Lotte van Elteren (red.)

IK AI

Over machtige algoritmen
en verantwoordelijkheid

ISVW Uitgevers
Dodeweg 8
3832 RD Leusden
www.isvw.nl

Titel: IK AI. Over machtige algoritmen en verantwoordelijkheid
Auteur: Lotte van Elteren (red.)
Correcties: Francien Homan
Vormgeving omslag: Elianne Koolstra
Vormgeving binnenwerk: Elgraphic
Druk: Wilco, Amersfoort

1e druk

ISBN: 978-90-834369-4-4
NUR: 730

MIX
Papier | Ondersteunt verantwoord bosbeheer
FSC® C004472

ChatGPT en de deugd van het denken

Een aristotelische reflectie op generatieve AI in de klas

Anco Peeters

Zoals wel vaker met de introductie van nieuwe technologieën – denk aan de hype rond iPadscholen een jaar of tien geleden – proberen veel bedrijven, goeroes en zelfverklaarde experts voet aan de grond te krijgen met het inzetten van kunstmatige intelligentie (AI) in het onderwijs. Dat is niet verwonderlijk: wie het lukt om scholen te overtuigen van het nut van een specifiek stukje AI kan daar potentieel veel mee verdienen.

Aanstichter van deze opvallende interesse is het inmiddels even beroemde als beruchte ChatGPT. ChatGPT verraste eind 2022 het grote publiek op minstens twee manieren. Ten eerste met het inzicht dat AI indrukwekkende dingen kan doen met taal. ChatGPT laat je bijvoorbeeld snel gepersonaliseerde tekst genereren en bestaande teksten herschrijven. Ten tweede met het aantonen dat GPT, *Generative Pre-trained Transformer*, waar ChatGPT de voorkant van is, en andere taalmodellen die deze diensten mogelijk maken, ook voor gewone gebruikers extreem laagdrempelig zijn te maken: werken met GPT is net zo simpel als het uitwisselen van tekstberichten via je telefoon.

Waar discussies over het gebruik van AI in het onderwijs in het begin nog vooral gingen over het laten schrijven van essays,

is inmiddels duidelijk dat studenten AI veel breder inzetten: bijvoorbeeld voor het 'verbeteren' van hun schrijfstijl, het samenvatten van tekstbronnen en het opnemen van bronverwijzingen.[1] En dit gaat alleen nog maar over tekstuele opdrachten: binnen de verscheidenheid aan vormen van generatieve AI, waaronder ook GPT valt, vindt men AI-systemen die niet alleen tekst, maar ook spraak, muziek, afbeeldingen en zelfs video kunnen maken. Het zal geen verbazing wekken dat handige leerlingen en studenten enthousiast aan de slag gaan met de verschillende mogelijkheden die voorgaande ontwikkeling schept voor het leuker en gemakkelijker, maar zeker niet altijd beter, maken van schoolopdrachten.

De debatten over AI in het onderwijs omvatten een breed palet aan standpunten waarin de nuance soms moeilijk te vinden is. Zo meldde *The Guardian* in januari 2023 dat Australische universiteiten ervoor kozen om hun toetsen aan te passen en in een aantal gevallen slechts toetsen met alleen pen en papier toe te staan om zo plagiaat door middel van generatieve AI te voorkomen.[2] Het leverde een aantal meewarige reacties op. John Naughton, hoogleraar Publiek begrip van de technologie, vergeleek in dezelfde krant ChatGPT met het werkbladprogramma Excel door te zeggen dat beide slechts een stuk gereedschap zijn, bedoeld om bestaande menselijke capaciteiten te verbeteren. In *de Volkskrant* zag hoogleraar Wetenschapscommunicatie Ionica Smeets veel overeenkomsten tussen ChatGPT en de opkomst van de rekenmachine.[3] Zowel Smeets als Naughton concluderen dat de nieuwe AI-gereedschappen in functie niet veel zullen verschillen van eerdere technologieën en dat we, na de eerste schokgolven, een evolutie in het onderwijslandschap zullen zien waarbij generatieve AI in het veranderde onderwijs is opgenomen.

Ik zie de toekomst van generatieve AI in de klas wat somberder

in. Begrijp me niet verkeerd, ik ben enorm onder de indruk van de nieuwste technologische ontwikkelingen in mijn vakgebied en denk dat het informeren van leerlingen, studenten en docenten over deze gereedschappen geen betoog behoeft. Maar ik denk óók dat we chronisch onderschatten hoe generatieve AI ons in onze leer- en denkprocessen zal vormen en hoe op haar beurt generatieve AI weer ontworpen wordt om op deze veranderingen in te spelen. Mijn stelling is dan ook, contra Smeets en Naughton: ChatGPT is níét zoals een rekenmachine.[4]

In wat volgt zal ik eerst de werking van generatieve AI, met een focus op taalmodellen, toelichten. Vervolgens bespreek ik twee ongemakken voor het inzetten van generatieve AI in de klas. Het eerste ongemak is de moeilijkheid van het op een verstandige manier kunnen omgaan met AI, vanwege een gebrek aan praktische kennis over AI dat wij als individuen hebben. Het tweede ongemak is de kortzichtigheid waarmee we generatieve AI inzetten, terwijl het ons kritisch denken ondermijnt. Hierna bespreek ik hoe de aristotelische deugdethiek ons helpt om op een verantwoordelijke manier de inzet van generatieve AI in het onderwijs te benaderen. Ik besluit met enkele kritische opmerkingen en een schets van hoe het misschien anders kan.

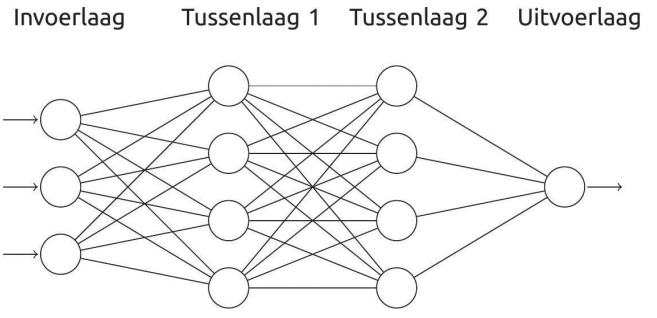
Generatieve AI: hoe werkt het?

