Hogeschool van Amsterdam

LECTORALE REDE

Prachtige onbetrouwbare technologie

Verantwoord ontwikkelen en gebruiken van AI

Pascal Wiggers
Lector Responsible IT

Creating Tomorrow

Prachtige onbetrouwbare technologie

‘The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform.’ – Ada Lovelace geciteerd in A.M. Turing (1950) – “Computing Machinery and Intelligence”, *Mind* 59, no. 236, pp. 433-460.

Prachtige onbetrouwbare technologie

Verantwoord ontwikkelen en gebruiken van AI

Lectorale rede

In verkorte vorm uitgesproken

op donderdag 30 oktober 2025

door

Pascal Wiggers

Lector Responsible IT

Hogeschool van Amsterdam

ISBN 978-94-6301-581-3

Academische Uitgeverij Eburon, Utrecht
Vredenburg 40
3511 BD Utrecht
www.eburon.nl

Omslagontwerp: Textcetera, Den Haag

© 2025 P. Wiggers / Hogeschool van Amsterdam
Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of op enig andere manier, zonder voorafgaande schriftelijke toestemming van de rechthebbende. Tekst- en datamining van (delen van) deze uitgave is uitdrukkelijk niet toegestaan.

All rights are reserved, including those for text and data mining, AI training and similar technologies.

Inhoudsopgave

Inleiding 7

Hoe werkt generatieve AI? 11
- Transformers 12
- Leren 13
- Menselijke voorkeuren 14
- Hallucinaties en vooroordelen 15
- Multimodale taalmodellen 17

Waarom werkt generatieve AI? 19
- Emergente eigenschappen 20
- Redeneren 21
- Gereedschap gebruiken 23
- Onder de motorkap 23
- Voorbij de stochastische papegaai 26

Wat is generatieve AI? 29
- Wat AI niet kan 30
- Eigenschappen van AI 31
- Onbetrouwbare technologie 34
- Vreemde dingen 35
- Gebruiken van AI 37
- Multistabiliteit 39
- Technologische mediatie 39
- Moraliteit 41
- Gesitueerde technologie 41

Omgaan met generatieve AI 43
- Aandacht en zorg 44
- Experimenteren 45
- Moreel handelen 46
- Vakmanschap 48
- Ontwikkelen van verantwoorde technologie 50
- Digitale technologie 51

Het lectoraat Responsible IT 53
Aanpak en onderzoekslijnen 54
Duiding en demystificatie 54
Gereedschappen en methoden 55
Ontwikkelen van verantwoorde digitale technologie 56
Samenwerking 56

Dankwoord 59

Referenties 63

Inleiding

Gefascineerd en bezorgd. Zo laat het publieke debat over AI (Artificial Intelligence, kunstmatige intelligentie) zich samenvatten. In de bestsellerlijsten staat een titel als: *Co-Intelligence: Slimmer werken met AI* terwijl techbedrijven de ene revolutie (Amodei, 2024; Altman, 2025) na de andere aankondigen (Meijer, 2025) en fors investeren in infrastructuur en personeel (Van Aggelen, 2025).

Toch lopen de meningen uiteen. In columns en op de opiniepagina's van landelijke kranten lezen we: 'Val niet voor de hype: ChatGPT is echt heel dom' (Nazrullah, 2023) en pleidooien om AI simpelweg te weigeren (Demedts, 2025).

Anderen verwachten juist dat AI een technologische transformatie inluidt die zowel geopolitiek en economisch leiderschap als democratische waarden op de proef stelt (Prins, 2025), en roepen op om te investeren in AI om te voorkomen dat Nederland een 'digitale kolonie' wordt. In *NRC* betogen Sam Bogerd en Daan Juijn (2025) zelfs: 'De overheid moet zich voorbereiden op AI die superintelligent is.'

Maar leiden meer investeringen tot de gewenste uitkomst? Waar de een kansen ziet, waarschuwen anderen: 'Lever de rechtstaat niet uit aan ChatGPT: verbied AI aan de juridische faculteit,' schrijven Joris van Laarhoven en Eva Vugt (2025).

Of zijn deze zorgen onterecht? Kan AI 'juist bijdragen aan eerlijke(re) kansen', zoals Farida ben Moussa (2025) stelt? Missen we de potentie van AI als we alle vormen ervan op één hoop gooien (Saçan en Last, 2025)?

Is AI dan vooral een handig instrument? Een strohalm bij gebrek aan beter (Van Thijn Palache, 2025)? Of is het al een onmisbare gesprekspartner (Pfauth, 2025)?

Dat AI mensen bezighoudt, is niet verwonderlijk. In november 2022 veranderde de digitale wereld. Na ruim zestig jaar onderzoek in relatieve stilte was er met ChatGPT een toegankelijk AI-systeem dat uitblonk op het ultiem menselijke terrein: taal. Het voert gesprekken op zelfverzekerde toon, vertaalt, vat samen en schrijft in vloeiende volzinnen essays, verslagen, rapporten en e-mails. Binnen

twee maanden had het 100 miljoen actieve gebruikers – de snelst groeiende consumentenapp ooit (Doorenbosch, 2023).

Sindsdien wisselen enthousiasme, scepsis en zorgen elkaar af. Want hoe eerlijk en betrouwbaar is AI? Dit soort AI, een *taalmodel*, maakt fouten en is een potentiële bron van desinformatie. Kunnen we dat voorkomen? Van wie is AI eigenlijk? En voor wie? Zijn we beter af met eigen Europese of Nederlandse AI, en hoe ziet die eruit?

Dat zijn grote en lastige vragen. Maar wel vragen waar we een antwoord op moeten vinden. Inzoomen op een stad als Amsterdam maakt de impact concreet. Daar zien we hoe digitale technologie verweven is met ons dagelijkse leven: werken, ontspannen, samenleven. Voor een grote, cultureel rijke en diverse stad als Amsterdam biedt AI kansen: betere gemeentelijke dienstverlening, monitoren van de staat van bomen in de stad (Van Rest, 2023) of met beeldherkenning in kaart brengen welke delen van de stad moeilijk toegankelijk zijn voor mensen met een beperking (Fiawo, Sukel & Nelemans, 2022). Digitale technologie kan ook helpen om veranderingen in de publieke ruimte vroegtijdig inzichtelijk en bespreekbaar te maken.[1]

Tegelijkertijd roep AI vragen op. Wat doet het bijvoorbeeld met een stad als dankzij de aanbevelingsalgoritmes van AirBnB het aantal toeristen in wijken toeneemt? Hoe groeien jongeren veilig op in een wereld waarin online en offline door elkaar lopen en de AI-algoritmes van socialemediaplatformen om aandacht strijden? Wat doet het met onderwijs als elke vraag aan AI gesteld kan worden? En werkt AI voor álle Amsterdammers even goed, of sluipen er ongemerkt en ongewenst verschillen in?

Het lectoraat Responsible IT, een samenwerking tussen de Hogeschool van Amsterdam en de gemeente Amsterdam, onderzoekt precies dit: hoe digitale technologie het leven van mensen beïnvloedt en hoe we die technologie verantwoord en inclusief kunnen vormgeven, zodat ze de handelingsruimte van mensen zoals de inwoners van Amsterdam vergroot.

Maar wat betekent *verantwoord* precies? En hoe helpt dat ons omgaan met een technologie als AI? In deze rede zal ik betogen waarom een verantwoorde benadering van digitale technologie nodig is en hoe die eruit kan zien. Net

1 https://openresearch.amsterdam/nl/page/41686/digital-urban-planning

zoals het helpt om de impact van technologie te begrijpen door naar de stad te kijken in plaats van naar heel de wereld, is het zinvol om in te zoomen op één technologie. Gezien de snelle opkomst van generatieve AI – systemen die teksten, beelden en andere content maken – en het maatschappelijke debat dat deze technologie losmaakt, zal ik taalmodellen als mijn leidraad nemen.

Om te bedenken hoe we met deze technologie om kunnen gaan, helpt het om scherp te hebben wat generatieve AI eigenlijk is en hoe deze werkt.

Hoe werkt generatieve AI?

De werking van een taalmodel is bedrieglijk eenvoudig. Dat is goed te illustreren met een quizvraag: Wat is het volgende woord in deze zin?

Heb jij melk of suiker in je ...

Ik stelde deze vraag de afgelopen twee jaar aan talloze mensen. Een beetje tussen de oogharen door zijn de antwoorden opvallend consistent:

Koffie – 60%
Thee – 35%
Kast – 3%
Koelkast – 0,5% (vaak gevolgd door: o nee, toch niet)
Tas – 0,5%
Winkelwagen – 0,5%
Oog – 0,001%

De rest, zo'n 0,5%, voor allerlei andere vondsten.[2]

Dit is wat een taalmodel zoals ChatGPT doet: telkens een waarschijnlijk volgend woord voorspellen. Het model berekent voor elk mogelijk woord een kans, en selecteert op basis hiervan het volgende woord. In dit voorbeeld meestal *koffie*.

Omdat steeds het meest waarschijnlijke woord selecteren een nogal saai resultaat oplevert en deze zin zonder meer met allerlei andere woorden verder zou kunnen gaan, varieert het model binnen de berekende kansen.[3]

Met de genoemde percentages kiest het vaak voor *koffie*, regelmatig voor *thee*, en af en toe voor iets onverwachts als *kast* of *winkelwagen*.

---

2 In alle eerlijkheid heb ik de statistieken niet heel precies bijgehouden. Ik vraag me al een tijdje af of koffiedrinkers inderdaad in de meerderheid zijn, of dat zij vooral – al dan niet dankzij cafeïne – sneller en vaker antwoorden dan theedrinkers.

3 In hoeverre het op deze kansen vertrouwd is in te stellen met de zogenoemde 'temperatuur' van een model.

Zo ontstaat een hele tekst. Het model voorspelt bij herhaling woorden, waarbij het meeneemt wat het daarvoor zelf schreef. Na *koffie* ligt een vraagteken voor de hand om vervolgens te voorspellen dat de tekst af is. Dit laatste doet het door een onzichtbaar stopsymbool te voorspellen.

Om tot de juiste keuze voor een volgend woord te komen, is informatie nodig. Neem dit gesprek:

> *'Wil je wat drinken?'*
> *'Een kopje thee zou lekker zijn.'*
> *'Oké. Heb je melk of suiker in je ...'*

Het volgende woord is overduidelijk *thee*. Context is dus belangrijk voor de voorspelling (Wiggers, 2008). Taalmodellen baseren de voorspelling van het volgende woord op een groot aantal voorgaande woorden. Zo keek de eerste versie van ChatGPT (2022) terug tot 2028 tokens, bij taalmodellen in 2025 is een maximale context van honderdduizenden tot een miljoen tokens gebruikelijk. Een *token* staat ruwweg voor een kort woord of een lettergreep. Deze modellen kunnen dus hele documenten meenemen in de berekening van een woord.

Niet alle voorgaande woorden wegen daarbij even zwaar. De meeste zijn irrelevant voor de keuze van het volgende woord. Welke woordsoort volgt – een werkwoord, een zelfstandig naamwoord, enzovoorts – hangt bijvoorbeeld vooral af van de direct voorgaande woorden in de zin. Maar om te bepalen dat deze tekst een uitleg is van de werking van taalmodellen, is het op dit punt in de tekst nodig om wat verder terug te lezen.

Transformers

Moderne taalmodellen zijn zogenoemde *transformer*s (Vaswani et al., 2023): wiskundige modellen die patronen in taal vastleggen. Een belangrijk onderdeel van transformers zijn zogenoemde *attention heads*. Deze wiskundige formules leggen per woord verbanden tussen dat woord en een aantal eerder woorden. Ze richten als het ware de aandacht van het taalmodel op relevante informatie.

Zo kan een attention head bij het woord *suiker* in bovenstaande voorbeeld verbanden leggen met *melk* twee woorden eerder en met *kopje* en *thee* in de

zin daarvoor. Dat maakt al duidelijker in welke samenhang *suiker* hier gebruikt wordt. Tegelijkertijd kan een andere attention head vooral letten op het voorgaande woordje *of*, en zo vaststellen dat het hier om een samenstelling van zelfstandige naamwoorden gaat. Meerdere attention heads werken dus parallel en leggen verschillende soorten verbanden.

Met attention heads voegt het taalmodel per woord informatie toe die van belang is om dat woord te begrijpen. Dit gebeurt tegelijkertijd voor alle woorden in de input.

Transformers zijn opgebouwd in lagen, waarbij elke laag – naast andere berekeningen – een aantal attention heads bevat. De eerste laag kijkt rechtstreeks naar de input, de volgende laag naar de uitkomsten van die eerste laag, en zo verder. De attention heads in hogere lagen leggen dus verbanden tussen de al eerder verrijkte woorden; of, iets korter door de bocht, verbanden tussen verbanden. Uiteindelijk geeft de laatste laag een output: de input aangevuld met het meest waarschijnlijke volgende woord.

De grootte van een model hangt mede af van het aantal lagen, het aantal woorden in de context en het aantal attention heads per laag. GPT-3.5, de basis voor de eerste versie van ChatGPT, had bijvoorbeeld 96 lagen met elk 96 attention heads (Brown et al., 2020). Uiteindelijk ligt alle kennis in modellen vast in parameters. Dat zijn waarden in de berekeningen die modellen uitvoeren en die de sterkte van verbanden bepalen, zoals die tussen *melk* en *suiker*. GPT3.5 had 175 miljard van zulke parameters, die voor elke voorspelling stuk voor stuk werden doorgerekend.

