld in het zonlicht aanschouwen.

Net als in het oude Griekenland kunnen we de betovering van AI alleen trotseren door weer filosofische vragen te stellen. Maar hoe vinden we verlichting in de schaduwwereld van ChatGPT? Om hier antwoorden op te vinden moeten we voorbij gaan aan vragen over de waarheid en ons richten op de vraag: wat is nieuw?

Wanneer is iets nieuw of origineel? Intuïtief zeggen we: als het er niet eerder was. Ieder mens is bijvoorbeeld uniek, en iedere geboorte is de start van iets compleet nieuws. Zo zien we vaak ook nieuwe ontwikkelingen in de kunst. Mondriaan maakte volstrekt nieuwe kunst door te breken met wat vooraf kwam en een opening te creëren naar een nieuwe kunstervaring.

Pas kreeg voor het eerst een 'AI-kunstenaar' een prijs met een digitaal kunstwerk dat grotendeels gemaakt is door de AI-applicatie Midjourney. Spelontwikkelaar Jason Allen verbeterde eindeloos een tekstinpuut tot er een – volgens internet – prachtig digitaal kunstwerk uit kwam. Het stuk, met de hoogdravende titel *Théâtre d'Opéra Spatial*, sleept de kijker mee naar een futuristische mix van Gustav Klimt, een Japans samurai-kasteel en *Star Wars*. Enkele vage, sierlijk geklede figuren kijken door een raam uit over een stedelijke en sneeuwrijke woestenij.

Dit digitale theaterstuk leverde Allen de eerste prijs op in een competitie voor beeldende kunst in de Amerikaanse staat Colorado, maar ook een kanonnade van reacties op Twitter. Velen waren boos over de toekenning van de prijs, en vergeleken het met een marathon winnen in een Lamborghini. De boosheid kwam vooral voort uit de overtuiging dat Allen zijn talent als kunstenaar niet had ingezet om iets origineels te maken. Hij had simpelweg een opdracht gegeven.

Kunstmatige intelligentie daagt onze gevestigde ideeën uit over originele poëzie, kunst en literatuur. ChatGPT lijkt in staat compleet



Of het nu woorden zijn of pixels, generatieve AI-systemen zijn getraind om voorspellingen te doen

nieuwe gedichten en verhalen op papier te zetten. DALL-E produceert digitale afbeeldingen, zoals nieuwe schilderijen in de stijl van Van Gogh of Mondriaan.

Hoewel teksten en afbeeldingen verschillen, werken ChatGPT en DALL-E volgens dezelfde principes. Of het nu woorden zijn of pixels, generatieve AI-systemen zijn getraind om voorspellingen te doen. ChatGPT voorspelt welk woord na een voorafgaand woord 'redelijkerwijs' moet verschijnen. DALL-E voorspelt welke configuratie van een digitaal plaatje past binnen een statistisch patroon.

Het verschil tussen mensen en machines zit 'm in de manier waarop ze zich tot de toekomst verhouden. AI is afhankelijk van enorme hoeveelheden data: miljarden woorden uit boeken, artikelen en het internet, en plaatjes van mensen, katten en auto's. Elementen uit deze dataset worden aan elkaar gerelateerd door waarschijnlijkheid. Aan het eind van de zin 'de premier van Nederland is' vind je eerder 'Mark Rutte' dan 'hagelslag'.

In *The New Yorker* schreef journalist Ted Chiang dat ChatGPT stukjes tekst op het internet parafraseert, terwijl je op Google juist citaten vindt. Bestaande tekst wordt als het ware samengebonden tot iets makkelijker verteerbaars, zoals een JPEG opzettelijk bepaalde informatie weglaat om een digitaal plaatje samen te drukken in een kleiner dataformaat. ChatGPT doet hetzelfde (en meer) als een JPEG, maar dan met het hele internet. Doordat informatie niet wordt gekopieerd maar opzettelijk wordt weggelaten en geremixt, ontstaat er de indruk van iets nieuws.

Maar is dit dezelfde 'nieuwheid' die mensen ervaren en produceren? Was Mondriaan niet ook gewoon een heel goede 'remixer'? Nam hij niet juist elementen op uit de bestaande kunst, om daaruit iets nieuws te creëren? Sommigen zullen door het zuur bijten en dit simpelweg beamen. Net als generatieve AI-systemen zijn mensen niets anders dan probabilistische machines, die bestaande informatie remixen tot iets nieuws.

