

De onttovering van AI

De onttovering van AI

*Een pleidooi voor het gebruik
van gezond verstand*

Ronald Meester en Marc Jacobs



MAZIREL
PERS

Ontwerp omslag: ST. DUO, Rotterdam
Ontwerp binnenwerk: Crius Group, Hulshout

ISBN 978 94 6456 341 2
e-ISBN 978 94 6456 342 9
NUR 740

Mazirel Pers is een imprint van Walburg Pers

© 2024 R. Meester & M. Jacobs, p/a Uitgeversmaatschappij Walburg Pers,
Zutphen

© 2024 Uitgeversmaatschappij Walburg Pers, Zutphen

www.walburgpers.nl

Alle rechten voorbehouden. Niets uit deze uitgave mag worden veele-
voudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar
gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch,
door fotokopieën, opnamen of enige andere manier, zonder voorafgaande
schriftelijke toestemming van de uitgever.

Marc draagt dit boek op aan
John, Anke, Hanneke en Frank

Ronald draagt het op aan zijn gezin

Inhoud

Deel I AI en Techniek

- | | |
|-------------------------------------|----|
| 1. De liefdesbrief | 11 |
| 2. Een klein overzicht van AI | 35 |
| 3. Hoe werkt het nou eigenlijk? | 55 |
| 4. Over bewustzijn en reductionisme | 79 |

Deel II AI en de mens

- | | |
|---|-----|
| 5. Techniek dringt zich op, en AI al helemaal | 107 |
| 6. Kennis, vaardigheden en overtuiging | 129 |
| 7. Verantwoordelijkheid en regulatie | 149 |

Deel III AI in de praktijk

- | | |
|---|-----|
| 8. Een verstandige en gebalanceerde omgang met AI | 169 |
| 9. Een wereld vol AI | 187 |
| 10. Een overweging over AI | 215 |

- | | |
|-------|-----|
| Noten | 223 |
|-------|-----|

Deel I

AI en Techniek

1. De liefdesbrief

Het is geen eenvoudige aangelegenheid om kritisch te schrijven over de razendsnelle ontwikkelingen op het gebied van kunstmatige intelligentie (AI¹) zonder in allerlei clichés te vervallen. Elke ontwikkeling kent haar eigen onheilsprofeten, cultuurpessimisten, doemdenkers en vroeger-was-alles-beter-verzuchters. En natuurlijk roept ook elke ontwikkeling een groep adepten in het leven, die ons probeert te vertellen hoe fantastisch deze ontwikkeling wel niet is, en hoe we daar met z'n allen zeker van zullen gaan profiteren.

Voor sommigen, zoals bijvoorbeeld historicus en futuroloog Yuval Harari, is een ontwikkeling zoals AI een voorbode van een al dan niet angstaanjagende tijd waarin de mensheid de controle over de wereld uit handen gaat geven ten gunste van autonome digitale entiteiten. Anderen, zoals bijvoorbeeld ondernemer Elon Musk, zijn vooral bezig met likkebaardend onderzoeken hoe deze AI hun *business* kan versterken en uitbreiden. Er ligt namelijk misschien wel (heel) veel geld en controle in het verschiet.

Wij zien de ontwikkeling van AI vooral als een nieuwe stap in de 'vermodelisering' en technocratisering van de samenleving. We hebben al eerder kritisch geschreven over de rol die wiskundige modellen volgens ons moeten hebben,² maar de huidige ontwikkelingen roepen om hun eigen, diepgaande duiding. Wij zien het gebruik van AI-modellen óók als een technische of technocratische ontwikkeling. Deze technische ontwikkeling staat natuurlijk niet op zichzelf, en we zullen dan ook een wat hoger gezichtspunt innemen, en laten zien dat wat er nu

aan de hand is parallellen vertoont met bijvoorbeeld de industriële revolutie en de verregerende mechanisering die daarna heeft plaatsgevonden.

Natuurlijk is digitalisering niet hetzelfde als mechanisering. Daar waar mechanisering in eerste instantie vooral fysiek werk uit handen nam, en daarna de computer vooral een hulp werd bij rekenwerk en organisatie, lijkt moderne AI zich op een ander terrein te profileren. Je vraagt een taalprogramma als ChatGPT³ iets, en je krijgt antwoord. Je vraagt ChatGPT om een samenvatting van een bepaald boek, en die rolt er in een oogwenk uit. Je vraagt ChatGPT naar een bewijs voor de stelling van Pythagoras, en het systeem⁴ komt onmiddellijk met diverse bewijzen. Je vraagt het systeem om een liefdesbrief aan je geliefde te schrijven en opeens ben je een dichter. Je vraagt ChatGPT om een stuk code in, zeg, Python te schrijven en je hoeft het antwoord alleen nog maar even te controleren. Je vraagt ChatGPT om een vertaling te maken en het is zo gebeurd. Je kunt ChatGPT ook vragen naar bijvoorbeeld alle boeken die een auteur heeft geschreven, of naar de onbekende dichter van een bijzondere regel die je ergens hebt onthouden. Hoewel het systeem (veel) fouten maakt, gaat er ook heel veel goed. Je kunt zoveel verschillende dingen doen dat het soms bijna wel magie lijkt. Het is dan ook niet verwonderlijk dat veel mensen systemen als deze allerlei menselijke eigenschappen zoals bewustzijn en bepaalde emoties toedichten; daarover later (veel) meer.