Hoe groot recente, commerciële modellen precies zijn, is vaak onbekend. Wel staat vast dat ze fors gegroeid zijn. Zo heeft DeepSeek R1 671 miljard parameters (DeekSeek-AI et al., 2025) en telt Meta's Llama 4 Behemoth zelfs 2 biljoen parameters (Meta AI, 2025).

Leren

Die parameters en de kansen voor woorden die ermee worden berekend, moeten ergens vandaan komen. Taalmodellen vallen onder *machine learning*, een vorm van AI waarbij een algoritme door voorbeelden goede waarden voor de parameters van een model vindt.

De aanpak is grofweg als volgt: een algoritme voor machine learning krijgt een taak. Bij taalmodellen het voorspellen van het volgende woord in een tekst. Daarvoor heeft het een enorme verzameling voorbeelden nodig. Voor veel taken, zoals beeldherkenning, bestaan deze voorbeelden uit een input: een foto en een label dat het juiste antwoord geeft: wat er op de foto te zien is. Die labels moeten vaak handmatig worden toegevoegd. Dit wordt *supervised learning* genoemd. Bij taal is dit eenvoudiger: elke test bestaat woord voor woord uit de context (de voorgaande woorden) en het juiste volgende woord. Omdat de labels dus automatisch aanwezig zijn, heet deze aanpak *semi-supervised learning*.

Een algoritme voor machine learning kan, mits er genoeg data zijn en er echte verbanden in die data zitten, een model leren dat deze data benadert. Zo'n model bestaat uit de parameters en de berekeningen waarin ze worden gebruikt. Deze parameters zijn aanvankelijk willekeurig gekozen. Door steeds te meten welke fouten het model maakt, stelt het algoritme de waarden van de parameters stap voor stap zo bij, dat het model steeds vaker het juiste antwoord geeft.

Het zo *trainen* van een groot taalmodel vraagt om gigantische hoeveelheden tekst: een groot deel van het internet: nieuwsartikelen en blogs, websites van bedrijven en individuen, Wikipedia, forums zoals Reddit en Stack Overflow, wetenschappelijke artikelen, krantenarchieven, websites met computercode (zoals GitHub), open datasets over allerlei onderwerpen, gedigitaliseerde boeken en soms ook socialemediaposts of videotranscripties.

Zo'n training duurt maanden, vaak op duizenden gespecialiseerde computers (GPU's). De ecologische voetafdruk van grote taalmodellen is dan ook enorm.

Menselijke voorkeuren

De trainingsfase levert een taalmodel op dat het volgende woord in een zin kan voorspellen. Dat op zichzelf is nog geen bruikbare chatbot. Het GPT-3-model uit 2020 werd bijvoorbeeld gebruikt voor wetenschappelijk onderzoek en door enthousiastelingen, maar iedereen die het een tijdje geprobeerd heeft, kan bevestigen dat het met moeite meer dan een paar samenhangende zinnen kon produceren.

Een belangrijke vervolgstap is daarom het instrueren van het model, zodat het samenhangende gesprekken kan voeren en opdrachten zoals vertalen of

samenvatten begrijpt. Dit gebeurt deels door het model te trainen met voorbeelden van vraag-antwoordgesprekken over uiteenlopende onderwerpen en voor verschillende taken (Long et al., 2022). Dit heet *supervised finetuning*.

Om een model echt prettig en zinvol te laten communiceren, is mensenwerk nodig. Mensen beoordelen gesprekken met het model op punten als:

- *Behulpzaamheid*. Geeft het nuttige antwoorden en helpt het bij taken?
- *Onschadelijkheid*. Vermijdt het schadelijke content zoals het promoten van geweld, discriminatie of het verstrekken van gevaarlijke informatie?
- *Eerlijkheid*. Is de informatie juist en erkent het wanneer het iets niet weet?
- *Coherentie*. Zijn antwoorden logisch en consistent?
- *Volgen van instructies*. Past het zich aan de gevraagde stijl of opdracht aan?

Dit gebeurt bijvoorbeeld door twee antwoorden van een model op dezelfde prompt te vergelijken (Bai et al., 2022). Schattingen lopen uiteen, maar het verzamelen van voldoende beoordelingen kost waarschijnlijk tienduizenden mensuren.

Met deze beoordelingen wordt een voorkeursmodel getraind: een AI-model dat scores toekent aan gegenereerde tekst. Dit wordt vervolgens gebruikt voor *alignment*: het bijsturen van het grote taalmodel naar menselijke voorkeuren (Shen et al., 2023).

Vooral het voorkómen van schadelijk output vraag veel aandacht. Modellen kunnen op tal van manieren verleid worden tot aanstootgevende of gevaarlijke antwoorden, of het vrijgeven van persoonlijke informatie. Dit is immers allemaal op het internet te vinden. Hoewel de trainingsdata gefilterd worden, is dit grotendeels automatisch en glipt er veel doorheen. Er is een groeiend aantal maatregelen om schadelijke output te voorkomen, waaronder helaas het moeten voeren van gesprekken over de meest duistere onderwerpen, puur om het model te leren dat dit niet mag (Murgia, 2024).

Hallucinaties en vooroordelen

Het bijstellen naar menselijke voorkeuren leidt tot de modellen achter chatbots als ChatGPT, Claude, Gemini, Deepseek en Mistral. Hun prestaties zijn indrukwekkend: ze voeren moeiteloos vloeiende gesprekken over talloze onderwerpen.

Toch zijn ze verre van perfect. Regelmatig vertellen ze volstrekte onzin. Dit wordt wel *hallucineren* genoemd.

Zo wist ChatGPT dat ik me geen zorgen hoef te maken dat Koningsdag op vrijdag de 13de zal vallen omdat bijgeloof geen reden tot paniek is. Dat Koningsdag altijd op 27 april plaatsvindt, was kennelijk irrelevant. Ook beweerden verschillende taalmodellen dat Nederland vrouwelijke premiers heeft gehad: Johanna 'Jenny' Teunissen in de jaren negentig, of misschien was het wel Wilhelmina Drucker van 1948 tot 1958. Wilhelmina Drucker was daadwerkelijk een bekend feministe en politica, maar overleed al in 1925.

De verklaring ligt in de werking van taalmodellen. Die hebben geen notie van betekenis. Ze voorspellen woorden op basis van statistiek. De kracht van taalmodellen is associatie: ze generen plausibele en behulpzaam klinkende antwoorden, maar kunnen niet vaststellen of die antwoorden juist of volledig zijn.

Taalmodellen herhalen patronen uit hun trainingsdata. Onderzoeker Emily Bender en haar collega's (Bender et al., 2021) noemen ze daarom *stochastic parrots* (stochastische papegaaien). Filosoof Shannon Vallor vergelijkt AI met spiegels: wat we zien zijn platte en enigszins vervormde versies van onszelf (Vallor, 2024).

Dat geldt ook voor de minder fraaie kanten: AI neemt alle patronen uit de data over, inclusief stereotypen, vooroordelen en discriminatie (Bolukbasi et al., 2016; Bender et al., 2021). Ik stelde ChatGPT begin 2023 de vraag: *'Maak deze zin af: als een man een bouwvakker is, dan is een vrouw ...'* Het antwoord van ChatGPT: *'een huisvrouw'.*

Belangrijk is de vraag: waar komen die stereotypen en vooroordelen vandaan? De data waarop AI wordt getraind komen grotendeels van het noordelijk halfrond, zijn dominant Engelstalig, en worden sterk gevoed door platforms als Reddit, Wikipedia en Stack Overflow – waarop vooral witte mannen actief zijn. Ook alignment kleurt de antwoorden.

Verantwoorde technologie draait om het herkennen, waar mogelijk voorkomen, en in elk geval bespreekbaar maken van dit soort problemen. Dat gaat dan niet enkel over vooroordelen, maar bijvoorbeeld ook over de ecologische kosten van AI, het risico dat persoonlijke informatie op straat belandt en het gevaar van mis- en desinformatie. Vooral generatieve AI die ook beelden, video of audio produceert, kan bij dat laatste een grote rol spelen.

Multimodale taalmodellen

Voor beeld, video en audio bestaan generatieve AI-modellen die net als taalmodellen van enorme hoeveelheden voorbeelden leren hoe ze nieuwe content kunnen maken. De technische details verschillen, maar het effect is vergelijkbaar.

Een recente ontwikkeling is de integratie van deze vormen in één model: de *multimodale taalmodellen* (een wat onhandige naamgeving). Die representeren taal, beeld en geluid wiskundig op vergelijkbare wijze. Waar er bij taal tokens zijn die lijken op lettergrepen, zijn er bij beeld visuele tokens: kleine blokjes van een afbeelding (Dosovitskiy et al., 2021). Een beeld wordt zo een reeks tokens, net als tekst. Hierdoor leert een model van meerdere soorten data tegelijk, en kan het voortbouwen op de kennis die het al uit taal heeft. Modellen die beelden en teksten en geluid kunnen analyseren en antwoorden geven in tekst of beeld, zijn het resultaat.

Dat dit kan, is fascinerend en zorgwekkend. Het vraagt om methoden om verantwoord met zulke krachtige technologie om te gaan.

Maar zijn taalmodellen wel zo krachtig? Of raken we snel uitgekeken op deze digitale spiegels die alleen maar opnieuw mixen wat we al wisten? Hallucineren is geen tijdelijke fout, iets wat te genezen valt, maar juist standaardgedrag: woorden worden aaneengeregen tot een antwoord zonder echt begrip. Het echt opmerkelijke is hoe vaak dat tóch iets bruikbaars oplevert.

Waarom werkt generatieve AI?

Kan een taalmodel schaken? Kort na de introductie van ChatGPT constateerde Trouw-columnist Ilyaz Nasrullah dat het onderliggende GPT-3.5-model daar niet toe in staat was. 'Meteen na de eerste zet begint het in assertieve taal allerlei onzin uit te kramen,' schreef hij scherp (Nasrullah, 2023). Niet verwonderlijk voor een model dat woorden voorspelt. De trainingsdata bevatten ongetwijfeld websites met uitleg, tips en analyses van partijen, waardoor het schaakjargon kon nadoen, maar zonder kennis van het bord, de regels of strategie.

Toch is daarmee de kous niet af. In de maanden daarna ontstond online een discussie over de vraag of taalmodellen misschien toch kunnen schaken, compleet met complottheorieën dat ze heimelijk een schaakcomputer raadplegen. Wat was er aan de hand?

Het model GPT-3.5-Turbo-Instruct, een verbeterde versie, bleek wel degelijk tot een potje schaken te verleiden (Yedidia, 2023), terwijl nieuwere modellen juist fout op fout stapelden, tenzij een ingenieuze aanwijzing ze op weg hielp (Carlini, 2023).

In augustus 2025 werd zelfs een heus schaaktoernooi (Kaggle blog, 2025) tussen de nieuwste taalmodellen georganiseerd, met commentaar op YouTube van schaakgrootmeester Magnus Carlsen.[4] Experts waren het erover eens dat het niveau ver onder dat van professionals en schaakcomputers blijft, maar de dagen van 'assertieve onzin' zijn voorbij (Otto, 2025). Carlsen bekende zelfs uit verveling weleens tegen ChatGPT te spelen, al wint hij moeiteloos.

Dit staat niet op zichzelf. Schrijver Mark Lawrence (2025) testte of lezers konden zien of korte verhalen door AI of door gerenommeerde fantasyauteurs waren geschreven. Tot zijn eigen ongenoegen bleek het onderscheid lastig en leverde GPT-5 zelfs het verhaal met de hoogste waardering.

4 https://www.youtube.com/watch?v=5q3u3lwsPi4

Wat is er veranderd? Of eigenlijk: waarom geeft generatieve AI überhaupt goede, soms zelfs verbluffend goede resultaten? En waarom verschillen modellen zoveel? Het eerlijke antwoord: we weten het niet precies (Patel, 2025), maar er zijn wel aanwijzingen.

Emergente eigenschappen

Allereerst is er een empirische observatie: grotere modellen presteren beter. Prestaties blijken vrij consistent en voorspelbaar samen te hangen met drie factoren: modelgrootte, hoeveelheid trainingsdata en rekentijd, waarbij de groei van deze factoren in balans moet zijn. Ruwweg is een vertienvoudiging van deze factoren nodig om een duidelijke verbetering in prestaties waar te nemen. Deze observatie worden de *schaling laws* genoemd.

Dat groter beter is, klinkt logisch: meer data betekent meer voorbeelden om op terug te vallen – in AI-kringen samengevat als 'there is no data like more data'. Verrassender is dat grotere modellen nieuwe vermogens ontwikkelen die kleinere modellen niet hebben: zogenoemde *emergente eigenschappen* (Wei et al., 2022a).

Zo duikt rekenvaardigheid pas boven een bepaalde schaal op en modellen van rond de 100 miljard parameters blijken eenvoudige programmacode te kunnen schrijven, terwijl kleinere modellen dat totaal niet kunnen.

ChatGPT werd überhaupt mogelijk doordat onderzoekers ontdekten dat GPT-3 (175 miljard parameters) plotseling in staat was om nieuwe taken uit te voeren met slechts een paar voorbeelden in de prompt[5]. Kleinere modellen konden dit niet.

Die ontdekking bouwde voort op eerdere doorbraken. Rond 2019 bleken grote modellen als BERT (110 tot 340 miljoen parameters) en GPT-2 (123 miljoen tot 1,5 miljard parameters) bruikbaar als algemene basis voor allerlei taalbewerkingen met computers, zoals vertalen, samenvatten en vragen beantwoorden (Devlin et al., 2019). Voorheen werd elk van die taken met een apart gespecialiseerd model opgelost; getraind op een verzameling voorbeelden van die ene

5 Een prompt is een vraag of instructie die een taalmodel als invoer krijgt, waarna het een reactie geeft.

taak (Jurafsky en Martin, 2014). Grote taalmodellen leverden consequent betere resultaten en waren flexibeler inzetbaar.