Wat is generatieve AI eigenlijk? Er wordt nog weleens mysterieus gedaan over hoe de systemen functioneren die zelfstandig teksten, plaatjes, geluid en wat dies meer zij maken. Met name aanbieders en ontwikkelaars van deze systemen lijken in hun bedrijfsmodel op te hebben genomen dat ze zichzelf presenteren als een soort alchemist met ingewikkelde en magische formules voor het creëren van goud. Heel verwonderlijk is dat niet, want de mythische status

van generatieve AI draagt bij aan aanhoudende aandacht in het publieke discours en het verwerven van klanten en investeringen. Maar, even afgezien van de technische uitwerking, zijn de principes die aan generatieve AI ten grondslag liggen niet zo ingewikkeld als men graag doet geloven. Een beetje basiskennis over die principes helpt ons beter te begrijpen wat de GPT's van deze wereld doen en waarom ze soms de mist in gaan.

Wat in de volksmond allemaal onder de noemer 'kunstmatige intelligentie' wordt geschaard, omvat in werkelijkheid een bonte verzameling aan algoritmen, dataverwerkingsmethoden, statistiek, cognitiewetenschap, filosofie en zelfs robotica. Een belangrijk vakgebied binnen de AI is dat van de zogenoemde zelflerende systemen (machine learning). Dit vakgebied draait om het verwerken en analyseren van grote hoeveelheden data. Daar heeft men verschillende instrumenten voor en één daarvan is het maken van een rekenmodel dat een 'artificieel neuraal netwerk' (ANN) heet.

Zo'n ANN is losjes gebaseerd op de werking van biologische hersenen en bestaat uit twee hoofdingrediënten: knooppunten ('neuronen') en verbindingen (zie afbeelding voor een simpel ANN). De knooppunten zijn gerangschikt in lagen. Ten eerste is er een invoerlaag voor de data die het netwerk in gaan, bijvoorbeeld de zin: 'Hoe versloeg David Goliath?' Dan is er aan de achterkant een uitvoerlaag voor de gevraagde uitvoer, bijvoorbeeld: 'Met een slinger', met daartussenin een x-aantal tussenlagen die het echte werk doen, zoals het verwerken van woorden uit de invoerzin tot een uitvoerzin.



Waar een biologisch neuraal netwerk fysieke beperkingen kent – een schedel kan immers maar zoveel inhoud bevatten – laat een artificieel neuraal netwerk zich slechts beperken door de rekenkracht van de machines waar het op draait. Dat betekent dat je een artificieel netwerk zo groot of 'diep' kunt maken als je wilt door extra tussenlagen toe te voegen. Hoe groter de hamburger van tussenlagen wordt, des te beter het netwerk patronen kan herkennen in de data die je het voert; daarom noemt men dit type leren ook wel *deep learning*. Het idee is dat elke tussenlaag, na training, leert om een bepaalde eigenschap uit de data te halen, bijvoorbeeld dat de woorden 'David', 'Goliath' en 'slinger' vaak samen voorkomen.

Het type netwerk waar grote taalmodellen als GPT uit bestaan zijn transformeernetwerken. Speciaal aan deze transformers is het zogenoemde aandachtsmechanisme dat ze in staat stelt om op een relatief eenvoudige manier de belangrijke kernwoorden uit een invoer te halen en ze in een context te plaatsen. Hierdoor kan het netwerk het woord 'slinger' inschatten als wapen en niet als onderdeel van een klok. De transformer-architectuur werd in 2017 voorgesteld als efficiënte opvolger van bestaande netwerken die probeerden álle woorden uit een invoer aan elkaar te koppelen.[5]

Neurale netwerken leren grofweg door ladingen trainingsdata door te werken en op basis daarvan de verbindingen tussen

knooppunten te versterken of verzwakken. Je kunt een netwerk bijvoorbeeld een paar Bijbelboeken voeren. In het specifieke geval van transformers wordt met de data dan doorgaans een spelletje 'vul het ontbrekende woord in' gespeeld. Men geeft het netwerk de zin 'Goliath droeg op zijn hoofd een ___' en de woorden 'muts', 'helm' en 'koffiekopje'. Wanneer het netwerk de verschillende mogelijke uitkomsten vergelijkt met de trainingsdata, zal het bijvoorbeeld voor 'muts' een gemiddelde, voor 'helm' een hoge, en voor 'koffiekopje' een lage waarschijnlijkheidswaarde geven. Op deze manier leert het netwerk dat verbindingen tussen bepaalde woorden en zinsconstructies waarschijnlijker zijn dan andere. Herhaal dit en het netwerk leert taalpatronen herkennen.[6]

Eenmaal getraind kun je zo'n netwerk vragen om dit spelletje nog eens te spelen maar dan op grotere schaal. Je vertelt het netwerk dan: 'Vertel mij een Bijbels verhaal over een ogenschijnlijk onbeduidend persoon die een overweldigende vijand verslaat.' Het netwerk zal op basis van de geleerde voorspellingen zinnen maken die het beste passen bij de gevraagde invoer. Als het netwerk goed is getraind zal er een uitvoer uit komen die enigszins lijkt op het bekende verhaal.

Andere vormen van generatieve AI werken in principe vergelijkbaar, maar daar heeft het transformernetwerk bijvoorbeeld geleerd om gekleurde vlakjes te contextualiseren en zo intelligibele beelden te genereren – zoals van een reus die door een steen wordt geraakt. In dit geval is er natuurlijk een vele malen complexer netwerk nodig dan hiervoor beschreven. En de hoeveelheid data om het initiële beoordelende netwerk te trainen zal eerder enkele bibliotheken aan boeken beslaan dan slechts (een gedeelte van) de *Bijbel*. Hiermee wordt de trainingsduur vanzelf langer. Toch is dit in beginsel hoe generatieve AI werkt.

Met dit voorbeeld laten enkele fundamentele problemen met

zelflerende systemen zich ook eenvoudig verklaren. Ten eerste ligt het voor de hand dat de kwaliteit van de data waarmee het eerste netwerk wordt getraind merendeels verantwoordelijk is voor de kwaliteit van de data die het tweede netwerk oplevert. In vaktermen heet dat 'garbage in, garbage out', oftewel, als er aan de voorkant rommel in gaat, dan mag je ook aan de achterkant rommel verwachten. Zo is het algemeen bekend dat AI-modellen vooroordelen die in data verscholen liggen, verwerken in de patronen die ze leren. Zo zal een ongefilterd taalmodel dat getraind is op data van het internet vrolijk de meest misogyne en racistische uitspraken doen in een echo van wat het geleerd heeft.

Ten tweede is er de vraag van eigenaarschap: van wie zijn de data die gebruikt zijn om te trainen? En van wie zijn de data die geproduceerd worden? Een aantal bekende Amerikaanse schrijvers heeft inmiddels rechtszaken aangespannen om bedrijven aan te klagen die ongevraagd romans als trainingsmateriaal gebruiken bij het trainen van AI-modellen en daarmee het schrijverschap onder druk zetten.[7] Ten slotte is het trainen en draaien van een taalmodel enorm energie-intensief. Zo vraagt het verwerken van één enkele opdracht aan GPT zo'n tien keer meer energie dan een zoekopdracht bij Google. Met de huidige energiekosten die taalmodellen eisen is het daarom nog maar de vraag of ze op langere termijn duurzaam inzetbaar zullen zijn.