Wij gaan niet onmiddellijk roepen dat AI het begin van het einde is, en we gaan ook niet roepen dat het onmiddellijk verboden zou moeten worden. We omarmen het echter ook niet helemaal. We zijn, kort door de bocht, niet zozeer begaan met de vraag of de AI-modellen ons zullen

gaan overnemen, maar wel met de vraag wat wij bereid zijn zélf op te geven in onze omgang met AI. De vragen die opdoemen zijn *levensbeschouwelijke* vragen naar wat mensen de moeite waard vinden, vragen over in hoeverre mensen bereid zijn verantwoordelijkheid te nemen voor hun beslissingen, daden en gedachtegoed, en *last but not least* vragen over de omgang met modellen en techniek en het vermeende gemak ervan.

Wat maakt dat we het juist hierover willen hebben? Dat een hijskraan meer kan tillen dan een mens vinden we doorgaans niet erg; dat een rekenmachine veel sneller kan rekenen dan wijzelf ook niet echt. Zelfs een schaakcomputer die de wereldkampioen verslaat accepteren we, met als argument dat het nu eenmaal zo is dat schaken een mathematisch object is met vaste regels, en dat een computer gewoonweg duizelingwekkend veel sneller kan rekenen dan wijzelf. Computers doen dingen die we niet zelf zouden kunnen, althans niet in de tijd die ons gegeven is. We kunnen ons tevreden stellen door te denken dat we – met voldoende tijd – al die berekeningen in principe ook zelf hadden kunnen maken. Er gebeurt, met andere woorden, niets principieel ‘vreemds’ als een computer voor ons rekent. Er is hierbij overduidelijk geen sprake van magie, maar slechts van brute rekenkracht. Als je dieper rekent kun je bij schaken betere zetten vinden; een rekenpartij die we in principe zelf ook hadden kunnen uitvoeren, ware het niet dat dit door de enorme omvang ervan volkomen onmogelijk is om binnen redelijke tijd te doen.

Met AI lijkt nu echter een ander sentiment te zijn ontstaan. Als de eerste tekenen niet bedriegen, zijn we nu ook bereid om ons door AI te laten leiden bij het adresseren van allerlei andersoortige problemen. Als we

ChatGPT iets vragen, komt het programma terug met een *talig* antwoord. Dat kan een antwoord zijn dat op zichzelf weer nieuwe vragen oproept, en zo kun je een ‘gesprek’ gaan voeren met het algoritme. En dan wordt het interessant, want er lijkt een groot verschil te bestaan tussen een uitkomst van een grote berekening enerzijds en een antwoord in taalvorm en zelfs een heel gesprek anderzijds. Rekenen is onpersoonlijk, maar taal is dat zeer zeker niet. Dat geeft dan al snel de illusie dat we bij interactie met een programma als ChatGPT ook daadwerkelijk aan het communiceren zijn met een bewuste entiteit, die begrijpt wat we willen vragen en bij twijfel om opheldering vraagt. Net als een mens dat zou doen.

Dit is, en we kunnen dat niet vroeg genoeg benadrukken, een enorme illusie. Een taal spreken enerzijds en communiceren anderzijds zijn echt heel verschillende dingen. We zullen in de volgende hoofdstukken uitleggen hoe ChatGPT en vergelijkbare programma’s feitelijk werken, en dan zal de lezer begrijpen dat ChatGPT óók niets anders doet dan rekenen, op basis van informatie die wij het programma zelf hebben aangeleverd. ChatGPT voert absoluut niets uit waarvan wij zouden zeggen dat het ‘intelligent’ of ‘slim’ is; je kunt stukken taal construeren op basis van heel veel voorbeelden, heel veel rekenkracht, en slimme kansrekening en statistiek.