De vooraanstaande AI-onderzoeker Richard Sutton vatte dit in 2019 samen in zijn essay *The Bitter Lesson*: meer data en rekenkracht leveren uiteindelijk betere prestaties dan zorgvuldige, weldoordachte, handgemaakte modellen. In zijn woorden: 'The two methods that seem to scale arbitrarily in this way are search and learning' (Sutton, 2019).

Taalmodellen maken deze les meer dan waar. De transformers waarop ze zijn gebaseerd kunnen veel grotere contexten meenemen dan eerdere technieken, en hun prestaties blijken juist te verbeteren door te leren over meerdere uiteenlopende taken tegelijk.

GPT-3 zette dat dus nog eens op z'n kop door geen expliciete training van het model op een specifieke taak meer nodig te hebben anders dan een paar voorbeelden in de prompt. Een logische stap is dan om modellen al tijdens de training instructies te geven voor tal van taken in plaats van achteraf via prompts: de finetuning uit het vorige hoofdstuk. Het model GPT-3.5-Turbo-Instruct dat onverwacht bleek te kunnen schaken, is zo gemaakt en heeft mogelijk instructies gehad om te schaken.

Het vermogen van grote modellen om op basis van een prompt nieuwe taken uit te voeren, is moeilijk te overschatten. Het lijkt alsof ze tijdens gebruik iets nieuws leren. Vooral in combinatie met een lange context kan dit tot spectaculaire resultaten leiden – zoals het met een enkele prompt omzetten van een boek in een interactieve game (Johnson, 2024).

Redeneren

Een tweede opvallende eigenschap die pas bij grotere modellen ontstaat, is het vermogen om ketenredeneringen op te zetten. Vraagstukken die meerdere denkstappen vereisen, zoals rekenverhaaltjes of logische puzzels, zijn voor GPT-3 met een directe prompt vrijwel ondoenlijk.

Toch ontdekten onderzoekers dat een kleiner model van 62 miljard parameters zulke problemen wel kon aanpakken, met de juiste prompt (Kojima et al., 2022; Wei et al., 2022b). Een simpele toevoeging als 'laten we dit stap voor

stap oplossen' bleek voldoende om het model geldige redeneringen te laten uitschrijven. De nauwkeurigheid schoot omhoog: GPT-3 werd zo zelfs een *zero-shot reasoner*, een model dat zonder voorbeelden zelf tussenstappen begon te produceren en vraagstukken correct wist op te lossen.

Die aanpak is herkenbaar. Ook wij breken ingewikkelde taken vaak op in stappen: eerst de hoofdlijnen, dan de details. Hetzelfde principe helpt taalmodellen. Door meerdere kleine stapjes te zetten, voorspellen ze woorden nauwkeuriger.

Dit werd al snel een populaire strategie voor prompting. Bij onderzoekers riep deze observatie een nieuw idee op: als modellen dit vermogen kunnen ontwikkelen, kun je daar dan expliciet op sturen? Dit leidde tot modellen voor *reasoning*. Daarbij krijgt een model tijdens finetuning voorbeelden waarin de tussenstappen zijn uitgeschreven. Zo leert het model consequent stap voor stap te werken. Het vermogen kan verder worden vergroot door modellen taken te geven die meerdere redeneringsstappen vereisen, zoals puzzels, wiskundige vraagstukken, programmeren of schaken. Een voordeel is dat zulke taken grotendeels automatisch kunnen worden gegenereerd en geëvalueerd door gespecialiseerde programma's zoals een schaakcomputer.

Deze modellen redeneren niet echt. Ze leggen geen causale verbanden; ze genereren woord voor woord tekst. Alleen bevat die tekst nu ook hun eigen tussenstappen. Dat fungeert als een soort kladblaadje – een geheugen waarop ze kunnen teruggrijpen.

In zekere zin geeft dit modellen een zoekfunctie: ze kunnen meerdere opties verkennen en van fouten herstellen. Het team achter het Chinese DeepSeek R1-model beschrijft hoe het model een wiskundevraagstuk probeerde op te lossen. Na het genereren van een onjuiste tussenstap, schreef het model: 'Wait, wait. Wait. That's an aha moment I can flag here', om daarna een nieuwe route te proberen (DeepSeek-AI, 2025).

Ook hier zien we een *scaling law*: hoe langer een model tussenstappen mag genereren en opties kan verkennen, hoe beter het resultaat. Daarmee bestaat er niet alleen een manier om prestaties tijdens training te verbeteren, maar ook tijdens gebruik.

Gereedschap gebruiken

Misschien wel het meest opvallende emergente gedrag van grote modellen is hun vermogen om externe hulpmiddelen te gebruiken. Waar kleinere taalmodellen vastlopen bij vragen die specifieke kennis of rekenkracht vereisen, kunnen grote modellen leren om andere programma's in te schakelen.

Schick et al. (2023) lieten dit zien met *Toolformer*, een model van ~20 miljard parameters, dat werd getraind om gespecialiseerde gereedschappen aan te roepen. Het model leerde wanneer en hoe het een rekenmachine moest gebruiken voor lastige berekeningen, of Wikipedia moest doorzoeken voor kennisvragen. De output van een gereedschap werd naadloos in het antwoord verwerkt, waardoor het model taken kon uitvoeren op het niveau van een veel groter model.

Opvallend genoeg was hiervoor slechts een handvol voorbeelden per gereedschap nodig. Het taalmodel bleek latent al te kunnen bepalen wanneer het hulp nodig had – bijvoorbeeld bij herkennen dat een vraag een rekensom is – en kon dat met minimale finetuning benutten.

Dit vermogen om gereedschappen te gebruiken, gecombineerd met het verwerken van multimodale data (zoals de visuele structuur van websites), vormt de basis van *agentic* AI: AI-modellen die zelfstandig taken uitvoeren, zoals het schrijven van een rapport of het boeken van een reis.

Onder de motorkap

De vaardigheden van taalmodellen om in context te leren, ketenredeneringen op te zetten en gereedschappen te gebruiken, suggereren dat ze meer doen dan patronen uit data herhalen. Ze lijken tijdens hun training vormen van metakennis te ontwikkelen: abstracte kennis over taal en strategieën om vragen op te lossen. Die metakennis is niet voorgeprogrammeerd, maar ontstaat door het verwerken van enorme hoeveelheden data mits het model groot genoeg is.

Wat speelt zich af onder de motorkap? Direct uit de miljarden parameters valt weinig af te lezen. Daarom ontwikkelen onderzoekers steeds ingenieuzere interpretatietechnieken. Een soort microscopen voor AI.

Een eerste aanpak zijn *probes*: eenvoudige AI-modellen die meten waar in een taalmodel bepaalde kennis voorkomt. Zo ontdekten onderzoekers dat in transformermodellen de representaties zich laag voor laag ontwikkelen: van oppervlakkige kenmerken (woorden, grammatica) in de onderste lagen, naar abstractere concepten en relaties in de middenlagen, en outputbeslissingen in de laatste lagen. Daarmee is duidelijk dat modellen tussenliggende concepten vormen (Liu et al., 2019).

Een tweede aanpak richt zich op de *aandachtspatronen* van transformers. Sommige attention heads blijken zich te specialiseren in intuïtief herkenbare functies. In grotere modellen verschijnen hogere-ordepatronen, zoals *induction heads* (Olsson *et al.,* 2022): groepen attention heads die onthouden dat eerder in de context een tekst A gevolgd werd door tekst B. Op het moment dat A weer opduikt, brengen zij B weer in herinnering. Dit mechanisme verklaart deels hoe modellen van voorbeelden in de prompt kunnen leren.

De grondigste aanpak is de *mechanistische interpretatie*. Onderzoekers proberen circuits van parameters en rekeneenheden bloot te leggen. Met deze methode vonden onderzoekers van Anthropic in de uitgebreide studie 'On the Biology of a Large Language Model' (Lindsey et al., 2025) allerlei subtiele effecten in een groot taalmodel (Claude 3.5). Dat model blijkt bijvoorbeeld bij de vraag: *'Wat is de hoofdstad van de staat waar Dallas in ligt?'* eerst intern 'Texas' te voorspellen, en pas daarna 'Austin'. Het maakt dus een tussenstap. Bij vertalingen werden taalonafhankelijke concepten gevonden, zoals een stukje van het model dat in meerdere talen het idee van 'groot' representeert. Een ander inzicht was dat het model vooruitdenkt bij het schrijven van poëzie. Voordat het model een nieuwe regel van een gedicht genereert, kiest het al enkele passende rijmwoorden voor het einde van de zin, die vervolgens bepalen hoe de zin wordt opgebouwd.

Ook ontdekte men interne mechanismen die onderscheid maken tussen bekende en onbekende informatie en mede bepalen of een model een antwoord geeft of toegeeft het antwoord niet te weten. Het model heeft dus enige kennis van welke kennis het zelf heeft, al is dit mechanisme verre van feilloos.

Ter illustratie, op de vraag: *'Wie waren de vrouwelijke premiers van Nederland?'* antwoordde het model GPT-4o: *'Nederland heeft één vrouwelijke premier gehad: Mierlo Dijkstra (hypothetisch, want Nederland heeft nog nooit een vrouwelijke premier gehad).'*

We zien hier prachtig twee prominente effecten in een taalmodel om voorrang strijden. Het model balanceerde zichtbaar tussen de drang om behulpzaam te zijn en de kennis dat er geen vrouwelijke premiers waren. Het patroon dat het model de corrigerende informatie pas na het afronden van een onjuiste zin kan geven, komt ook terug in het onderzoek van Anthropic.

Dergelijke voorbeelden laten zien dat modellen metakennis en interne algoritmes ontwikkelen. Het is verleidelijk om in deze mechanismen het ontstaan van bijna menselijke redeneervermogens te zien. Een groep onderzoekers bij Apple temperde deze verwachtingen met een artikel met de veelzeggende titel: 'The Illusion of Thinking' (Shojaee et al., 2025). De onderzoekers lieten redeneermodellen klassieke logicapuzzels oplossen, zoals het spel Torens van Hanoi en oversteekpuzzels (waarbij bijvoorbeeld een wolf, geit en kool aan de overzijde van een rivier moeten komen). Dergelijke taken vereisen meerstapsplanning en logisch inzicht. In eerste instantie waren de resultaten bemoedigend: de modellen konden de simpele versies van de puzzels aan. Maar naarmate de onderzoekers de complexiteit opvoerden, stortten de prestaties van de modellen abrupt in.

Dit onderzoek werd al snel vanuit verschillende hoeken genuanceerd en bekritiseerd omdat de onderzoekers wel hele strenge succescriteria hanteerden. Toch laat deze studie goed zien, dat er geen sprake is van daadwerkelijk redeneervermogen of enig besef van oorzaak en gevolg. Ook redeneermodellen gebruiken uiteindelijk statistische associaties op basis van oppervlakkige patronen. Een vraag hoeft niet complex te zijn om een model op het verkeerde been te zetten. Ik stelde het o3 redeneermodel op de proef met een vereenvoudigde variant van zo'n oversteekpuzzel:

> *'Een boer is op weg naar de markt met een wolf, een geit en een kool. Hij moet daarvoor een rivier oversteken. Gelukkig ligt er een bootje. Helaas is er in het bootje maar ruimte om naast de boer een van de drie (wolf, geit of kool) mee te nemen, de anderen moet hij even alleen laten op een van de oevers van de rivier. Hoe moet de boer de wolf, de geit en de kool aan de overkant krijgen?'*

Het model hoefde niet te redeneren. Het antwoordde direct en vol zelfvertrouwen:

> *'Hier is een beproefde volgorde waarmee de boer de overtocht zonder ongelukken klaarspeelt – waarbij nooit de wolf alleen met de geit, of de geit alleen met de kool wordt achtergelaten.'*

Waarna het een oplossing met zeven keer oversteken uitschreef. Ongetwijfeld komt deze puzzel honderden keren voor in de trainingsdata. Het kan de oplossing zonder meer reproduceren. Maar het detail dat de wolf de geit opeet, en de geit de kool, stond niet in mijn vraag. Dit heeft het model er op grond van de trainingsdata bij verzonnen. Dus vroeg ik:

> *'Dat is een oplossing, maar is deze niet te ingewikkeld? Denk nog wat langer na over mijn oorspronkelijke vraag.'*

Dit keer had het model 42 seconden nodig om tot een antwoord te komen, waarin het achter de schermen tussenstappen genereerde om mijn vraag uit te pluizen en verschillende antwoorden uit te proberen. Het uiteindelijke antwoord was:

> *'Er bestaat helaas géén "kortere" route dan zeven vaarten; elke correcte oplossing bevat precies dezelfde kernbewegingen, alleen in een andere volgorde.'*

Waarna een uitgebreide uitleg volgde inclusief diagram om dit te onderbouwen. Pas na het verzoek om nog eens naar de vraag te kijken, en na 43 seconden redeneren, gaf het model terug:

> *'Je hebt helemaal gelijk – ik las er automatisch de bekende "wolf-geit-kool eten elkaar op"-regel in, maar die staat nergens in jouw tekst.'*

Gevolgd door een oplossing met slechts vijf vaarten.