De aard van taalmodellen heeft een belangrijke consequentie. Omdat een taalmodel slechts leert om vaak voorkomende woordpatronen te leren als een soort veredelde spellingschecker, heeft het geen idee van de betekenis van wat het produceert. Dat het niet snel een zin als 'Gandhi probeerde in 1944 Hitler te overtuigen de oorlog te stoppen' zal maken is niet omdat een taalmodel deze personen en hun rol in de geschiedenis kent, maar omdat dit soort zinconstructies simpelweg weinig tot niet in de trainingsda-

ta zal voorkomen. Omdat taalmodellen niet werken op basis van betekenis en redeneren op basis van kansberekening, verkopen ze echter wel met enige regelmaat de grootste onzin. Dit wordt weleens 'hallucineren' genoemd, maar eigenlijk is dat een verkeerde term, omdat het impliceert dat GPT ook zinvolle, niet-hallucinerende gedachten kan hebben. Beter is het om een term van de Amerikaanse filosoof Harry Frankfurt te lenen: GPT is niet hallucinant of zelfs een leugenaar, maar een *bullshitter*.[8] Een leugenaar moet immers besef hebben van de waarheid om die te kunnen verdraaien. Een bullshitter geeft überhaupt niets om de waarheid, en dat is precies wat er met GPT aan de hand is. Hoewel bedrijven enorm veel investeren in het verhogen van de betrouwbaarheid van hun taalmodellen blijft dit een inherent probleem. Het is daarom belangrijk om taalmodellen niet te zien als veredelde zoekmachines, maar eerder als *bullshittende* verhalenvertellers.

Ongemak 1: verstand op nul met taalmodellen

Er blijken genoeg haken en ogen te zitten aan de ontwikkeling en het gebruik van generatieve AI. Toch zijn velen enthousiast over het inzetten ervan in het onderwijs en experimenteren talloze leerlingen en leraren met het gebruik ervan. Ongetwijfeld liggen daar in een aantal gevallen, vooral wanneer het enthousiasme uit de technologiesector komt, financiële belangen aan ten grondslag. Het kunnen verkopen van een AI-systeem – vaak met bijbehorende ondersteuning en dienstverlening – aan een school of scholenverbond is nu eenmaal een aanlokkelijk perspectief voor bedrijven. Dit is echter niet genoeg reden om de groeiende aanvaarding van AI in het onderwijs te verklaren. Taalmodellen geven bijvoorbeeld ook potentieel veel mogelijkheden voor docenten en leerlingen om anders om te gaan met kennisverwerking. Maar zelfs

met kennis van de werking, ogenschijnlijk onschuldig gebruik en de inzet van denkbeeldige energiezuinige taalmodellen doemen er problemen op. Laten we daarvoor kijken naar een specifieke toepassing.

Eerder noemde ik dat scholieren en studenten ChatGPT en andere taalmodellen gebruiken voor het opnemen van bronverwijzingen. Ze kunnen zo een taalmodel opdracht geven om op bepaalde plekken in tekst te noteren dat er een bronverwijzing vereist is, of die zelfs direct laten invoegen. Met name deze laatste toepassing is riskant, want een taalmodel kent geen fundamenteel onderscheid tussen bestaande en niet-bestaande bronnen. Het model kan slechts op basis van kansberekening bepaalde bronelementen die goed bij elkaar lijken te passen bij elkaar zetten, want dat is wat het geleerd heeft.

Om een voorbeeld te pakken: bij de naam 'Simone de Beauvoir' is de kans redelijk groot dat een taalmodel met de titel *La Deuxième Sexe* op de proppen komt. Gelukkig klopt deze relatie, want De Beauvoir heeft inderdaad dat boek geschreven. Maar als je een taalmodel vraagt om zelf een lijstje met boeken van De Beauvoir te maken, dan kun je ook de titel *L'Étude de l'homme* krijgen, een boek dat zij niet geschreven heeft en dat, voor zover ik kan achterhalen, slechts een bekende frase uit het werk is van Blaise Pascal.[9] Pascal was weliswaar ook een bekende Franse filosoof, maar wel een die in een heel ander gedachtegoed en tijdsgewricht opereerde dan De Beauvoir.

Wellicht komt dit op de lezer over als een wat suf voorbeeld. Iemand met een beetje verstand van zaken zal toch weten dat zij juist met dit soort gebruik goed moet uitkijken? Tot op zekere hoogte klopt dit, maar de valkuilen doemen natuurlijk niet op in ideale situaties, maar juist in situaties waarin een schrijver géén expert is op het onderwerp, geen tijd neemt voor factchecken of niet goed

de onderliggende werking en risico's van taalmodellen begrijpt. En zelfs als al deze problemen niet spelen kan het nog fout gaan.

Dit is geen fictief scenario: in december 2024 kreeg een hoogleraar aan de Universiteit van Stanford – nota bene met een leeropdracht over misinformatie – het voor elkaar om in zijn bijdrage als wetenschappelijk expert aan een gerechtelijk dossier verzonnen bronverwijzingen op te nemen.[10] De hoogleraar is een ervaren gebruiker van GPT en is goed op de hoogte van de risico's van bullshitting waar taalmodellen zo in uitblinken. Waarom ging het dan toch mis?

De hoogleraar gebruikte GPT vaak om kladversies van teksten te schrijven. Dat deed hij ook in dit geval. Hij gaf het model een lijstje met kernpunten, samen met de opdracht om deze punten uit te werken tot leesbare tekst. Daarnaast kreeg GPT de instructie om, bij wijze van herinnering tijdens later redigeerwerk, de frase '[cite]' op te nemen op plekken waar bronverwijzingen nodig zouden zijn. Daar ging het mis. Op twee plekken in de tekst plaatste het taalmodel niet een citatie, maar een verzonnen bronverwijzing die zich vermomde als een daadwerkelijk bestaande. De hoogleraar, die 600 dollar per uur betaald kreeg voor het uitwerken van zijn verklaring als expert, zag bij de latere controle en uitwerking van de tekst over het hoofd dat deze bronverwijzingen verzonnen waren.

Normaal gesproken zou je bij het opmaken van de bibliografie er zelf achter moeten kunnen komen dat de verwijzingen over niet-bestaande bronnen gaan. Maar tot overmaat van ramp – en ironie – liet de hoogleraar in misinformatie zijn bronnenlijst óók door GPT genereren. Weliswaar weer met behulp van gegevens die hij zelf aanleverde, maar daar stonden dus ook de verzonnen verwijzingen uit de lopende tekst tussen. Toen de rechtbank zijn bijdrage onder ogen zag en verwijzingen naar niet-bestaande zaken vond, liep de hoogleraar tegen de lamp.