Filosofen als Hannah Arendt hebben dit punt echter bestreden. Een mens anticipeert volgens haar niet zoals een machine voorspelt. Wij ervaren de tijd niet als een compressie of remix van het recente verleden. Mensen hebben een levend verleden, in herinnering, en toekomst, in anticipatie. Dat stelt ons in staat om te verrassen, om spontaan iets nieuws te doen.

AI-systemen zijn daadwerkelijk creatief, zoals blijkt uit het kunstwerk van Allen. Er is nu blijkbaar zelfs een vorm van creativiteit waarin deze systemen beter worden dan mensen. Deze vorm van creativiteit is alleen inherent gelimiteerd, omdat die berust op statistische voorspelbaarheid, en omvat niet alles wat origineel is aan menselijk handelen en menselijke expressie.

ChatGPT omringt ons dus met een nieuwe, begrensde vorm van creativiteit. Dat kun je jammer vinden, maar is het ook gevaarlijk? Op deze vraag is geen simpel antwoord te geven. De concrete gevolgen zijn soms makkelijk aan te wijzen – zoals wanneer de chatbot een verkeerd antwoord geeft – maar het gaat hier om een vage tendens.

Wat duidelijk is bij iedere technologie, en dit schreef Plato al over de uitvinding van het schrift, is dat we te maken hebben met een *pharmakon*, wat zowel geneesmiddel als vergif betekent. Het is duidelijk wat de mogelijk fantastische toepassingen zijn van deze tool, maar de keerzijde (het vergif) blijft tot nog toe grotendeels onderbelicht.

Velen zullen de tendens omarmen als een geneesmiddel en AI zien als een sleutel tot een cybernetische utopie. Als je gelooft dat alles wat wij mensen maken in zekere zin al een remix is, zal ChatGPT enkel een versnelling bewerkstelligen. Deze versnelling wordt bejubeld door zowel links als rechts, van ideeën over een vol geautomatiseerd luxe communisme tot het transhumanisme dat halsreikend uitkijkt naar de singulariteit.

Daarnaast is er het geruststellende optimisme van Jesse Frederik en Ian Bogost, die aangeven dat ChatGPT dom genoeg is om ongevaarlijk te zijn – geneesmiddel noch vergif. Bogost noemt in *The Atlantic* als reden dat ChatGPT geen epistemisch systeem is (gericht op kennis vergaren) maar een esthetisch systeem (gericht op creatieve expressie). Het programma *speelt* met tekst in plaats van dat het tekst begrijpt, en komt daarom vaak op de proppen met leuke maar incorrecte creatieve probeersels, of – als het in een keurslijf wordt gedrukt – met saaie, standaard middelbare-schoolessays.

Maar zijn esthetische systemen wel zo ongevaarlijk? Dit brengt ons terug bij de kritiek van de socratische filosofen. Net als bij de sofisten is het gevaar van ChatGPT dat het ons betovert en daarmee de diversiteit van het menselijk leven overwoekert met een techniek. Deze betovering kan erg gevaarlijk zijn, niet alleen omdat het ons kan misleiden maar ook omdat we ons 'zelf' erin kwijt kunnen raken.

ChatGPT opereert als een steeds verder uitdijende feedbackloop. Deze loop zal zich voorbij de kunst bewegen, omdat de voorspellingskracht van een programma als ChatGPT op talloze wijzen kan worden ingezet. Zo heeft een rechter in Colombia de chatbot onlangs al gebruikt bij het opstellen van een rechterlijk oordeel.

Het gevaar zit 'm niet zozeer in afzonderlijke toepassingen, maar in de tendens om aspecten van ons leven te overwoekeren. We kennen dit in praktische zin misschien van de rol van digitale technologie in ons dagelijks leven. Natuurlijk kun je nog een papieren kaart gebruiken om de weg te vinden, maar deze capaciteit wordt overwoekerd door Google Maps. Natuurlijk kun je nog spontaan je grote liefde tegenkomen in het alledaagse leven, maar deze mogelijkheid wordt overwoekerd door datingapps als Tinder. Paradoxaal genoeg nemen cybernetische systemen onze individuele vrijheid niet af, maar confronteren ze ons met een tendens die ons collectief beperkt.