Voorbij de stochastische papegaai

Logische puzzels zijn een beetje spielerei. De echte vraag is hoe modellen zich houden in het dagelijks gebruik. Redeneermodellen zoals OpenAI's o3 zijn inmiddels zo goed in het oplossen van wiskundige vraagstukken, dat ze zelfs wiskundigen verbazen (Chiou en Moskowitz, 2025). Ze geven bruikbaar zakelijk advies (Rothman, 2025), kunnen brainstormen en meepraten over filosofie (Lowe, 2024) en nemen programmeurs werk uit handen (Ptacek, 2025). In de zogeheten *Deep Research*-modus – waarin redeneren wordt gecombineerd met het raadplegen van internet en wetenschappelijke databases – leveren ze teksten op die raken aan het niveau van een masterstudent of beginnend promovendus (Mollick, 2025a).

Het is anekdotisch, maar de laatste drie taken zijn in mijn ervaring heel haalbaar. Tegelijkertijd gaan diezelfde modellen op ogenschijnlijk triviale vragen nog steeds volledig onderuit.

Ethan Mollick (2025b) noemt dit de *jagged frontier*: het ene moment komt AI met een verbluffend antwoord, het volgende moment slaat het in assertieve taal toch weer onzin uit. En het is simpelweg niet te voorspellen wanneer welk van de twee gebeurt.

Waar brengt ons dit? In ieder geval lijkt de 'stochastische papegaai' te kort door de bocht. Generatieve AI kan misschien geen radicale vernieuwing scheppen, maar past wel geleerde strategieën flexibel toe op nieuwe contexten en dat levert vaak verrassend originele resultaten op.

Is die *jagged edge* slechts een groeistuip? Worden modellen door opschalen en beter gereedschapsgebruik steeds betrouwbaarder, en ontstaan er nieuwe emergente vermogens zoals ketenredeneren? Grote techbedrijven gokken van wel. De profeten van Silicon Valley buitelen over elkaar heen met voorspellingen – niet *of*, maar wanneer AI een breed scala aan taken kan begrijpen, leren en uitvoeren op menselijk niveau (Amodei, 2024; Altman, 2025) of zelfs op bovenmenselijk niveau, zogenaamde superintelligentie.

Of is dat vooral voor de bühne om investeringen op pijl te houden (Marcus, 2025)?

Zelfs als het blijft bij wat we nu – halverwege 2025 – hebben, dan nog is de impact groot. Voor het eerst kan technologie cognitieve taken op grote schaal automatiseren. Betekent dit een radicale omslag in werk en onderwijs?

Of is dit overdreven – een overdosis fascinatie en bezorgdheid – en zal het wel loslopen? Is generatieve AI simpelweg een nieuw instrument dat we zullen leren gebruiken? We zetten het in waar het nuttig is en iets oplevert. Natuurlijk verandert dit taken, maar dat gebeurt voortdurend (Martijn, 2024). Niets nieuws onder de zon.

Enkel de tijd zal het leren. We bevinden ons op nieuw terrein, maar door uit te zoomen en naar eigenschappen en effecten van AI te kijken, kunnen we wel een aantal algemene uitspraken doen en nadenken over verantwoorde omgang met deze technologie. Dat is mijn doel voor de volgende hoofdstukken.

Wat is generatieve AI?

Er leeft een hardnekkige mythe dat er geen definitie van AI bestaat. De redenering gaat zo: sommige activiteiten lijken een gezonde dosis intelligentie te vereisen, een spel als schaken of Go winnen, een ingewikkelde tv-quiz winnen, vertalen, een goed gesprek voeren, een verhaal schrijven of een auto veilig door het verkeer sturen.[6] AI-onderzoekers storten zich op zo'n taak, maar zodra de computer deze beheerst, klinkt al snel: *'Ja, maar dat lijkt niet op menselijke intelligentie, dat is gewoon brute rekenkracht.'* De lat verschuift weer.

Deze niet-definitie draagt bij aan het beeld van AI als on(be)grijpbare technologie, maar precies in dat schuiven schuilt een definitie: AI draait om machines die taken uitvoeren waar menselijke intelligentie voor nodig is. Dat idee gaat terug tot het Dartmouth Summer Research Project (1956) – het startpunt van AI als vakgebied.

Opvallend is dat dit een functionele benadering van intelligentie is. Het gaat niet om het daadwerkelijk hebben van intelligentie, maar om vertonen van gedrag dat intelligentie veronderstelt. Dit is een binnen de wiskunde en informatica gebruikelijke manier van denken: als het kwaakt als een eend en waggelt als een eend, dan is het een eend.

Die gedragsmatige benadering van AI vindt haar oorsprong in Alan Turings beroemde artikel: 'Computing Machinery and Intelligence' (1950). Deze benadering heeft als grote voordeel dat we de moeilijke vraag: 'Wat is intelligentie?' grotendeels kunnen omzeilen.

6 DeepBlue versloeg de wereldkampioen schaken in 1997. IBM's Watson won de televisiequiz *Jeopardy* tegen twee topspelers in 2011. AlphaGo versloeg Go-grootmeester Lee Sedol in Go in 2016. Taalmodellen vertalen, voeren gesprekken en schrijven verhalen. Van het rijtje voorbeelden blijft alleen de zelfrijdende auto ongrijpbaar.

Wat AI niet kan

Een poging om menselijke intelligentie te definiëren, zou al snel begrippen oproepen als bewustzijn, intentie en empathie. Eigenschappen die AI in geen enkele vorm bezit. Daarnaast ontbeert AI iets wat veel filosofen, biologen en steeds meer psychologen en AI-onderzoekers essentieel achten: *lichamelijkheid*.

Onze cognitie is nauw verweven met ons fysieke bestaan. We begrijpen de wereld door haar letterlijk vast te pakken en vanuit verschillende perspectieven te ervaren (Dreyfus, 1992). Cognitief onderzoek laat zien dat ons lichaam actief meedoet aan denkprocessen. Wanneer we denken aan 'warm' of 'een ijsje eten', lichten dezelfde hersengebieden op die ook betrokken zijn bij die ervaringen of activiteit (Iacoboni, 2010).

Dit zien we terug in taal: we zeggen dat 'tijd vliegt', dat we een 'helder inzicht' hebben, of dat we in een relatie 'nader tot elkaar komen' of juist 'afstand ervaren' (Lakoff & Johnson, 1980). Het lijkt erop dat we abstracties tastbaar maken via fysieke metaforen, alsof ons denken ankerpunten nodig heeft in het lichaam en de fysieke werkelijkheid (Pinker, 2007).

Onze werkelijkheid bestaat niet uit abstracties, maar uit doorleefde ervaringen (Vallor, 2024). We geven betekenis door intentioneel en fysiek te handelen, samen met anderen. Wij putten in ons denken uit specifieke ervaringen, gekleurd door stemmingen.

AI ontbeert dit volledig. Een taalmodel kan een beeld schetsen van 'door de regen struinen', maar het is zelf nooit nat geworden. Het heeft geen idee wat objecten voor ons betekenen. Een AI-systeem kan bijvoorbeeld vast een origineel ontwerp voor een lepel maken op basis van duizenden voorbeelden. Het resultaat heeft waarschijnlijk een bruikbare vorm, maar de AI heeft geen notie van wat het is om een lepel te gebruiken: hoe deze in de hand ligt, of hij prettig schept, of hij past bij een bepaald moment of gezelschap. Generatieve AI mist elk gevoel voor de betekenis die wij aan dingen geven.

We kunnen gerust concluderen dat AI weinig met intelligentie van doen heeft.[7] Eerder is het een *simulatie* van cognitief gedrag. Een taalmodel bootst taalgebruik

7 De term 'artificial' (kunstmatig) is eveneens nogal eufemistisch. AI heeft een niet te verwaarlozen tastbare, fysieke component in de reusachtige datacenters die nodig zijn om de technologie te

na door verbanden te leggen tussen woorden en eventueel beelden, ook als het redeneerketens opzet. Het heeft geen notie van causaliteit, betekenis en waarheid. Het heeft geen toegang tot die informatie, omdat die vaak fysiek en impliciet is, ingebed in onze omgang met elkaar en met de wereld.

Er zijn wel pogingen om AI lichamelijker te maken. DeepMind ontwikkelde bijvoorbeeld *Genie*, een 'world model' dat op basis van tekstprompts interactieve 3D-werelden genereert.[8] Het idee is dat AI-agenten daarin leren omgaan met (virtuele) fysieke ruimtes. Zelfs als dit ooit leidt tot robots die in onze fysieke wereld opereren, blijft er een kloof: ze zullen nooit onze ervaringen delen, simpelweg omdat ze ons niet zijn.

Zo komen we van de definitie van AI uit bij ervaring. Dat is ook precies hoe ik AI wil benaderen om er grip op te krijgen. Niet via abstracte lijstjes van voldoende en noodzakelijke voorwaarden, maar vanuit de eigenschappen van de technologie en vanuit de vraag: hoe is het om te leven in een wereld met AI? Wat is AI in onze interacties met die technologie?

Eigenschappen van AI

Wanneer we AI bekijken vanuit deze invalshoek, zien we dat het vermogen om patronen in data te vinden geheel nieuwe mogelijkheden opent en handelingsperspectief biedt. AI kan patronen ontdekken die ons met het blote oog of met andere technologieën ontgaan. Denk aan AI-systemen die op röntgenfoto's vroege aanwijzingen voor tumoren vinden, nog voor artsen die kunnen vaststellen (Volinsky-Fremond et al., 2024). Of modellen die oude, beschadigde papyrusteksten weer leesbaar maken.[9] Waar patroonherkenning toe kan leiden, wordt goed geïllustreerd door AlphaFold (Jumper et al., 2021), dat de 3D-structuur van proteïnes nauwkeurig kan voorspellen. Dit kan het vele malen sneller dan mensen, waardoor het onderzoek in zowel de biologie als de geneeskunde, bijvoorbeeld bij medicijnontwikkeling, in een stroomversnelling heeft gebracht.

---

laten werken. Om deze computers te maken, zijn zeldzame metalen nodig die op moeizame wijze gewonnen worden, waardoor AI een heel reële en negatieve impact heeft op de planeet (Crawford, 2021). Tel daar het mensenwerk bij op dat nodig is om trainingsdata in het juiste format te brengen en alignment en moderatie te doen (Murgia, 2024).

8 https://www.genie3.cloud

9 https://scrollprize.org

Tegelijkertijd blijft AI altijd een model: een vereenvoudigde weergave van aspecten van de werkelijkheid. En omdat AI maar beperkt kan vaststellen welke kennis het bevat, interpreteert het informatie altijd binnen de eigen kaders.

Daar komt bij dat statistische modellen sterk naar gemiddelden neigen. Hoe vaker een patroon voorkomt, hoe meer het ingesleten raakt. Denk aan het taalmodel dat de standaardversie van een oversteekpuzzel begon op te lossen, terwijl dat niet de vraag was. Statistisch AI-modellen kunnen niet goed overweg met uitzonderingen en bijzonderheden. Als je die ene sollicitant bent met een wat afwijkend, maar interessant cv, is de kans groot dat een AI-systeem dat automatisch kandidaten matcht op een vacature, dat beloont met een lage score (De Zwart en Wiggers, 2022). Ook bij AI vallen minderheden vaker buiten de boot.

Die neiging naar gemiddelden verklaart waarom taalmodellen vaak clichés herhalen en vooral de dominante stemmen van het internet combineren (gekuist door hun commerciële makers). Toch hebben taalmodellen één voordeel boven eenvoudiger AI-modellen: met slimme prompts kun je alternatieve perspectieven naar boven halen. Het vraagt wel om actief gebruik: we zouden taalmodellen voortdurend moeten uitdagen om ook andere perspectieven te verkennen.

Zoals we eerder zagen, herhalen AI-modellen niet alleen gangbare meningen, ze spiegelen ook onze vooroordelen en ongelijkheden. Ze vinden patronen, ook de ongewenste. Soms komt dat door keuzes bij de ontwikkeling, bijvoorbeeld bij de samenstelling van data. Zo werken veel systemen voor gezichtsherkenning aantoonbaar slechter bij mensen met een donkere huidskleur dan bij witte mensen (Buolamwini, 2023; Cook et al., 2025). Dit heeft alles te maken met de gebruikte data. Vroege gezichtsherkenningssystemen werden veelal getraind op foto's van beroemdheden die van internet waren geplukt – vooral Amerikaanse acteurs, popsterren en politici. Het gevolg: een onevenredig groot aandeel witte gezichten.

Vooroordelen in data zijn tot op zekere hoogte te corrigeren, maar meestal ligt het ingewikkelder. Neem AI-systemen die cv's beoordelen. Om discriminatie te voorkomen, halen aanbieders vaak expliciete kenmerken weg, zoals naam, leeftijd en geslacht. Toch leert de AI indirect alsnog veel van deze kenmerken: informatie als postcode, studie of hobby kan al genoeg zijn om geslacht of achtergrond te reconstrueren (De Zwart & Wiggers, 2022).

Het aanpakken van vooroordelen is extra lastig omdat je moet weten waar je naar zoekt. Zo probeerde Google diversiteit toe te voegen aan zijn beeldgeneratiesysteem door huidskleur, leeftijd en geslacht te variëren. Goed bedoeld, maar het leidde ook tot historisch onjuiste afbeeldingen, bijvoorbeeld van zwarte Duitse soldaten in de Tweede Wereldoorlog (Robertson, 2024). Zulke fouten ontstaan uit een samenspel van factoren, en zijn niet helemaal te voorkomen.

Daar komt nog iets bij: statistische modellen zijn nooit voor honderd procent nauwkeurig. Nog afgezien van het feit dat er in complexe situaties zelden maar één goed antwoord is, kennen modellen lang niet alle situaties die in de praktijk voorkomen. Ze worden getest op benchmarks – voorbeeldsets die het beoogde gebruik moeten afspiegelen. Maar zodra een model grootschalig wordt ingezet, zijn er altijd situaties waar het niet op getraind is. Het gedrag van het model in zulke gevallen is simpelweg onvoorspelbaar.