De les uit deze anekdote is niet dat GPT dingen verzint en dat mensen dat soms over het hoofd zien. Het gaat er ook niet om dat moedwillig plagiëren of de boel willen bedonderen slecht is, zoals een scholier die antwoorden op een schoolopdracht laat genereren en doet alsof die antwoorden van hem zijn. De hoogleraar heeft op basis van zijn jarenlange ervaring en kennis waarschijnlijk een routine opgebouwd voor het uitwerken van expertbijdrages. Dat hij eerst een lijst met punten opstelt en vervolgens een taalmodel vraagt op basis daarvan een eerste kladversie uit te werken (met seinposten in de tekst waar bronverwijzingen nodig zijn) zou bovendien gezien kunnen worden als efficiënt werken. Natuurlijk, je kunt terecht morele vraagtekens zetten bij een expert die een belangrijk deel van zijn inhoudelijke bijdrage à $ 600/u uitbesteedt, maar diezelfde vraagtekens zou je kunnen neerzetten als hij dat werk door een onderzoeksassistent had laten doen en de uitkomst niet of slecht gecontroleerd had. Beide gevallen van uitbesteding rieken naar luiheid, slordigheid en hebzucht, maar de werkroutine als zodanig kan in principe, met meer aandacht, ook zonder deze problemen werken.

Wat deze casus echt interessant maakt, is dat zélfs misformatie-experts, van wie je mag verwachten dat ze verstandig en beheerst met generatieve AI zullen omgaan, toch zulke enorme missers kunnen maken. De meeste reacties op de werkwijze van de hoogleraar met GPT zullen variëren van 'Goh, best slim en efficiënt bedacht', tot 'Hmm, ik verwacht dat een expert voor dat bedrag zijn werk zelf secuur doet'. Maar het ligt helemaal niet voor de hand om van tevoren te voorspellen dat deze specifieke fout zou optreden, hoewel de fout achteraf makkelijk te verklaren is en voorkomen had kunnen worden. Toch zijn het kunnen voorspellen van mogelijke problemen en het kunnen uitleggen van hoe goed te werken met generatieve AI belangrijke voorwaarden voor wat wij door-

gaans onder verstandig gebruik verstaan. Ik zal in de voorlaatste paragraaf terugkomen op het begrip 'verstandig', wanneer ik met de blik van Aristoteles' ethiek naar AI-gebruik kijk.

Ongemak 2: generatieve AI en ongezonde gewoonten

Het tweede ongemak ligt in het verlengde van het eerste. Ons gebrekkig vermogen om in te schatten welke risico's AI oplevert, belet ons om AI te gebruiken op een manier die ons laat groeien in onze vaardigheden. Ik leg dit uit aan de hand van een aantal toepassingen.

Laten we teruggrijpen op de analogie van ChatGPT als rekenmachine. Ionica Smeets gebruikte deze analogie ter ondersteuning van haar voorspelling dat acceptatie van ChatGPT in het onderwijs een kwestie van tijd is en dat onze taak vooral is om na te denken over hoe we AI het beste inzetten. Onderzoeker Hamsa Bastani, specialist op het gebied van zelflerende systemen in het onderwijs van de Universiteit van Pennsylvania, houdt zich precies met deze vraag bezig. Zij zette met haar collega's een veldexperiment op waarbij ze onder duizend middelbare scholieren onderzocht hoe de leerprestaties in het wiskundeonderwijs onder invloed van GPT veranderden.[11]

Bastani verdeelde haar onderzoekspopulatie in drie groepen. De eerste groep gebruikte een standaardvariant van GPT, genaamd 'GPT Base', de tweede gebruikte de variant 'GPT Tutor' en de derde kreeg geen AI-ondersteuning. De werking van GPT Base is vrijwel identiek aan die van ChatGPT: je kunt het vragen voorleggen en krijgt een reactie terug. GPT Tutor werkt net even anders: in plaats van dat scholieren meteen een antwoord op een wiskundeprobleem kunnen opvragen, geeft GPT Tutor een aantal

mogelijke antwoorden (waar het goede antwoord tussen zit), een beschrijving van vaak voorkomende fouten die mensen bij dat probleem maken én hints om tot een oplossing te komen. De drie groepen scholieren moesten vervolgens wiskundeproblemen oplossen, waarbij de eerste twee groepen dat eerst met GPT-hulp deden en daarna zonder.

De voorlopige resultaten van Bastani's onderzoek in het artikel 'Generatieve AI Can Harm Learning' zijn razend interessant. Scholieren die hulp van GPT Base krijgen, scoren 48% beter dan scholieren die geen hulp krijgen. Voor GPT Tutor ligt dat percentage zelfs op 127%. Maar wanneer de ondersteuning van GPT Base wordt weggehaald, keldert de score van deze groep scholieren en scoren ze 17% lager dan degenen die geen AI-hulp kregen. De echt interessante uitkomst is echter dat scholieren die eerst GPT Tutor gebruikten en daarna zelfstandig moesten werken, nauwelijks anders scoorden dan de scholieren die nooit hulp hebben gehad. De onderzoekers geven als mogelijke verklaring dat de GPT Base-scholieren AI als kruk gebruikten en daardoor niet goed leerden lopen toen ze het gebruik van de kruk moesten ontberen. De GPT Tutor-scholieren daarentegen kregen niet direct de antwoorden, maar juist hulp en uitleg, waarmee ze leerden zélf actief de wiskundeproblemen aan te pakken.

Een van de conclusies die Bastani trekt, is dat GPT niet uniek is in het opwerpen van een dilemma tussen ergens minder goed in worden versus het sneller of beter uitvoeren van een taak. In die zin is GPT natuurlijk precies zoals een normale rekenmachine die onze puur rekenkundige vaardigheden ondermijnt. Waar ze wél een verschil ziet is dat GPT voor een veel breder scala aan intellectuele taken ingezet kan worden en daarmee ook meer impact kan hebben op onze intellectuele vermogens. Zo bestaat er een reële kans dat scholieren die langere tijd met GPT Base werken niet al-

leen geen wiskunde onder de knie krijgen, maar ook niet de probleemoplossende vaardigheden ontwikkelen die buiten de wiskunde belangrijk zijn. Denk bijvoorbeeld aan het opknippen van een groot probleem in deelproblemen en die stap voor stap op te lossen. In die zin is de ondermijning die GPT in gang zet veel ingrijpender en gevaarlijker dan die van een rekenmachine.