We beweren daarbij trouwens nadrukkelijk niet dat taal hierdoor gereduceerd zou worden tot kansrekening en statistiek, al wordt die gedachte hier en daar wel geopperd. Zo beweert wiskundige en AI-specialist Stephen Wolfram, de ontwerper van het softwarepakket Mathematica[®], dat we door ChatGPT juist veel leren over hoe taal in elkaar zit: ChatGPT zou ons linguïstische wetten doen ontdekken.⁵ Daar heeft hij, zo denken wij,

gedeeltelijk wel gelijk in. Geautomatiseerde data-analyse kan verbanden ontdekken waar wij dat als gewone mensen nooit voor elkaar hadden gekregen, en tot op zekere hoogte zal dat voor taal ook wel gelden. Maar dat wil natuurlijk helemaal niet zeggen dat taal gereduceerd zou kunnen worden tot dergelijke statistische wetten. Als je op een statistische manier naar taal gaat kijken, dan kun je statistische wetmatigheden ontdekken, maar dat is vooral het gevolg van de benadering die je zelf hebt gekozen. Het is nogal een vreemde conclusie om dan te denken dat taal door een dergelijke statistische analyse helemaal kan worden begrepen. Toch beweert Stephen Wolfram dat ChatGPT de ‘essentie van ons denken’ blootlegt.⁶ Dat is nogal een uitspraak.

Laten we dat punt nu eerst wat verder uitwerken, en beargumenteren dat Wolfram hierin onmogelijk gelijk kan hebben. ChatGPT werkt volgens een wel gedefinieerd wiskundig model, maar in plaats van getallen geeft het in woorden antwoord op vragen die je het programma stelt. En dan komen we dus, zoals we al schreven, in interessant vaarwater. We kunnen makkelijk volhouden (en verdedigen) dat onze identiteit niet in onze spierkracht ligt of in ons vermogen om berekeningen snel en foutloos uit te voeren. De manier waarop wij taal gebruiken om ons uit te drukken lijkt veel dichter te liggen bij ons wezen en onze identiteit: iedereen gebruikt taal op een andere manier. Als je ChatGPT nu vraagt naar de relatie tussen denken en taal, antwoordt het programma als volgt:

Er is een nauwe relatie tussen taal en denken. Taal is niet alleen een middel om informatie over te dragen, maar heeft ook een diepgaande invloed op ons denken en ons

vermogen om te redeneren. Een van de belangrijkste punten is dat taal ons in staat stelt om abstracte concepten te begrijpen en te communiceren. Door middel van taal kunnen we bijvoorbeeld praten over dingen die niet fysiek aanwezig zijn, zoals ideeën, emoties en concepten. Taal geeft ons ook de mogelijkheid om logisch te redeneren en complexe gedachten te formuleren. Daarnaast beïnvloedt taal ons vermogen om te leren en kennis te verwerven. We leren bijvoorbeeld nieuwe woorden en concepten door middel van taal en kunnen deze vervolgens gebruiken om ons begrip van de wereld te vergroten. Het is belangrijk om op te merken dat verschillende talen verschillende woorden en grammaticale structuren hebben, wat kan leiden tot verschillende manieren van denken en waarnemen van de wereld. Dit heeft belangrijke implicaties voor ons begrip van de realiteit en de vraag of we ooit een objectieve kijk op de wereld kunnen hebben als ons begrip ervan zo sterk wordt beïnvloed door onze taal en cultuur.

Je hoort in dit antwoord veel terug van de late Ludwig Wittgenstein. Volgens ChatGPT stelt taal ons in staat om bepaalde dingen te begrijpen, maar of dat begrip daarna ook weer in taal kan worden uitgedrukt, daarover zegt ChatGPT niets. Taal begrijpen betekent niet alleen maar dat je moet kijken naar de vraag of iets waar of onwaar, logisch of onlogisch is, maar ook of het gesproken woord succesvol is wat betreft intentie. Dat was het inzicht van bijvoorbeeld taalfilosoof John Austin in de jaren 50 en 60 van de vorige eeuw.⁷ Maar voor intentie is een *mind* nodig, een bepaald soort bewustzijn. Bovendien is het volgens diezelfde Austin ook nog eens zo dat taal niet alleen over de wereld gaat, maar ook over het zelf dat in die wereld acteert: als je iets zegt, dan doe je ook iets. Taal

acteert eigenlijk als een model dat we gebruiken om de wereld begrijpelijker te maken. Taal is uniek in de zin dat iedere persoon zijn of haar eigen manier heeft om zich uit te drukken. Taal creëert daarmee ook iets van werkelijkheid. Daarmee sluiten we aan bij onze algemene visie op modellen waarin deze niet alleen iets over de wereld zeggen, maar ook hun eigen werkelijkheid creëren.⁸

Het is absoluut zo dat taal een middel is om ons begrip van de wereld te vergroten, maar dat wil nog niet zeggen dat begrip hiermee samenvalt. Er zijn verschillende manieren om 'om begrip heen te cirkelen', en niet elke taal is daar in elke omstandigheid even goed in; de culturele context is hierbij van groot belang.