Hoe kunnen we de betovering van ChatGPT trotseren? Alleen al het stellen van deze vraag klinkt wat conservatief en hopeloos nostalgisch. Technologie, zo geloven we nu en masse, heeft een eigen logica ('het komt er toch wel') en heeft, ondanks de onbedoelde bijwerkingen, over het algemeen een positieve impact.

Toch moeten we ons best doen om het gevaar van ChatGPT in toom te houden en de nieuwe sofisten niet te veel op het podium te hijsen. Filosofische vragen stellen is zeker het begin, maar

Net als om biologische groenten en fruit vragen we wellicht binnenkort om 'AI-vrije' schrijfsels

niet het einde van een oplossing. In tegenstelling tot de sofisten vergt het temmen van een technologie meer dan argumenten. Zonder wetten, technologische aanpassingen en instituties is geen enkel argument opgewassen tegen de aantrekkingskracht van AI.

Hoewel dit enkel voorzichtige suggesties zijn, denk ik dat er ten minste drie wegen leiden naar 'verlichting'. De eerste gaat over onze houding tegenover technologieën als ChatGPT. We moeten beseffen dat we niet machteloos staan, maar daadwerkelijk in staat zijn technologische vooruitgang een halt toe te roepen. De overheid speelt

hierin een sleutelrol, door technologische ontwikkeling niet louter aan te jagen maar ook in te dammen.

Dat dit succesvol kan zijn leert de geschiedenis. Ondanks het enorme gevaar van kernwapens zijn er wereldwijd robuuste afspraken gemaakt over de regulering van technologieën voor kernsplitsing, via verdragen en een uniek instituut als het Internationale Atoomagentschap. Ook technieken voor het klonen van mensen worden, hoewel het in principe mogelijk is, wereldwijd gelimiteerd. AI-systemen hoeven geen uitzondering te zijn. Als we het echt willen, kunnen we deze systemen effectieve beperkingen opleggen, iets wat de Europese Unie nu voorzichtig probeert met de opkomende 'AI Act'.

Een tweede weg gaat over de noodzaak om menselijke expressie te bevoordelen ten opzichte van de productie van kunstmatige intelligentie. AI is het maar om recht te doen aan de echte oorsprong van creatieve expressie, namelijk de duizenden artiesten die zich nu terecht bedreigd voelen. Het volstaat niet om disclaimers bij te voegen die zelf ook de schijn tegen zich hebben. Net als biologische groenten en fruit, vrij van pesticiden, vragen we wellicht binnenkort om 'AI-vrije' schrijfsels.

Tot nog toe (maar hoe lang nog?) kan dit door AI met AI te bestrijden. OpenAI heeft een programma uitgebracht dat kan herkennen of ChatGPT een tekst heeft geschreven. Maar naast dweilen met de kraan open is dit ook het alleen maar vergroten van de feedbackloop, met alle gevolgen voor het klimaat van dien (AI-systemen gebruiken enorme hoeveelheden stroom). We kunnen ook proberen om ons waardenpatroon zelf aan te passen: om minder waarde te hechten aan alles wat digitaal is en meer waarde toe te kennen aan ons materiële bestaan. Op zoek te gaan naar een echt schilderij, een live concert of een poëzie-slam.

Ten derde moeten we wellicht op zoek naar wat de techniekfilosoof Yuk Hui beschrijft als educatie van de sensibiliteit. Hij stelt in zijn nieuwste boek *Art and Cosmotechnics* dat de verschillende vormen van kunst, zoals we die vinden in de Griekse tragedie en de Chinese landschapskunst, ons iets kunnen vertellen over onze verhouding tot de technologie. We moeten voorbij gaan aan het idee dat kunst enkel representatie is, en opnieuw de vraag stellen hoe kunst een originele toegang biedt tot de wereld om ons heen.

Deze laatste weg vergt wellicht de meeste moeite, niet in het minst omdat onze wereld zo verzadigd is met moderne technologie. Om Mark Fisher te parafraseren: het lijkt gemakkelijker om ons het einde van de wereld voor te stellen dan het einde van de moderne technologie. Dat neemt niet weg dat we het moeten proberen, omdat de digitale schaduwwereld ons anders onherroepelijk zal opslokken. ■

Wessel Reijer is postdoctoraal onderzoeker aan de Universiteit van Wenen, gespecialiseerd in de filosofie en ethiek van technologie