Het zojuist besproken onderzoek slaat een belangrijke brug tussen het gebruik van GPT in het onderwijs en de bredere vraag hoe onderwijs de ontwikkeling van onze kritische denkvermogens kan ondersteunen of juist ondergraven. Ik durf te stellen dat het risico dat Bastani schetst met betrekking tot de inzet van GPT in wiskundeonderwijs nog vele malen groter is bij het meer voor de hand liggende gebruik van GPT in allerlei schrijfopdrachten. Dit type gebruik komt niet alleen meer voor, maar speelt ook een rol in veel meer verschillende soorten onderwijstaken en -opdrachten. Om dit te begrijpen is het zaak te kijken naar de verschillende gereedschappen die we voor schrijfwerk gebruiken.

De befaamde Amerikaanse fantasyauteur George R.R. Martin schrijft zijn epos *A Song of Ice and Fire* (op televisie uitgebracht als *Game of Thrones*) in het stokoude tekstverwerkingsprogramma WordStar 4.0, op een al even geriatrische MS-DOS-computer uit de jaren negentig. In diverse interviews geeft Martin aan dat hij de afleidingen van de moderne wereld hiermee kan buitensluiten, omdat de computer geen internetverbinding heeft. Verder hint hij erop dat zijn schrijfgewoonten zo vervlochten zijn met WordStar dat andere gereedschappen hem bij het schrijven zouden hinderen, omdat ze bijvoorbeeld willen ingrijpen op de spelling van de exotische namen in de werelden die hij schept.

Martins schrijfgewoonten en zijn wens om daar controle over te houden snijden een dieper punt van reflectie aan dat in verschillende filosofische tradities aandacht heeft gekregen. Vanuit de tech-

niekfilosofie betoogt de recent overleden Amerikaanse filosoof Don Ihde dat de schrijfgereedschappen die we gebruiken niet alleen de handeling van het schrijven vormen, maar ook de mentale activiteiten van de schrijver, en daarmee de tekst. Het relatief langzame schrijven met een pen nodigt ons bijvoorbeeld uit om tijdens het schrijfproces na te denken over de woorden en formuleringen die we hanteren, terwijl het snelle typen met een computer ons uitnodigt om losser en meer te schrijven, aangezien we de digitale tekst op elk willekeurig moment kunnen aanpassen. De Brits-Australische cognitiefilosoof Richard Menary stelt het nog scherper: schrijven ís denken. Mijn denkproces krijgt meer mogelijkheden door het externaliseren van mijn gedachten op papier. Niet alleen verschaf ik mezelf hiermee de mogelijkheid om mijn op papier gezette gedachten op een grotere schaal te manipuleren dan de gedachten die zich voor mijn geestesoog ontvouwen, bijvoorbeeld door het uitwerken van een betoog met argumenten in een logische structuur; door het schrijfproces dwing ik mezelf ook om mijn gedachten te verwerken en te structureren. Met andere woorden: de ietwat vage gedachten die ik heb over de inzet van generatieve AI in de klas krijgen pas echt concrete vorm door het schrijven van dit hoofdstuk.[12]

De bespiegelingen van Ihde en Menary worden ondersteund door empirisch onderzoek over de psychologie van het schrijven. In een simpel doch elegant experiment onderzochten Luuk Van Waes en Peter Jan Schellens de mentale aspecten van schrijvers die zich op een pen verlaten en zij die woorden met een computer produceren.[13] Het blijkt dat de eerste groep zijn aandacht in de tekst voornamelijk richt op het niveau van zinnen en paragrafen. De tweede groep vestigt daarentegen de aandacht meer op de individuele woorden en is geneigd meer te pauzeren en redigeren tijdens het schrijven zelf.

Wat kunnen deze twee inzichten – namelijk dat schrijven den-

ken is en dat ons schrijven (en daarmee ons denken) door ons schrijfgereedschap vorm krijgt – ons leren over de inzet van generatieve AI in het onderwijs? Allereerst geeft het ons reden om zorgen over de invloed van schrijven met behulp van GPT uiterst serieus te nemen. Dat deze zorgen reëel zijn, wordt krachtig geïllustreerd door een column van Victoria Livingstone in *Time*. Livingstone heeft, na twintig jaar werk in het hoger onderwijs, in 2024 haar baan als schrijfdocent opgezegd. Niet omdat ze denkt dat de promovendi die ze begeleidde betere teksten maken met een taalmodel, maar omdat haar studenten zich niet lieten overtuigen dat het leren ontwikkelen van een eigen, consistente schrijfstijl een proces is dat moeizame momenten kent, omdat het nu eenmaal een vaardigheid is die veel tijd en oefening vergt. Het wrange is dat beginnende schrijvers precies dáárom niet de expertise hebben om in te zien dat zoiets als ChatGPT je eigen schrijfwerk tot een generieke, statistisch acceptabele moes pureert. Livingstone haalt sciencefictionauteur Ted Chiang aan, die zegt dat het gebruiken van ChatGPT voor schrijfopdrachten net zoiets is als het meenemen van een vorkheftruck naar de sportschool: je wordt er niet (mentaal) fit van.[14]

Mijn zorgen hier zijn breder dan de recente waarschuwing van de Nederlandse techniekfilosoof Richard Heersmink in *Nature*, waarin hij zegt dat het gebruik van taalmodellen onze cognitieve vaardigheden kunnen verarmen als we er ons mentaal te veel op verlaten. Ik denk namelijk dat het hier gaat om meer dan het uitbesteden van delen van onze denkprocessen. Dat zou zijn alsof we de radertjes van ons denken nu buiten onszelf plaatsen zonder dat het functioneren van het denkproces wijzigt.[15] Nee, de manier waarop ons denken vorm krijgt en wat we denken verandert namelijk wezenlijk door het schrijven met GPT. Dat zien we bijvoorbeeld in onderzoek dat laat zien dat niet alleen ons schrijven, maar zelfs

onze meningen beïnvloed worden door te werken met een taalmodel. En er is weinig reden om aan te nemen dat voor andere vormen van generatieve AI, die geen tekst maar afbeeldingen, video of geluid of wat dan ook maken, deze invloed anders is. Een kunstenaar die met behulp van DALL-E kunst probeert te maken zal net zo goed het risico nemen dat haar creatieve proces en visie op esthetiek onder invloed van AI vervormen. Dit is reden tot zorg, want hiermee zet het gebruik van kunstmatige intelligentie de autonomie van ons denken en onze creativiteit onder druk. Zeker in de context van het cultiveren van het kritisch denkvermogen bij scholieren en studenten die zich later in het publieke domein moeten begeven is dit een groot probleem.[16]

Een aristotelische reflectie op AI in de klas

Het signaleren van voorgaande ongemakken roept de vraag op: wat nu? Hoe kunnen we het beste omgaan met de ongemakken waar generatieve AI ons mee opzadelt? Ik denk dat het werk van de Griekse filosoof Aristoteles (384–322) ons een antwoord kan geven. Althans, ik denk dat zijn inzichten ons handvatten kunnen geven die ons helpen de problemen die generatieve AI in het onderwijs oproept beter te duiden en daarmee aan te pakken.