Zoals elk model moet dus ook ChatGPT geïnterpreteerd worden, en dat kan alleen met bewustzijn. Er is niet alleen een zender maar ook een ontvanger. ChatGPT zelf begrijpt niets. Misschien is een analogie met de wiskunde verhelderend. In het begin van de vorige eeuw dachten de wiskundigen Bertrand Russell en Alfred North Whitehead dat wiskunde puur symbolisch was, en dat er geen betekenis aan hoefde te worden toegekend aan symbolen en operaties binnen de wiskunde. (Niet geheel toevallig was Russell de mentor van Wittgenstein, en heeft Russell laatstgenoemde waarschijnlijk in eerste instantie de verkeerde kant op gestuurd door taal aanvankelijk als pure logica op te vatten.) Die opvatting bleek onhoudbaar, omdat wiskunde zich zonder betekenis ook niet kan ontwikkelen. Wiskunde is niet alleen maar pure logica, laat staan dat taal dat zou zijn.

Begrip is niet te conceptualiseren en daarmee valt het automatisch buiten het bereik van ChatGPT en haar eventuele opvolgers. Wat bedoelen we daar precies mee? Welnu, als docenten en adviseurs zijn we natuurlijk

gewend om behoorlijk ingewikkelde en moeilijke dingen uit te leggen of uitgelegd te krijgen. We kunnen studenten introduceren in theorie, we kunnen voorbeelden geven, we kunnen uitleggen hoe we er zelf tegen aankijken, we kunnen ze op alle mogelijke manieren proberen te helpen. We kunnen studenten begeleiden bij hun zoektocht naar begrip, maar we kunnen ze niet iets laten begrijpen. Het laatste stukje van een reis naar begrip zal iedereen zelf, in alle eenzaamheid moeten afleggen. De kern is (dus) dat begrip nooit het gevolg is van een algoritmische activiteit. Begrip is emergent. Dat woord is afgeleid van het Engelse 'to emerge', en betekent zoiets als 'zomaar verschijnen'. In onze context verwijst het naar een bepaald fenomeen dat ontstaat uit andere fenomenen, maar niet in termen van die oorspronkelijke grootheden is uit te drukken. Het is, met andere woorden, niet reduceerbaar. Als twee wiskundigen allebei beweren dat ze een bepaalde uitspraak of gevolgtrekking begrijpen, dan is onmogelijk na te gaan of ze deze op dezelfde manier begrijpen. Er is geen algoritme dat tot begrip leidt. Begrip overkomt je, het emergeert uit de gegevens die je probeert te verwerken. ChatGPT gaat hierbij dus niet helpen, behalve wellicht bij het aanleveren van die gegevens.

Voor begrip is bovendien bewustzijn nodig. Je kunt alleen maar iets begrijpen als je je bewust bent van jezelf. We hebben echter geen idee wat bewustzijn precies is. Natuurlijk weten we dat we zonder brein geen bewustzijn kunnen hebben, maar welke rol ons brein daarin precies speelt is onduidelijk, net als de vraag of bewustzijn puur materialistisch te begrijpen zal zijn. Het zou daarom niet minder dan bizar zijn om te denken dat wij iets zouden kunnen maken, een algoritme bijvoorbeeld, dat

bewustzijn heeft. We weten de facto niet eens waar we het over hebben wanneer we over bewustzijn spreken. Sommige mensen denken dat bewustzijn emergeert als het netwerk in een brein maar voldoende complex is. Dat is bijvoorbeeld wat filosoof Daniel Dennett beweert,⁹ maar dat is echt niets anders dan pure speculatie.

Stephen Wolfram, met wie onze discussie over taal begon, beweert zoals we zagen dat ChatGPT de ‘essentie van ons denken’ blootlegt. Hij blijft helaas wat in het vage over wat die essentie van ons denken dan eigenlijk zou zijn; die essentie benoemt hij niet en dat is meer dan een veeg teken. Als we een ander mens vragen of deze zich bewust is van zichzelf, en deze persoon zegt ja, dan geloven we hem of haar. Toen we dit aan ChatGPT vroegen kregen we het volgende antwoord:

Als een AI-model heb ik geen bewustzijn zoals menselijke wezens dat hebben. Ik ben geprogrammeerd om vragen te beantwoorden en taken uit te voeren op basis van algoritmen en data die ik heb geleerd. Hoewel ik soms in staat ben om complexe taken uit te voeren en nuttige antwoorden te geven, ben ik niet in staat om emoties of bewustzijn te ervaren zoals mensen dat doen.

Als je ChatGPT vraagt of het denkt, dan zegt het programma het volgende:

Als AI-model ben ik in staat om geautomatiseerde taken uit te voeren en antwoorden te genereren op basis van de informatie en algoritmen die ik heb geleerd. Dit omvat het vermogen om patronen te herkennen, gegevens te analyseren en te verwerken, en bepaalde beslissingen te nemen op basis van die informatie.