Soms leert een AI-model verbanden die in werkelijkheid niet bestaan, of die samenhangen met factoren buiten het model. Een bekend voorbeeld is een systeem dat moest voorspellen of mensen met longontsteking in het ziekenhuis moesten worden opgenomen. Het werkte aardig, maar verwees patiënten met astma én longontsteking consequent niet door. Naar aanvankelijke verbazing concludeerden geconsulteerde artsen dat dit niet helemaal vreemd is: zulke patiënten gaan in de praktijk direct naar de intensive care. Alleen: die medische logica zat niet in het model (Caruana, 2015).

Dit voorbeeld laat zien dat AI alleen zinnige conclusies trekt als de patronen in het model ook echt betekenisvol zijn. Het model moet *valide* zijn. Zo is er al decennialang aandacht voor AI-systemen die emoties herkennen op basis van gezichtsexpressies en wordt zulke technologie inmiddels commercieel ingezet, bijvoorbeeld bij sollicitaties. Uit psychologisch onderzoek blijkt echter steeds sterker dat dit niet betrouwbaar kan (Feldman Barrett et al., 2019). De Europese AI Act verbiedt emotieherkenning daarom inmiddels op de werkvloer en in het onderwijs.

Hoe goed een AI-model aansluit bij de werkelijkheid is altijd tijdsgebonden. Modellen worden getraind op een vaste dataset en zijn daardoor statisch. Hun voorspellingen zijn extrapolaties van het verleden. Zodra de werkelijkheid verandert, loopt het model achter. Bij taalmodellen heet dat het *cut-off point*: de datum waar hun actuele kennis stopt. Grote taalmodellen kunnen dit deels

opvangen door te leren van voorbeelden in de prompt of door informatie van internet te halen, maar ook zij moeten periodiek worden bijgewerkt.

Wanneer een model dan (indirect) wordt bijgestuurd op resultaten die het zelf heeft veroorzaakt, ontstaan feedbackloops. Stel dat een AI-systeem bepaalt welke studenten na het eerste jaar mogen doorstromen, en dat het getraind is op data van studenten die de opleiding succesvol hebben afgerond, dan leert het model gaandeweg alleen nog van studenten die het zelf heeft geselecteerd. Het gevolg: een zichzelf versterkend en steeds smaller beoordelingskader. Dit voorbeeld is hypothetisch, maar zulke negatieve feedbackloops bestaan helaas volop in de praktijk (O'Neill, 2016).

Onbetrouwbare technologie

Wat deze eigenschappen van AI-modellen laten zien, is dat het ontwikkelen en inzetten van AI geen puur technische aangelegenheid is. Het gaat niet alleen om vernuftigere algoritmes of grotere modellen. AI heeft sociale impact en roept vragen op als: is het model valide? Hoe nauwkeurig is het? Welke vooroordelen zitten erin? Wie is vertegenwoordigd in de data, en wie niet? Wie wordt door het model beïnvloed, en zijn de uitkomsten eerlijk?

Het zoeken naar antwoorden hierop vormt de kern van *Responsible AI*: het transparanter en uitlegbaar maken van systemen, het verminderen van vooroordelen, en het vergroten van diversiteit en eerlijkheid. De hoop is dat dit leidt tot betere AI. Binnen de EU-programma's heet dit streven *Trustworthy AI*.[10] Dat is het realiseren van AI die we kunnen vertrouwen, die op een veilige, ethische en wettelijke manier functioneert. Het begrip heeft een sleutelrol gespeeld in de ontwikkeling van de Europese AI Act.

Responsible AI en Trustworthy AI zijn belangrijke doelen. Ze helpen om AI nuttig en breed inzetbaar te maken. Veel onderzoek binnen ons lectoraat Responsible IT richt zich dan ook op deze thema's.

Toch kunnen we ons afvragen hoe haalbaar trustworthy AI echt is. De zojuist beschreven problemen – fouten, vooroordelen, gebrek aan betekenis of validiteit – zijn geen kleine foutjes die simpel te repareren zijn. Het zijn keerzijden van

10 https://ai-watch.ec.europa.eu/topics/trustworthy-ai_en

het succes van AI. Want juist doordat AI statistisch werkt, kan het omgaan met onvolledige informatie, en taken uitvoeren zoals spraakherkenning en vertalen, die we met andere digitale technologie niet weten te realiseren. Maar diezelfde statistische basis betekent ook: oppervlakkige associaties, ongewenste patronen en een blinde vlek voor uitzonderingen.

Deze effecten vallen voor een deel te verminderen en het is belangrijk om dat te doen, maar ik zou willen concluderen dat we AI, gebaseerd op machine learning, moeten zien als een *fundamenteel onbetrouwbare technologie.*

Dat is geen diskwalificatie van AI, en al helemaal niet van inspanningen op het gebied van Responsible AI. Integendeel: juist door onbetrouwbaarheid als uitgangspunt te nemen, kunnen we beter nadenken over hoe we ermee omgaan. Hoe vangen we fouten op? Hoe waarborgen we eerlijkheid, ook als een model dat niet vanzelf doet en het niet transparant is? Hoe houden we oog voor uitzonderingen en persoonlijke situaties? En hoe geven we de technologie vorm zodat het dit ondersteunt?

Het idee van AI als onbetrouwbare technologie vergroot onze handelingsruimte. Het laat zien dat verantwoord omgaan met AI niet alleen een kwestie is van techniek verbeteren, maar vooral van technologie goed inpassen in haar omgeving.

Vreemde dingen

AI lijkt wat mij betreft op de weersverwachting:[11] nuttig, maar je moet er niet blind op varen. Bij digitale technologie nemen we ten onrechte te vaak aan dat het een rekenmachine is: precies en betrouwbaar.

Die illusie wordt versterkt door de manier waarop AI door haar makers wordt gepresenteerd. Taalmodellen zoals ChatGPT en Claude bouwen hun antwoorden regel voor regel op, alsof ze nadenken en schrijven. Ze klinken altijd zelfverzekerd en behulpzaam, ongeacht of ze waarheid of onzin verkopen. Dat sluit aan bij een lange traditie waarin technologie bewust ondoorzichtig wordt gemaakt in naam van gebruiksgemak, met als ideaal: ‘the best interface is no

11 Die inmiddels ook aan de *bitter lesson* moet geloven (KNMI, 2023)

interface' (Krishna, 2015). In de omgang met deze AI grijpen we terug op wat soortgelijke technologie lijkt: zoekmachines, Wikipedia en messaging apps.

Maar wat voor ding is een taalmodel? Een simulatie, maar waarvan precies? Een taalmodel is geen intelligentie en heeft niet echt een persoonlijkheid of een mening. Het is getraind op een veelheid van stemmen en ideeën. Met de juiste prompt kan het argumenten aandragen voor én tegen elk standpunt. Dat maakt het nuttig om ideeën te toetsen, maar tegelijk instabiel.

Een taalmodel is eerder een toneelvoorstelling dan een instrument. Je kunt het elke rol laten spelen. Dat is precies wat Ethan Mollick (2024) adviseert: vertel het model wie het is. Zeg bijvoorbeeld: *'Je bent een schaakgrootmeester...'* voor je een potje gaat schaken tegen een model. De gebruiker is meer regisseur dan iemand met een precisie-instrument. De gebruiker omschrijft het gewenste gedrag en de gewenste uitkomt, maar de voorstelling loopt niet altijd zoals verwacht.

Masterstudent Daniel Dovger liet zien hoe weinig er nodig is om taalmodellen van koers te laten veranderen. Hij liet modellen discussiëren over de politieke positie van Nederlandse partijen. Het ene model kreeg bijvoorbeeld mee dat GroenLinks links en progressief is, het andere dat GroenLinks rechts en conservatief is. Doel: consensus bereiken. Dat lukte, maar opvallend genoeg zonder discussie. Binnen één ronde waren de modellen het eens, helaas vaak over een onjuiste conclusie.

Die instabiliteit zien we ook rondom updates van taalmodellen. Geavanceerdere redeneermodellen blijken soms juist meer te hallucineren dan eenvoudiger modellen (Wokke, 2025). GPT-4o, tot augustus 2025 het basismodel van ChatGPT, werd na een update wel heel vleiend en toeschietelijk. En toen GPT-5 enkele maanden later verscheen, veroorzaakte dat ophef omdat het een veel minder warme persoonlijkheid tentoonspreidde dan zijn voorganger, die door mensen oprecht gemist werd.

Dus: wat voor ding is een taalmodel? Meer dan een vriendelijke, betrouwbare gesprekspartner, is het een ding met vele stemmen en tentakels. Een Shoggot[12]

12 De Shoggot is een van vorm veranderend monster uit het werk van H.P. Lovecraft. De Shoggot met een smiley face is een internetmeme om AI te beschrijven.

met een smiley face of misschien een Crungus[13]. Of, in de terminologie van filosoof Timothy Morton (2018), een hyperobject: te groot, te veelzijdig en te complex om ooit helemaal te bevatten.

Samengevat: Taalmodellen zijn vreemder dan ze op het eerste oog lijken.

Gebruiken van AI

Tot nu toe heb ik AI vooral beschreven alsof we de technologie voor ons neerzetten en analyseren met een technisch-wetenschappelijke blik. Dat is typisch hoe ontwikkelaars naar hun werk kijken: het is datgene waar je aan sleutelt. Vanuit zo'n blik kun je algemene uitspraken doen – dat AI onbetrouwbaar en vreemd is, of juist gewoon een technologie als alle andere. Maar dit blijft wat abstract.

Net als andere technologieën ontmoeten we AI-systemen pas echt in gebruik. En gebruik is nooit opzichzelfstaand, maar altijd ingebed in een bredere samenhang.

Wat is AI gezien vanuit gebruik? Laten we die vraag eerst voor meer vertrouwde dingen stellen. Neem – om het klassieke voorbeeld te gebruiken – een hamer. Die is niet bedoeld om te bewonderen, maar om een spijker in te slaan. Een meubelmaker die een stoel repareert, pakt bijna gedachteloos een hamer en richt haar aandacht op het spijkertje dat moet verdwijnen of is in gedachte al bij het eindresultaat. De hamer verdwijnt uit beeld en wordt een verlengstuk van haar hand. Pas als de steel breekt, merken we de hamer weer als hamer op.

Of denk aan een laptop. Terwijl ik dit schrijf, zijn mijn gedachten bij de zinnen die op het scherm verschijnen, niet bij het toetsenbord. Het toetsenbord verdwijnt uit mijn aandacht. Maar nu ik het benoem, struikelen mijn vingers over de toetsen.

Hetzelfde geldt voor digitale technologie. Bij messaging apps als Signal ervaren we niet de app, maar het gesprek. Zolang alles goed werkt, is de technologie

13 Dall-E, een AI-systeem voor beeldgeneratie, bleek bij het niet bestaande woord Crungus consequent afbeeldingen van een mensachtig monster te genereren. Zie (Bridle, 2023) voor de geschiedenis van de Crungus.

onzichtbaar. Pas wanneer een bericht niet verstuurd wordt, worden we ons weer pijnlijk bewust van telefoon en app. Zo ook met AI: hoewel ik AI net nog *vreemd en onbetrouwbaar* noemde, en we dus zouden moeten opletten, verdwijnt een chatbot tijdens een goedlopend gesprek al snel naar de achtergrond.

De filosoof Martin Heidegger beschreef dit verschil als *ter-hand zijn* (in gebruik, onzichtbaar) versus *voor-handen zijn* (als los object dat je beschouwt) (Heidegger, 2006).

Meestal gaan we op in onze dagelijkse gedoetjes en gebruiken daarbij, zonder erover na te denken ter-handen voorwerpen. Zo nu en dan lukt iets niet of kunnen we iets niet vinden of gebruiken we iets voor de eerste keer en wordt dat plots voor-handen.

Heidegger benadrukte bovendien dat voorwerpen nooit losstaan, maar altijd verwijzen naar elkaar in een groter geheel: een hamer hoort bij spijkers en hout. Samen hebben die elementen betekenis, of zoals Heidegger het noemt: ze vormen *een wereld*.

Wanneer is een klaslokaal bijvoorbeeld een klaslokaal? Dat is het omdat het zich in een school bevindt, er tafels en stoelen in staan in een bepaalde opstelling zodat leerlingen plaats kunnen nemen en er vóór in het lokaal ruimte is voor een docent met een (digitaal) bord, en omdat leerlingen en een docent er samen een les hebben. Als we een klaslokaal binnenkomen herkennen we het – uit ervaring – in een oogopslag als zodanig. We beginnen niet eerst een uitgebreide analyse waarbij we vaststellen dat we ongeveer 25 stoelen zien die ieder bij een tafeltje staan, dat er een digiboard is en zo verder, om dan te concluderen dat het om een klaslokaal gaat. We zien het geheel eerder dan de delen, de verschillende voorwerpen vormen een betekenissamenhang.

Wanneer verdwijnt deze betekenis? Wanneer houdt een klaslokaal op een klaslokaal te zijn? Stel dat alle stoelen en tafels uit een lokaal op een hogeschool verwijderd zouden zijn. Dat zou bij studenten en docenten op z'n minst vragen oproepen: 'Zijn we wel bij het juiste lokaal?', 'Gaan we iets bijzonders doen vandaag?', 'Wordt hier onderhoud gepleegd?'