In de wijsbegeerte van Aristoteles gaan theoretische inzichten hand in hand met empirische beschouwingen. Dit geldt dus ook voor zijn ethische opvattingen.[17] In zijn morele denken ligt de aandacht van de Griekse wijsgeer niet op wat onze handelingen, ten goede of ten kwade, in de wereld teweegbrengen (zoals utilisten doen) of met welke beweegredenen we ze uitvoeren (zoals plichtethici stellen). In plaats daarvan richt hij zich op de vraag: hoe kan een handeling het beste bijdragen aan een goed leven? Een goed leven is volgens Aristoteles gericht op *eudaimonia* (letterlijk: een

goede of gezonde geest). Deze gerichtheid op eudaimonia kenmerkt zich door een voortdurend streven naar het beste uit jezelf halen door het cultiveren van deugden.

Dat klinkt wat vaag, dus laten we kijken naar een voorbeeld. Een gepassioneerde pianist zal ernaar streven goed piano te kunnen spelen. Goed piano leren spelen is niet een vaardigheid die op een gegeven moment 'af' is. Zonder oefening zal de muzikant langzaam maar zeker minder goed worden in het bespelen van haar instrument. Goed pianospelen is een balanceeract van het vertalen van notenschema naar toetsaanslag, het inbedden van de muziek in de akoestiek van de ruimte en het aansluiten bij de toon van de gelegenheid en het gemoed van de aanwezigen. Uiteraard is dit geen uitputtende lijst. Waar het om gaat is dat goed pianospel gekenmerkt wordt door het oefenen van vaardigheden en de kunst om deze vaardigheid op de juiste manier in specifieke situaties te ontplooien. Een harmonieus samenspel van kennis en kunde in specifieke situaties vind je volgens de aristotelische deugdethiek niet alleen bij goede musici, sporters, docenten of zelfs breiers, maar vooral ook bij *moreel* goede mensen.

Moreel balanceren komt er, in de meeste gevallen, volgens Aristoteles op neer dat men het juiste midden tussen een tekort en een teveel moet vinden. Hiervoor kijkt hij naar mensen van wie men algemeen accepteert dat ze een deugd belichamen. Een moedige soldaat is niet laf, want dat is een tekort aan moed, maar ze is zeker ook niet *over*moedig. In beide gevallen kan zij namelijk haar medesoldaten in gevaar brengen door bijvoorbeeld zonder reden te vluchten of zonder coördinatie aan te vallen. Een gematigd iemand houdt maat tussen geheelonthouding en genotzucht in het plezier scheppen in eten, drinken, seks of andere activiteiten. Een empathische arts is een dokter die meevoelt met haar patiënten zonder zichzelf te verliezen in andermans lijden of zich zo klinisch

op te stellen dat een patiënt zich niet gehoord voelt. Niet voor niets bevindt de deugd zich in het midden en behoort ze, net als het pianospel, gecultiveerd te worden door voortdurende oefening tot ze een kwestie van gewoonte en automatisme is geworden.

Het oefenen van deugdzaamheid schept voor verschillende personen in verschillende situaties verschillende uitdagingen. De ene student geneeskunde zal makkelijker een empathische arts worden dan de ander, hetzij door de eigen achtergrond en lichamelijke en neurologische samenstelling, hetzij door de mogelijkheden die ze krijgt om empathie te cultiveren. Het helpt de dokter in spe om veel ervaring op te doen met hoe mensen reageren op verschillende ziektebeelden en om mentoren te hebben die laten zien hoe empathie in de gezondheidszorg eruitziet, zodat de student zich hieraan kan spiegelen. Voor sommigen, bijvoorbeeld mensen met een narcistische persoonlijkheidsstoornis, zal deze deugd echter een onbereikbaar ideaal blijven, of hoogstens een gesimuleerde imitatie zijn.

De Schots-Amerikaanse ethica Shannon Vallor presenteert in een artikel over eenentwintigste-eeuwse deugden de aristotelische ethiek als een krachtige theorie om technologische ontwikkelingen zoals AI moreel te duiden.[18] AI daagt ons namelijk uit om na te denken over een verzameling technologieën waarvan de capaciteiten zich niet of nauwelijks laten voorspellen en de intenties van de ontwikkelaars met een gezonde dosis achterdocht moeten worden bekeken. De deugdethiek laat ons nadenken over een verantwoordelijke inzet van AI in de klas door te vragen: welke vermogens oefenen we op welke manier door dit taalmodel te gebruiken en leidt dit tot goede of slechte gewoonten?

Nu kunnen we ook uitleggen waarom het tweede ongemak, dat van ongezonde (schrijf)gewoonten, problematisch is en waarom

het een slecht idee is als studenten het schrijven van een essay uitbesteden aan een taalmodel. Het gaat bij het schrijven van een essay namelijk niet om de tekst die gemaakt wordt. Waar het om draait is het oefenen van het helder en gestructureerd formuleren van je gedachten op een manier die verder gaat dan wat borrelpraat in de kroeg. Vol zelfvertrouwen tegen docenten verkondigen: 'Laat studenten prompts inleveren in plaats van een essay', zoals sommige AI-studenten doen, is daarom een denkfout.[19] Natuurlijk is het kunnen schrijven van een essay niet zaligmakend, maar het draagt duidelijk meer bij aan de ontwikkeling van ons kritisch denkvermogen – en daarmee aan eudaimonia – dan het opstellen van invoerteksten voor een taalmodel. Niet voor niets beschrijft onderwijsfilosoof Jan Bransen in zijn boek *Gevormd of vervormd? Een pleidooi voor ander onderwijs* zijn ideale school als een oefenruimte waar fouten gemaakt mogen worden en leeractiviteiten niet gericht zijn op prestaties maar op groei.

Toch laat Vallor een belangrijk probleem onbenoemd. Bij het bespreken van het eerste ongemak vertelde ik hoe uitdagend het kan zijn om 'verstandig' om te gaan met nieuwe technologieën zoals generatieve AI. Aristoteles munt voor wat wij '(gezond) verstand' noemen de term *phronesis*.[20] Deugdethici verschillen van mening over de vraag of phronesis een op zichzelf staande deugd is of dat dit eerder een algemene voorwaarde is voor het uitoefenen van deugden. Dat onderscheid is hier niet bijzonder van belang: ik ga voor nu uit van de tweede betekenis. Aristoteles begrijpt phronesis als de vaardigheid om in de juiste situatie de juiste inschatting te maken en daarnaar te handelen. Wat deugdzaam is, is immers situatieafhankelijk.