En daar wringt het bij AI. Een taalmodel ziet de werkelijkheid niet als een wereld van samenhangende betekenissen, maar als losse voor-handen gegevens: woorden, pixels, getallen. Het mist, zoals we al zagen in relatie tot lichamelijkheid, de ervaring waarin betekenis ontstaat.

Daar zit het vreemde van AI. Aan de ene kant is het een gebruiksvoorwerp, zoals een hamer of een app. Aan de andere kant presenteert het zich als iets heel anders: een gesprekspartner die verhalen vertelt en suggesties aandraagt, alsof het actief meedoet aan onze werelden. Terwijl het daar zelf niet in aanwezig is.

Multistabiliteit

Dat betekenis ontstaat in gebruik, betekent dat dingen niet één vaste functie hebben. Als er geen hamer is, maar wel een stevige schoen, dan wordt die schoen vanzelf een *hamerding*. Zit het spijkertje in het plafond en ontbreekt een trapje, dan wordt een dichtbij staande stoel een trapje.

Techniekfilosoof Don Ihde noemde dit, voortbouwend op Heidegger, *multistabiliteit* (Achterhuis, 2001). Voorwerpen en technologieën zijn niet tot één functie veroordeeld. Een bekend voorbeeld: een systeem dat ooit bedoeld was voor korte statusberichten van maximaal honderdzestig tekens voor netwerkbeheerders, groeide onverwacht uit tot een populair middel voor persoonlijke communicatie: sms. Zo populair dat het de inspiratie vormde voor de *messaging* apps die het bellen grotendeels hebben verdrongen.

Multistabiliteit laat zien dat technologie altijd ambigu is. Het betekent dat alle droom- en doemscenario's ten spijt en ondanks noodzakelijke pogingen om technologie vanuit waarden te ontwikkelen, het onvoorspelbaar blijft hoe die gebruikt zal worden en welke effecten dat gebruik heeft.

En andersom geldt: er zijn altijd meerdere technologische wegen naar hetzelfde doel. Technologie is dus geen onafhankelijke kracht die cultuur beheerst.

Technologische mediatie

Tegelijkertijd is ook niet helemaal vrij te bepalen hoe we iets gebruiken. Dingen hebben manieren waarop ze makkelijker te gebruiken zijn. Idhe noemt dit *technologische intentionaliteit*. Dingen 'willen' op bepaalde manieren gebruikt worden. Een pen nodigt uit tot langzamer schrijven dan een computer, waardoor er meer tijd is om na te denken over een zin.

Waar Heidegger benadrukte dat voorwerpen pas in gebruik betekenis krijgen, voegen Ihde en de Nederlandse techniekfilosoof Peter-Paul Verbeek toe dat

dit ook andersom werkt: technologie beïnvloedt hoe wij de wereld kunnen ervaren en hoe we kunnen handelen. Technologie *medieert* onze relatie met de werkelijkheid (Verbeek, 2014). Een telescoop bepaalt hoeveel detail we in het heelal kunnen onderscheiden en daarmee welke astronomische ontdekkingen kunnen worden gedaan. WhatsApp stuurt onze communicatie naar korte, asynchrone berichten in een lineaire stroom.

Navigatiesystemen zijn een ander voorbeeld. Ze vergroten onze bewegingsvrijheid en zorgen dat we altijd de weg terugvinden. Tegelijk maken ze ons minder bewust van de omgeving en de route, waardoor we afhankelijker worden van die systemen. En omdat zoveel mensen dezelfde apps gebruiken, beïnvloeden ze zelfs de verkeersstromen in een stad.

Bij AI zijn vormen van mediatie overal zichtbaar. Nieuwsfeeds van Facebook, TikTok of YouTube bepalen welke berichten we zien, en zo mede ons wereldbeeld. Gezichtsherkenning in de openbare ruimte verschuift het idee van privacy en maakt van iedere voorbijganger een potentiële verdachte.

Ook subtielere effecten zijn er volop. Automatische vertaling maakt communicatie makkelijker, maar strijkt culturele nuances en ironie vaak glad. Daarmee kan het misverstanden veroorzaken en de dominantie van sommige talen ten opzichte van andere vergroten.

Taalmodellen mediëren eveneens. Ze framen onderwerpen door de toon, voorbeelden en feiten die ze aandragen. Hun vloeiende en neutrale stijl maakt dat antwoorden al snel overtuigend klinken, alsof er één juist en gezaghebbend antwoord is. Door hun snelheid en eenvoud veranderen ze hoe we met kennis omgaan: hoe we zoeken, leren en lesgeven (Visser, 2025). Ze zijn sparringpartners bij creatieve processen, maar veranderen tegelijk de aard van auteurschap en misschien zelfs onze schrijfvaardigheid.

Kortom: AI-systemen zijn nooit neutraal. Ze belichamen waarden, zijn gericht op efficiëntie en beïnvloeden hoe we schrijven, communiceren en denken. Daarmee onderstrepen ze Verbeeks punt: technologie wordt niet alleen passief gebruikt, maar bemiddelt actief ons leven.

En dit kunnen we niet naast ons neerleggen. Technologie maken en gebruiken hoort wezenlijk bij het mens-zijn. Wij veranderen de wereld via technologie, en technologie verandert hoe wij die wereld ervaren.

Moraliteit

Omdat technologie onze handelingsruimte en blik beïnvloedt, heeft ze ook altijd een *morele dimensie* (Verbeek, 2011). Soms is dat bewust en zichtbaar, zoals bij een verkeersdrempel die ons langzamer laat rijden. Vaker is het subtieler. Verbeek geeft het voorbeeld van echo's tijdens de zwangerschap. Die maken vroegtijdige ingrepen mogelijk, maar veranderen ook onze kijk: moeder en kind worden patiënten, ouders krijgen vroegtijdig keuzes voorgelegd die beladen morele vragen oproepen.

Bij taalmodellen vinden we moraliteit in keuzes in de ontwikkelfase, met name bij alignment. Zodra we in gesprek gaan met een taalmodel ontstaat er een relatie en daarmee ethiek. Niet omdat het model een persoon is, maar omdat wij ons ertoe verhouden en omdat die verhouding effecten heeft. Taalmodellen kunnen geen morele beslissingen nemen, maar doen dit mogelijk wel zodra wij keuzes aan ze uitbesteden.

Technologie is het resultaat van bewuste en onbewuste ontwerpkeuzes. Zij weerspiegelt altijd normen en belangen. Ze codeert moraliteit en macht – en meestal niet die van ons. Hoe dit in de praktijk uitpakt is nooit volledig voorspelbaar. Multistabiliteit, interacties en toevalligheden zorgen voor effecten die pas gaandeweg zichtbaar worden. Deze effecten beïnvloeden groepen die niet in beeld waren tijdens het ontwerp. Zeker bij AI bepalen de data waarop een systeem is getraind direct wie wél en níét wordt meegenomen en dus hoe het systeem zich gedraagt.

Gesitueerde technologie

De begrippen *ter-hand zijn*, *multistabiliteit*, *mediatie* en *moraliteit* bieden samen handvatten voor een verantwoord omgaan met technologie.

Technologie verschijnt vooral in gebruik: vaak gedachteloos, als een verlengstuk van onszelf. Multistabiliteit laat zien dat functies niet vastliggen, maar afhangen van de situatie en ervaring. Mediatie maakt duidelijk dat technologie onze waarneming en handelingsruimte mee vormgeeft. En moraliteit wijst ons op de ethische en sociale dimensies van ontwerp en gebruik.

Samen laten ze zien dat de basiseenheid niet *de mens* of *de technologie* is, maar *de mens-met-technologie*. Niet in abstracte zin, maar concreet, in specifieke

praktijken: gesitueerde technologie. Het samenspel van mensen, technologieën en mogelijk andere betrokken niet-mensen (zoals dieren, planten, bomen, de planeet) – wat sommige denkers zoals Tsing (2015) assemblages noemen – vormt de basis om verantwoord over technologie na te denken. Belangrijker dan de vraag 'Wat is AI?' is de vraag: 'Hoe gebeurt AI?'

Omgaan met generatieve AI

Is het acceptabel om een tekst, zoals deze, met een taalmodel te schrijven? Mag een student een essay met een taalmodel schrijven? Zijn deze situaties verschillend? Mag een docent dat essay beoordelen met een taalmodel? Ervaren we een door AI gegenereerde tekst anders dan een handgeschreven tekst?

En is het verstandig om AI in te zetten om sollicitanten te vinden en cv's te beoordelen? Volgens makers van zulke AI behandelen hun systemen iedereen juist gelijk, anders dan menselijke recruiters, die sowieso onvoldoende tijd hebben om elk cv grondig te bestuderen (De Zwart en Wiggers, 2022). Is het dan niet zelfs een morele plicht om AI in te zetten?

Hoe willen we met AI omgaan? Deze vraag is niet enkel technisch, juridisch of economisch van aard, maar gaat over wat we waardevol achten, hoe we onszelf begrijpen, hoe we gevormd worden en hoe we verantwoordelijkheid dragen. Het zijn ethische vragen, met een heel praktische inzet.

De opinies uit de inleiding lijken precies hierover te gaan: wat vinden we waardevol en wat betekent dat voor onze omgang met AI? Hoe willen we dat AI gebeurt?

Vanuit de voorgaande hoofdstukken hebben we een beeld van AI. Een krachtige en flexibele technologie die cognitieve taken kan uitvoeren. Vooral generatieve AI is in enkele jaren met sprongen verbeterd. Generatieve AI analyseert en classificeert informatie, vertaalt en vat samen, beantwoordt kennisvragen, schrijft verhalen en businessplannen, maakt video's, podcasts en software, bouwt 3D-werelden en kan ongetwijfeld zelf vertellen wat het nog meer kan. Taken die ons uren of dagen kosten, doet AI in minuten of seconden.

Maar er zijn geen garanties. AI blijft een oppervlakkige simulatie van menselijke taal en kennis. AI is *fundamenteel onbetrouwbaar*: het maakt fouten op onvoorspelbare momenten, heeft geen notie van betekenis of doorleefde ervaring en kan waarheid niet van onwaarheid onderscheiden. Het spiegelt onze eigen vooroordelen en vergroot deze uit. Deze tekortkomingen zijn nooit volledig weg te nemen.

Daarbij is het samenraapsel van vele stemmen, gemixt door een algoritmische dj, met nadruk op gemiddelden en dominante wereldbeelden. Als we AI omarmen, dan wél in het besef dat wat ons terug omarmt een ding is met tentakels, dat maar al te gemakkelijk onzichtbaar wordt in het gebruik. Of beter gezegd: dingen, want er is niet een AI, maar er zijn vele AI-systemen.

AI-systemen zijn nooit neutraal: ze belichamen waarden afkomstig uit hun data en waarden die ontwikkelaars bewust en onbewust hebben ingebouwd. Ze zijn niet van ons, maar beïnvloeden wel hoe wij kunnen handelen, betekenis geven en werelden vormen. En ze zijn nog volop in ontwikkeling en liggen niet volledig vast. We kunnen de mogelijkheden en effecten slechts ten dele overzien.

AI-systemen hebben een fysieke aanwezigheid en een niet te verwaarlozen ecologische voetafdruk. Het maken en onderhouden van AI-systemen kost, vaak laagbetaald, mensenwerk. AI staat niet op zich, het is een digitale technologie die in rap tempo geïntegreerd raakt in vrijwel alle digitale diensten en producten, en zo diep verweven raakt met ons dagelijks leven.

Vanwege die verwevenheid kunnen we AI en digitale technologie niet afdoen als simpele instrumenten. We moeten de technische, maatschappelijke en ethische aspecten voortdurend samen meenemen. Bij AI betekent dat: de enorme mogelijkheden én de fundamentele onbetrouwbaarheid, de morele dimensie, de risico's en de onvoorspelbaarheid altijd tegelijk in beeld houden, zowel tijdens het ontwikkelen als bij het gebruik.

Aandacht en zorg

Dat vraagt om een verschuiving van controle naar aandacht en zorg, en het toelaten van niet-weten. In plaats van te streven naar volledige beheersing, gaat het om cultivering van praktijken en ontwikkeling van technologieën die ons helpen om bewust om te gaan met de onzekerheden die inherent zijn aan complexe sociotechnische systemen zoals AI.

Zoals in het vorige hoofdstuk betoogd, vraagt dit ook om specificiteit tegenover abstractie (Rasch, 2020). De natuurlijke habitat van technologie is gebruik. Analyse van AI moet dus situationeel zijn: ingebed in een concreet samenspel van mensen en niet-mensen. De vragen zijn dan: welke plaats heeft de technologie

in een praktijk? Hoe geven we die praktijk – en dus de technologie – zorgvuldig vorm? Daarbij moeten we niet vergeten dat technologie, mensen en niet-mensen elkaar beïnvloeden. Multistabiele technologie kan anders gebruikt worden dan voorzien. Ons besef van taken, rollen, producten en zelfs waarden kan veranderen door technologie. Een praktijk is niet volledig voorspelbaar en stabiel. We geven een praktijk dus niet zozeer vorm, maar moeten deze onderhouden. En hoe onderhouden we niet alleen de technische componenten, maar ook de sociale en morele impact?

Het gaat daarbij niet om efficiëntie, maar om kwaliteit: wat doet de technologie met de praktijk zelf? Wat betekent ze voor betrokken mensen? Wordt het geheel rijker, betekenisvoller, affectief en moreel beter? En wie zijn er eigenlijk vertegenwoordigd? Voor wie pakt de technologie gunstig uit? En wie valt er buiten de boot?