Denk terug aan de moedige soldaat: velen van ons hebben wellicht een intuïtief beeld bij de manier waarop een moedige soldaat zou moeten handelen, maar als de kogels je om de oren vliegen en

je onmiddellijk tot handelen moet overgaan, dan is het zo simpel nog niet om te weten wanneer je moet vluchten, standhouden of tot de aanval moet overgaan. Sterker nog, in een en dezelfde situatie kan het voor de ene soldaat moedig zijn om terug te trekken, bijvoorbeeld als een officier tegen bevelen van hogerhand in een aftocht organiseert om zijn soldaten niet onnodig te laten sneuvelen, terwijl het voor een ander lafheid betekent, bijvoorbeeld een soldaat die met zijn terugtrekking een paniekerige aftocht ontketent door het laatste restje moraal van zijn medesoldaten te breken.

Het probleem is: om verstandig te kijken naar welke gewoonten bijdragen aan een deugdelijke omgang met AI, hebben we bepaalde kennis nodig. Bijvoorbeeld kennis over wat er mis kan gaan bij het geven van een opdracht aan een taalmodel. Dat taalmodellen zaken uit de duim kunnen zuigen is inmiddels wel algemeen bekend; maar hoe zit het met de Stanfordhoogleraar die GPT vroeg in een tekst aan te geven waar er geciteerd moest worden en toen de (ongevraagde!) verzonnen bronverwijzingen over het hoofd zag? Het is gemakkelijk om met de kennis van nu te stellen dat dit een dom idee was. Maar het is nu eenmaal een feit dat precies de kracht van generatieve AI, namelijk dat ze op basis van kansberekening nieuwe media kan genereren, het ook moeilijk maakt om te voorspellen hoe de gevraagde uitvoer eruit zal zien en wat hierin mis kan gaan. Daarmee is het, zeker als gewone gebruiker, momenteel onmogelijk om *phronimos*, oftewel wijs, te zijn in het dagelijks gebruik van generatieve AI. En dat is precies de reden waarom ik zelf zo huiverig ben om grootschalig gebruik van bijvoorbeeld taalmodellen in de klas aan te raden.

Besluit

De ontwikkelingen in de kunstmatige intelligentie zijn in volle gang en we kunnen er slechts naar gissen hoe nieuwe toepassingen op dit gebied er over één jaar, vijf jaar of verder in de toekomst uit zullen zien. In het voorgaande benoemde ik twee ongemakken die generatieve AI oproept: ten eerste hoe we als gebruikers worden uitgedaagd om verstandig te zijn in het gebruik ervan; en, ten tweede, hoe het niet alleen delen van onze denk- en leerprocessen kan overnemen, maar die processen zelfs fundamenteel kan vervormen.

Dat het eerste ongemak zich niet zomaar laat oplossen is een fundamenteel probleem voor het gebruik van generatieve AI. Immers, we hebben ons gezonde verstand nodig om in te kunnen schatten in welke situaties en op welke wijze we AI goed kunnen gebruiken en daarmee het tweede ongemak te voorkomen. Deze paradox laat zich slechts oplossen door proefondervindelijk ervaring op te doen in het gebruik van generatieve AI. Het beste doen we dat op een transparante manier en in een omgeving waar experts begeleiding kunnen geven. De vraag is in welke mate we jonge geesten die op onderwijsinstellingen worden opgeleid als proefkonijn voor het ontwikkelen van deze kennis willen gebruiken.

Een veelgehoorde reactie op kritische reflecties op AI is: 'Dit is de toekomst dus we kunnen het maar beter gaan gebruiken.' Of de nog ergere Engelse variant: 'AI is here to stay.' Dit is een drogreden. AI is geen van bovenaf opgelegde natuurkracht die we slechts hebben te accepteren. We hebben zelf in de hand op welke manier we generatieve AI willen gebruiken. Ik wil nog een stapje verder gaan door te zeggen dat we zelf in de hand hebben óf we generatieve AI in de klas, of daarbuiten, willen inzetten. Het getuigt van phronesis

om nuchter te constateren dat we niet de boot missen als we niet meteen met AI in zee gaan. Scholen die de iPadfase hebben overgeslagen, doen het over het algemeen nog steeds meer dan prima.

Ben ik dan helemaal tegen de inzet van AI in het onderwijs? Nee, natuurlijk niet. Bij de faculteit waar ik werk, aan de Radboud Universiteit in Nijmegen, is het Nationaal Onderwijslab AI (NOLAI) gevestigd, waar collega's en ikzelf interdisciplinair onderzoek doen naar verschillende manieren om AI in het leerproces in te zetten. Zo kijken we bijvoorbeeld naar een manier waarop AI het maken van mindmaps kan ondersteunen door concepten en verbanden daartussen te visualiseren. In die projecten die NOLAI uitvoert is altijd ook aandacht voor ethische zorgen en wordt er na afloop van een project expliciet gevraagd: was dit een goed idee? Voegt deze toepassing iets toe?

Dit soort vragen mag wat mij betreft vaker gesteld worden. De ethiek van Aristoteles helpt ons om hierbij de aandacht te richten op hoe AI leerlingen als denkende en handelende wezens (mis)vormt. Niet voor niets plaatst de in de zomer van 2024 in werking getreden 'AI Act' AI-toepassingen die van invloed zijn op examinering in de hoogste risicocategorie.[21] Want verantwoordelijk gebruik van AI verlangt meer dan je afvragen welke tijdswinst het oplevert als we het hebben over de vorming van de kneedbare geesten die we aan ons onderwijs toevertrouwen.

Noten

1. Dat studenten dit doen komt naar voren in dit artikel in *Time*: Livingstone, V. (2024), 'I Quit Teaching Because of ChatGPT', *Time*, https://time.com/7026050/chatgpt-quit-teaching-ai-essay/. Studenten maken veel gebruik van Quillbot, te vinden via https://quillbot.com/. Het boek *Chatten met Napoleon* van Barend Last en Thijmen Sprakel (Amsterdam:

Boom 2024) geeft talloze voorbeelden van toepassingen van generatieve AI in het onderwijs, al hebben de auteurs naar mijn smaak zelf iets te gretig van AI gebruikgemaakt bij het schrijven van het boek. Het is mij bijvoorbeeld niet duidelijk wat de toegevoegde waarde is van chatten met een taalmodel dat doet alsof het een historisch figuur is (en daarbij waarschijnlijk zaken verzint), versus het opdelen van de klas in tweetallen en leerlingen een rollenspel te laten spelen waarbij ze elk een historisch figuur nadoen op basis van (controleerbare) informatie die ze zelf opzoeken.