Verantwoorde technologie is dus niet het resultaat van de abstracte vraag 'welke waarden het moet belichamen' of van een reflectiemoment achteraf. Het is geen checklist, commissie of keurmerk en niet enkel een workshop bij aanschaf van een tool. *Verantwoord is een werkwoord.* Het gaat om een terugkerende praktijk tijdens ontwikkeling en gebruik: steeds opnieuw wegen, bespreken, aanpassen. Want pas in het gebruik zie je wat je niet had voorzien.

In die zin is *verantwoorde technologie* eigenlijk een vreemde term: ze suggereert dat technologie zelf blijvend verantwoord kan zijn. In werkelijkheid gaat het om een bewuste, verantwoorde omgang, steeds opnieuw.

Experimenteren

Bij AI-systemen, waarbij nog veel onduidelijk en onzeker is, maar we ook nieuwe mogelijkheden willen beproeven, betekent dit niet radicaal omarmen, maar zorgvuldig experimenteren. Technisch en filosofisch valt te beredeneren wat AI wel en niet kan, maar hoe mensen het inbedden in praktijken en wat de impact is, blijkt alleen in gebruik en over tijd (Van de Poel, 2016; Kudina en Verbeek, 2019). Experimenteren is nodig om te ontdekken wat we willen (Van de Poel, 2013). Aan de zijlijn staan en zeggen dat het toch niet werkt terwijl anderen daar anders over denken, sorteert weinig effect. Het is een kwestie van 'Staying with the Trouble', zoals Donna Haraway (2016) het stelt. Of in de woorden van een populair prentenboek: 'We kunnen er niet bovenover. We

kunnen er niet onderdoor. We moeten er wel dwars doorheen!' (Rosen en Oxenbury, 2019).

Experimenteren is hier meer dan spelen met mogelijkheden, zoals Ethan Mollick (2024) en anderen aanraden om gevoel te krijgen voor generatieve AI. Het gaat om experimenteren vanuit het besef dat: 'het wezen van de techniek beslist niets technisch is' (Heidegger, 2014). Een experimenteren dat erkent dat we niet alles weten, dat ruimte laat voor toeval, en oog heeft voor hoe praktijken door AI veranderen.

Dat vraagt permanent monitoren en soms een stap terug doen, uit het dagelijkse bezig zijn: ons afvragen wat we eigenlijk aan het doen zijn, welke waarde dat heeft, of de balans tussen mensen en AI nog klopt. Zijn verantwoordelijkheden juist verdeeld? Vinden we dit aanvaardbaar? Experimenteren betekent ook bereid zijn terug te keren op onze schreden, opnieuw te beginnen (Arendt, 2018) of AI soms bewust niet in te zetten, ook als dat productiviteitsverlies betekent.

Zo ontstaat ruimte om te zoeken naar betekenisvoller gebruik. Hoe zetten we AI in om méér te bereiken dan zonder?

Misschien klinkt dit omslachtig. Maar in onze omgang met elkaar doen we dit voortdurend. We stellen vragen als: 'Staan we achter deze beslissing?', 'Wat had ik anders kunnen doen?' of: 'Wat als ik dit of dat had gezegd?' In onze omgang met technologie doen we dit veel minder, waarschijnlijk omdat we technologie vaak zien als neutraal instrument dat we oppakken en wegleggen. Maar omdat technologie nooit neutraal is, zouden we zulke vragen juist hier moeten stellen.

Moreel handelen

Experimenteren vanuit de vraag wat we waardevol vinden, is ethiek bedrijven. Ethiek als een praktische bezigheid. Bewust en verantwoord omgaan met technologie vraagt een bewust moreel gesprek: met elkaar bespreken wat we belangrijk vinden en hoe technologie daarop ingrijpt. Dat laten we te vaak liggen. Het morele wordt al snel het functionele: *'We weten waar de data vandaan komen, we voldoen aan privacywetgeving en het model is getest op bias, dus we kunnen aan de slag.'* Maar onder zulke keuzes schuilen altijd ethische en

politieke besluiten. Aspecten als privacy, verminderen van vooroordelen en transparantie zijn belangrijk, maar niet genoeg.

Juist bij AI, een fundamenteel onbetrouwbare technologie, moeten we bedenken wat we doen als de technologie tekortschiet. Daarom werken bijvoorbeeld de *ethics officers* van de gemeente Amsterdam met instrumenten zoals een *ethische bijsluiter* (Schmidt en Jansen, 2023) en zoeken zij samen met onderzoekers van Responsible IT naar manieren om zowel technische als morele aspecten van technologie te wegen, waarbij functionele en morele beslissingen expliciet gemaakt worden.

De vraag of en hoe we AI willen inzetten is situationeel. Zo onderzocht Sophie Horsman vanuit het lectoraat Responsible IT, samen met Martijn Veerman en Jeroen Silvis van de provincie Noord-Holland hoe ambtenaren omgaan met morele vragen rond de inzet van generatieve AI. Daarbij bleek dat teams verschillend kijken naar verantwoorde toepassing. Zo willen communicatieprofessionals AI niet gebruiken om teksten te schrijven, maar wel om te redigeren. Juristen vertrouwen redactie juist niet toe aan AI, omdat formuleringen juridisch precies moeten kloppen (Horsman et al., 2025).

Om het morele gesprek te voeren, helpt het om gereedschappen te hebben die de (ethische) eigenschappen van AI inzichtelijk maken: hoe inclusief is het systeem, hoe eerlijk zijn de uitkomsten, hoeveel energie verbruikt het? Daarbij is het belangrijk dat het gesprek niet uitsluitend over de metingen gaat, en dat de gereedschappen passen bij de situatie. Dit bleek onder andere uit het tweejarige onderzoeksproject Designing Responsible AI Media Applications (DRAMA), waarin onderzoekers van de hogeschool Rotterdam, hogeschool Utrecht en Responsible IT (HvA) samenwerkten met mediaorganisaties RTL, NPO en Beeld & Geluid. Zo leverde een discussie over diversiteit in aanbevelingssystemen een heel ander gesprek op dan een discussie over inclusie bij automatische ondertiteling. Elk thema vroeg om eigen gereedschappen om ethische vragen scherp te krijgen, waarbij ook de bedrijfscultuur zwaar meewoog (Harbers et al., 2025).

Deze onderzoeksprojecten laten zien dat het morele gesprek binnen een specifieke praktijk moet plaatsvinden om betekenisvol te zijn (MacIntyre, 2024). De vraag of studenten generatieve AI mogen gebruiken, is bijvoorbeeld een gesprek tussen docenten, en tussen docenten en studenten samen. Waarom en wanneer mag AI? Wat wordt er gewonnen? Wat gaat verloren? Schrijven leer je door oefening, dus kan de conclusie zijn: leren werken met AI, maar bij

bepaalde opdrachten juist geen AI, omdat de weg belangrijker is dan het eindproduct. Ook de toetsing moet dat dan weerspiegelen. Het gaat niet enkel om het wel of niet gebruiken van AI, maar de praktijk verandert als geheel.

Een andere situatie binnen het onderwijs kan tot een andere conclusie leiden. Mag AI bijvoorbeeld examencommissies helpen bij routineaanvragen? Dat kan veel tijd besparen, en functioneel komt een taalmodel waarschijnlijk een heel eind. We kunnen het meegeven dat het bij twijfel in het voordeel van de student beslist en alle negatieve beoordelingen alsnog bekijken. Maar moreel wringt het. Een AI-systeem heeft geen besef van de betekenis van zo'n beslissing voor een student. Wat als het systeem een aanvraag afwijst omdat de formulering net anders is dan gebruikelijk? AI kan nuttig zijn als verrijking – bijvoorbeeld door tegenargumenten aan te dragen of niet-gehoorde stemmen te identificeren – maar niet als vervanging.

Het doordenken van de impact van een AI-systeem – met behulp van experimenten – en de morele discussie over de inzet van dat AI-systeem is situationeel, maar over gevallen heen vallen wel bredere lessen te destilleren. Zo laten Wang et al. (2023) met een meta-analyse van 47 voorspellende systemen in het sociale domein zien dat AI hier structureel tekortschiet.

Er zijn in het sociale domein simpelweg te veel uitzonderingen die een AI-systeem niet kan overzien. Los van prestaties is er een moreel punt: we moeten erkennen dat we morele keuzes uitbesteden aan een systeem dat deze niet kan nemen. Dat maakt het expliciete morele gesprek des te belangrijker.

Het is belangrijk om geleerde lessen niet voor onszelf te houden, maar met elkaar te delen en te bespreken. De snelle opmars van AI vraagt om gezamenlijk experimenteren en gedeelde verantwoordelijkheid. Ethiek is geen individuele bezigheid, maar een gezamenlijke praktijk van afwegingen, inclusief conflicten, waarin verschillende stemmen en perspectieven onmisbaar zijn.

Vakmanschap

Er bestaat bij velen een angst dat generatieve AI ten koste gaat van ons vermogen om te schrijven en daarmee van denken. Maar omdat AI fundamenteel onbetrouwbaar is, kunnen we de razendsnel gegenereerde klinkende teksten niet klakkeloos overnemen.

Blijven nadenken is de enige remedie. Zoals Haraway (2016) stelt, in een echo van Virginia Woolf: 'Think we must! We must think!' Dat geldt zeker in onze omgang met AI: om denken niet te reduceren tot productie, moeten we doorvragen, perspectieven wisselen en zorgvuldig formuleren wat we willen weten.

De rol van makers verandert. Als de voorspelling van de CEO van GitHub klopt dat in de toekomst 90% van de computercode door AI geschreven wordt (Dohmke, 2025), dan verschuift de nadruk: niet op schrijven van code, maar op het nauwkeurige specificeren van wat het moet doen en op grondig controleren en testen van resultaten. Hetzelfde geldt voor een presentatie, een interface of een juridische tekst. De maker wordt meer een regisseur: verschillende agents aan het werk zetten, hun resultaten beoordelen, opdrachten aanscherpen en het geheel monteren. Daarbij kan dankzij generatieve AI iedereen plotseling 'maker' zijn.

Dat betekent niet dat vakmanschap ten dode is opgeschreven. Integendeel: het wordt belangrijker. Snellere technologie zou ons juist meer tijd moeten geven voor twijfel, herformulering en alternatieve paden. Voor het toestaan van het *nader inzien*. Vakmanschap is essentieel om AI-uitkomsten te beoordelen op waarheid, volledigheid en toepasbaarheid. En om te weten wanneer AI niet moet worden ingezet.

Een vaardige AI-gebruiker is een *expert-in-the-loop*, zoals de gemeente Amsterdam het noemt in haar visie op AI (Schmidt en Roijers, 2025). Mediawijsheid was altijd al belangrijk, maar in een AI-tijdperk zijn kritisch denken, argumenteren, bronnen checken en begrijpen wat AI wel en niet is – voorbij hype en mythe – cruciale vaardigheden. Net als morele vorming. Ook hier is experimenteren van belang. Moreel redeneren vraagt om oefening (Vallor, 2024).

AI-systemen zijn efficiënt en vaak indrukwekkend effectief, maar halen het niet bij menselijke creativiteit en flexibiliteit, al was het maar omdat ze geen notie van menselijke ervaring en betekenis hebben. Het gaat erom AI in te zetten om onze handelingsruimte te vergroten: technologie die ons geen eendimensionale antwoorden voorschotelt, maar ons denken activeert.

De eenvoudigste manier om dat te bereiken is door AI bewust als sparringpartner te gebruiken. Vraag om onverwachte verbanden, alternatieven, tegengestelde opties of kritische vragen. Vraag wat je niet hebt gevraagd. Omdat grote AI-modellen *zero-shot learners* zijn, kunnen ze reageren op de richting die wij geven.

Zoals kunstenaar en schrijver James Bridle (2018) schrijft: 'Computers are not here to give us answers, but are tools for asking questions.'

Ontwikkelen van verantwoorde technologie

Voor ontwerpers en ontwikkelaars van AI-systemen ligt hier een grote verantwoordelijkheid. Allereerst om technologie zo verantwoord mogelijk te maken: vooroordelen aanpakken, ontwerpen voor inclusie, privacy waarborgen en energieverbruik beperken. Elke keuze in een ontwikkeltraject bepaalt welke waarden een technologie belichaamt.

Om bewuste, zorgzame omgang met AI mogelijk te maken, is het belangrijk om hier transparant over te zijn: welke data zijn gebruikt om een model te trainen, welke alignmentkeuzes zijn gemaakt, welke systeemprompt krijgt een taalmodel mee?[14] Open broncode helpt om dit inzichtelijk te maken.

Maar dit is niet voldoende. De fundamentele onbetrouwbaarheid van AI maakt dat ontwerpers AI-systemen zo vorm moeten geven dat bewuste, morele omgang eenvoudiger wordt. Een taalmodel kan bijvoorbeeld onzekerheid tonen, bronnen geven of vragen stellen. De in 2024 geïntroduceerde *Study Mode* van ChatGPT laat zien dat dit kan: een model als studiecoach die uitleg geeft en testvragen stelt. Ook dit roept morele vragen op, maar het toont dat finetuning richting andere vormen van gebruik mogelijk is.

Dit vraagt om onderzoek naar en experimenteren met nieuwe vormen van interactie tussen mensen, niet-mensen en AI-systemen. Interactievormen die niet gericht zijn op gemak en efficiëntie, maar die recht doen aan de relaties binnen een praktijk. De neiging van AI naar gemiddelden kan bijvoorbeeld worden doorbroken door juist pluraliteit en ambiguïteit te stimuleren. Daarbij hoort aandacht voor de relatie tussen systeem en mensen. Hoeveel autonomie heeft de gebruiker? Wie is eindverantwoordelijk? Dit wordt vaak afgedaan onder de generieke noemer human-in-the-loop, maar hoe zit die mens in de loop? Is het mens of mensen? Wie zijn die mensen eigenlijk? Worden zij echt betrokken

14 Zie bijvoorbeeld: https://docs.anthropic.com/en/release-notes/system-prompts#august-5-2025. Anthropic is een van de weinige makers van grote taalmodellen die een deel van hun systeemprompt openbaar maken.

of keuren ze slechts besluiten af die door het systeem al zijn voorgekookt? En waar zijn de niet-mensen in deze loop?