2. Cassidy, C. (2023), 'Australian universities to return to "pen and paper" exams after students caught using AI to write essays', *The Guardian*, https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays.
3. Smeets, I. (2023), 'De discussie over ChatGPT in het onderwijs doet me denken aan die over rekenmachines', *de Volkskrant*, https://www.volkskrant.nl/wetenschap/de-discussie-over-chatgpt-in-het-onderwijs-doet-me-denken-aan-die-over-rekenmachines~b40e0005/; Naughton, J. (2023), 'The ChatGPT bot is causing panic now – but it'll soon be as mundane a tool as Excel', *The Guardian*, https://www.theguardian.com/commentisfree/2023/jan/07/chatgpt-bot-excel-ai-chatbot-tech.
4. De lezing die ik gaf over geen rekenmachine zijn, is te vinden via https://www.youtube.com/watch?v=jngzUEQ_tR0.
5. Voor meer over de ontwikkeling van transformernetwerken, zie: Zhang, A. (2023), *Dive into Deep Learning*, Cambridge University Press, https://d2l.ai/chapter_attention-mechanisms-and-transformers/index.html, en: Poole, D.L. & Mackworth, A.K. (2023), '8.5 Neural Models for Sequences', in: *Foundations of computational agents*, https://artint.info/3e/html/ArtInt3e.Ch8.S5.html.
6. Voor wie meer over dit leermechanisme wil leren (*self-supervised learning*) heeft IBM hier een uitleg: https://www.ibm.com/think/topics/self-supervised-learning.

7. Voor meer over vooroordelen (biases) in generatieve AI, zie het nieuwe boek van Shannon Vallor: *The AI Mirror* (Oxford: Oxford University Press 2024). Voor de rechtszaken van auteurs tegen AI-bedrijven, zie het artikel 'Amerikaanse schrijvers klagen bedrijf achter ChatGPT aan voor schending auteursrechten', *De Standaard*, https://www.standaard.be/cnt/dmf20230921_91170868.
8. Frankfurt, H.G. (2005), *On Bullshit*, Princeton: Princeton University Press.
9. Dit is op basis van een vraaggesprek dat ik had met een lokaal model van Hermes (13B), een variant van LLaMa2.
10. Council, S. (2024), 'Stanford "lying and technology" expert admits to shoddy use of ChatGPT in legal filing', *Sfgate*. Link naar de Stanfordhoogleraar: https://www.sfgate.com/tech/article/stanford-expert-gpt-minnesota-deepfakes-19954595.phps.
11. Het artikel van Bastani, die eerder publiceerde in *Nature*, is op dit moment nog niet gepubliceerd in een wetenschappelijk tijdschrift, dus de resultaten zijn onder voorbehoud. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024), 'Generative AI Can Harm Learning', *The Wharton School Research Paper*, https://doi.org/10.2139/ssrn.4895486 / https://hamsabastani.github.io/.
12. Ihde, D. (1990), *Technology and the lifeworld: From garden to earth*, Bloomington: Indiana University Press; Menary, R. (2007), 'Writing as thinking', *Language Sciences* 29(5), pp. 621–632, https://doi.org/10.1016/j.langsci.2007.01.005.
13. Van Waes, L. & Schellens, P.J. (2003), 'Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers', *Journal of Pragmatics* 35(6), pp. 829–853, https://doi.org/10.1016/S0378-2166(02)00121-2.
14. Livingstone, V. (2024), 'I Quit Teaching Because of ChatGPT', *Time*, 2024; Chiang, T. (2024), 'Why A.I. Isn't Going to Make Art', *The New Yorker*, https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-

going-to-make-art. Chiang schreef het korte verhaal 'Story of Your Life' – over buitenaardse wezens en de invloed van taal op cognitie – dat in 2016 op meesterlijke wijze door Denis Villeneuve verfilmd werd als *Arrival*.

15. Mijn kritiek op Heersmink is een echo van de meer algemene kritiek die stromingen in de fenomenologie en belichaamde cognitie uiten op het functionalistische denken in de cognitiefilosofie dat het mentale karakteriseert als software die draait op de hardware van de hersenen. Zie voor een ietwat technische bespreking van dit debat mijn artikel over geheugen en virtual reality: Peeters, A., & Segundo-Ortin, M. (2019), 'Misplacing memories? An enactive approach to the virtual memory palace', *Consciousness and Cognition 76*, 102834–102834, https://doi.org/10.1016/j.concog.2019.102834.
16. Heersmink, R. (2024), 'Use of large language models might affect our cognitive skills', *Nature Human Behaviour*, https://doi.org/10.1038/s41562-024-01859-y; Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023), 'Co-Writing with Opinionated Language Models Affects Users' Views', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, https://doi.org/10.1145/3544548.3581196.
17. De twee hoofdwerken die Aristoteles over ethiek schreef zijn de *Ethica Nicomachea* en de *Ethica Eudemia*.
18. Vallor, S. (2021), 'Twenty-First-Century Virtue: Living Well with Emerging Technologies*', in: Ratti, E. & Stapleford, T. A. (red.), *Science, Technology, and Virtues*, Oxford: Oxford University Press, pp. 77–96, https://doi.org/10.1093/oso/9780190081713.003.0005.
19. Met de term 'prompt' wordt een invoer, zoals een vraag of opdracht, aan een taalmodel aangeduid. Artikel: Noij, M. (2024), 'Spence houdt zich bezig met de ethiek van AI: "Veel angst voor AI komt voort uit onwetendheid"', *Vox*, https://www.voxweb.nl/bericht/spence-houdt-zich-bezig-met-de-ethiek-van-ai-veel-angst-voor-ai-komt-voort-uit-onwetendheid.

20. In de traditie wordt *phronesis* soms ook met de Latijnse term *prudentia* aangeduid, wat vaak vertaald wordt als 'praktische wijsheid'.
21. Europese wetgeving over de inzet en het gebruik van kunstmatige intelligentie.

Overige bronnen

Agarwal, D., Naaman, M. & Vashistha, A. (2025). 'AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances' (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.2409.11360.

Williams-Ceci, S., Jakesch, M., Bhat, A., Kadoma, K., Zalmanson, L., & Naaman, M. (2024). 'Bias in AI Autocomplete Suggestions Leads to Attitude Shift on Societal Issues', https://doi.org/10.31234/osf.io/mhjn6.