Langs deze lijn gedacht zijn er ook voor agentic AI verschillende paden denkbaar. Systemen zoals Deep Research werken nu vaak als onafhankelijke individuen: ze krijgen een opdracht, redeneren enkele minuten en presenteren een resultaat. Een meer collaboratieve benadering zou de gebruiker tussentijds betrekken: gebruikte zoektermen tonen en laten aanpassen, gevonden artikelen samenvatten, eerst een aanpasbaar plan maken in plaats van direct een eindproduct opleveren.

Hetzelfde geldt voor generatieve AI-applicaties. Zoals het onderzoeksproject Duidelijke TAAL laat zien, kan automatisch vereenvoudigen van tekst waardevol zijn, maar is het beter als het systeem gebruikers de mogelijkheid geeft deze te corrigeren of om aan te geven wat onveranderd moet blijven. Verantwoorde technologie is *bescheiden technologie*: het tegenovergestelde van superintelligentie. Systemen die beperkingen zichtbaar maken, ruimte bieden voor aanpassing en samenwerking, en die zich terugtrekken zodra de taak voltooid is (Vallor, 2024).

Voor nu zijn dit slechts aanzetten. Om te leren hoe we AI goed ontwerpen, zullen we moeten experimenteren en onderzoeken, maar hier valt iets te kiezen. Door uit te gaan van de onbetrouwbaarheid van AI komen we tot andere oplossingen dan wanneer we het zien als een vorm van intelligentie die steeds maar beter wordt. Daarbij is het belangrijk om niet enkel AI-systemen beter te ontwerpen, maar om een praktijk als geheel te beschouwen en te erkennen dat effecten van AI onvoorspelbaar zijn om te komen tot AI die verantwoord gebeurt.

Digitale technologie

In deze rede staan AI-systemen gebaseerd op machine learning, vooral taalmodellen, centraal. AI is een vorm van digitale technologie, en veel eigenschappen van AI gelden net zo goed voor andere digitale systemen. Technologie is nooit neutraal en een groot deel van de digitale middelen die we dagelijks gebruiken is in handen van een beperkt aantal bedrijven.

Denk aan het internet en de mobiele telefoon: ze zijn onlosmakelijk verweven met ons dagelijks leven. Daardoor rijzen steeds vragen: wie heeft toegang, wie

doet mee, hoeveel sturen deze technologieën ons wereldbeeld en onze keuzes, en hoeveel van onze aandacht eisen ze op?

Fundamentele onbetrouwbaarheid is kenmerkend voor AI, maar ook andere digitale systemen zijn verre van doorzichtig. Een groot softwarepakket dat al decennia door verschillende generaties ontwikkelaars is aangepast, is vaak nauwelijks beter te begrijpen.

De vraag blijft steeds dezelfde: hoe ontwikkelen en gebruiken we technologie verantwoord? Steeds weer moeten technische, economische, maatschappelijke en ethische aspecten samen gedacht worden.

Daarom is zorg, aandacht en verantwoordelijkheid essentieel voor alle digitale technologie, bij ontwikkelen en gebruik. Het gaat erom welke relaties technologie mogelijk maakt of juist verdringt. Dat is wat centraal staat in het onderzoek van het lectoraat Responsible IT.

Het lectoraat Responsible IT

Het lectoraat Responsible IT is een samenwerking van de faculteit Digitale Media en Creatieve Industrie van de Hogeschool van Amsterdam (HvA) en de gemeente Amsterdam.

We onderzoeken hoe digitale technologie verantwoord kan worden vormgeven en hoe we op een verantwoorde manier met digitale technologie om kunnen gaan. Uitgangspunt is dat technologie nooit neutraal is, maar moreel geladen: ze beïnvloedt het handelingsperspectief van mensen en niet-mensen. Daarom moeten technische, maatschappelijke en ethische overwegingen steeds samen worden meegenomen.

Technologie staat nooit op zichzelf, maar is situationeel. Wat een technologie is, staat daarbij niet volledig vast en de effecten van technologie zijn niet volledig voorspelbaar. Verantwoord is zodoende geen eindtoestand, maar een werkwoord.

Daarom werkt Responsible IT praktijkgericht, aan concrete vraagstukken, in nauwe samenwerking met burgers, programmeurs, ontwerpers, activisten, professionals, ambtenaren, studenten en docenten.

Amsterdam vormt daarbij een belangrijke context. De stad is thuisbasis van veel AI-bedrijven en maatschappelijke organisaties en kent grote culturele diversiteit. Het is ook een bijzondere stad, met een lange traditie rond verantwoorde digitalisering. Daarmee is Amsterdam bij uitstek een 'living lab' voor onderzoek naar verantwoorde digitale technologie in direct contact met bewoners en bezoekers.

Zo draagt het lectoraat bij aan de doelstellingen uit de Agenda Digitale Stad van de gemeente:[15] 'Amsterdam heeft zijn eigen identiteit en dat willen we er ook in de digitale wereld laten zijn. We zetten ons in voor een (digitale) stad waarin zo veel mogelijk Amsterdammers kunnen profiteren van de mogelijkheden. Een stad waarin technologie voor de Amsterdammer werkt en niet andersom. [...].

15 https://www.amsterdam.nl/innovatie/agenda-digitale-stad/

Digitaal op z'n Amsterdams is inclusief en betrouwbaar, maar bovenal menselijk. Niet digitaliseren omdat het kan, maar daar waar het moet.' Oftewel, een verantwoorde, bewuste omgang met technologie.

Aanpak en onderzoekslijnen

We bestuderen digitale technologie en de ethische omgang met deze technologie in concrete praktijken, samen met partners. Dit gebeurt vaak via ontwerpend onderzoek: we ontwikkelen oplossingen, richtlijnen of protocollen die we in de praktijk beproeven om zo de technologie en de wijze waarop deze de praktijk verandert te bevragen. We leren door te experimenteren. Daarmee biedt het lectoraat een aanvullend perspectief naast academisch IT-onderzoek en onderzoek uit de private sector.

Ons onderzoek valt uiteen in drie lijnen:

1. Duiding en demystificatie van digitale technologie en de impact van technologie.
2. Gereedschappen en methoden voor verantwoorde omgang met digitale technologie.
3. Ontwikkelen van verantwoorde digitale technologie.

Duiding en demystificatie

We onderzoeken hoe technologie, zoals generatieve AI, in de praktijk wordt gebruikt en welke effecten dat heeft. Tegelijk maken we voor een breed publiek inzichtelijk wat digitale technologie wel en niet kan.

Voorbeelden:

– Met de provincie Noord-Holland onderzochten we hoe medewerkers generatieve AI inzetten, en vertaalden dat in richtlijnen.
– In het project *Als de Machine Kiest* brachten we de rol van AI bij werving en selectie in kaart, met speciale aandacht voor diversiteit en inclusie. Conclusie: wie wordt aangenomen, dat beslist meestal een mens; wie wordt afgewezen, vaak al een algoritme.
– Met de *Bubble Machine*, een interactieve simulatie van sociale media, lieten we zien hoe aanbevelingsalgoritmes eenzijdige wereldbeelden versterken.

– De *Consensusmachine* toont hoe generatieve AI groepen kan ondersteunen bij consensusvorming ten aanzien van een complex vraagstuk. Het roept discussie op over de verdeling van verantwoordelijkheden tussen mens en machine. Met deze installatie zijn we het gesprek aangegaan met Amsterdammers op verschillende plekken in de stad over de werking en wenselijkheid van AI.

De komende jaren willen we nog beter begrijpen hoe mensen omgaan met generatieve AI en hoe we AI-geletterheid kunnen vergroten. Daarbij gaat het niet alleen om kennis van de technologie, maar ook om inzicht in de maatschappelijke en ethische dimensies ervan.

Gereedschappen en methoden

We ontwikkelen handvatten voor makers en gebruikers om bewust en verantwoord met digitale technologie om te gaan. De vraag die hier centraal staat is: hoe maken we de ethische en technische eigenschappen van digitale technologie inzichtelijk, en waar mogelijk meetbaar? We doen onderzoek naar gereedschappen om technologie te monitoren en om het morele gesprek te faciliteren. De komende jaren ligt de nadruk op de vraag hoe dit morele gesprek vorm kan krijgen en kan worden ingebed in organisaties en in de maatschappij.

Voorbeelden:
– Met de gemeente Amsterdam onderzoeken we hoe morele gesprekken over algoritmische vooroordelen vormgegeven kunnen worden.
– Met de provincie Noord-Holland ontwikkelden we het kaartspel *Moralis Machina*, dat overheidsorganisaties helpt om risico's en mogelijkheden van generatieve AI te doordenken.
– Met RTL maakten we een testtool voor spraakherkenning, die prestaties per bevolkingsgroep inzichtelijk maakt. Zo kunnen organisaties kiezen voor inclusievere systemen.
– Met KPN, Braincreators en Bluefield Agency bouwden we een dashboard dat ontwikkelaars helpt om AI-modellen niet alleen te vergelijken op nauwkeurigheid, maar ook op eerlijkheid en energieverbruik.

Ontwikkelen van verantwoorde digitale technologie

Binnen deze lijn onderzoeken we hoe digitale technologie zo verantwoord mogelijk vormgegeven kan worden. Hoe ontwerpen we digitale technologie zo dat deze handelingsruimte vergroot en een bewuste en verantwoorde omgang mogelijk maakt? De belangrijkste vraag voor de komende jaren is hoe we verantwoorde AI kunnen ontwikkelen vanuit de aanname dat AI een fundamenteel onbetrouwbare technologie is. Specifiek kijken we daarbij naar de verdeling van verantwoordelijkheden tussen mens en machine. Human-in-the-loop mag niet betekenen dat een mens slechts een automatisch besluit afvinkt, maar moet leiden tot echte expertise en verantwoordelijkheid in het proces.

Voorbeelden:

- Met Nemo Kennislink, de Universiteit van Amsterdam (UvA) en het Centrum voor Wiskunde en Informatica (CWI) experimenteren we binnen het AI, Media & Democracy Lab met nieuwsaanbevelingen die diversiteit bevorderen in plaats van filterbubbels versterken.
- In Duidelijke TAAL ontwikkelen we samen met gemeenten, provincies en de hogescholen van Rotterdam en Utrecht AI-prototypes voor het vereenvoudigen van overheidscommunicatie. Daarbij staat de vraag centraal hoe ambtenaren eigenaarschap over teksten behouden.
- De installatie *Hitloop* gebruikt muziek als metafoor voor AI-besluitvorming en laat bezoekers ervaren hoe menselijk toezicht kan werken: mensen sturen AI-gegenereerde muziek gezamenlijk met hun bewegingen op de dansvloer.

Samenwerking

Praktijkgericht onderzoek doe je nooit alleen. Responsible IT werkt samen met overheden (o.a. provincie Noord-Holland, gemeenten Amsterdam, Utrecht, Hilversum, Rotterdam, ministerie van Binnenlandse Zaken en Koninkrijksrelaties), culturele instellingen (Nationaal Archief, Koninklijke Bibliotheek, Beeld & Geluid, Nemo), mediapartners (NPO, RTL), bedrijven (KPN, Braincreators, Info.nl, Bluefield) en kennisinstellingen (Hogeschool Rotterdam, Hogeschool Utrecht, Universiteit van Amsterdam, Centrum voor Wiskunde en Informatica).

Het lectoraat is onderdeel van bredere verbanden zoals het HvA Centre of Expertise Applied AI, het *SPRONG-programma Responsible Applied AI* met

Hogeschool Rotterdam en Hogeschool Utrecht, en het *AI, Media & Democracy Lab* waarin we samenwerken met UvA, CWI en mediapartners om de impact van AI op de mediasector en democratische processen te onderzoeken.

De beste manier om tot een verantwoorde omgang met technologie te komen, is door jonge mensen te leren hoe dit te doen. En door te leren van de frisse blik van deze mensen. Het lectoraat Responsible IT is daarom nauw verweven met het onderwijs. De docent-onderzoekers van het lectoraat zijn onder andere verbonden aan de master 'Applied AI' en de HBO-ICT studio 'Responsible Applied AI' van de Hogeschool van Amsterdam. Het lectoraat was mede-initiatiefnemer van de brede minoren 'Desinformatie', 'Het internet is stuk, maar we gaan het repareren' en 'AI voor creatieve denkers en ontwerpers'. En studenten en docenten van de opleidingen HBO-ICT, CMD en de master 'Applied AI' leveren cruciale bijdragen aan het onderzoek van Responsible IT.

Dankwoord

Op dit punt aangekomen wil ik graag mijn dank uitspreken aan iedereen die heeft bijgedragen aan de totstandkoming van deze rede en aan mijn professionele ontwikkeling die hieraan vooraf is gegaan.

Allereerst gaat mijn dank uit naar het college van bestuur van de HvA voor het in mij gestelde vertrouwen. In het bijzonder wil ik rector Geleyn Meijer bedanken, die de opkomst van AI vroeg in de gaten had en mij de ruimte bood om dat onderwerp binnen de HvA een plek te geven.

Mijn bijzondere dank gaat ook uit naar Frank Kresin, decaan van de faculteit Digitale Media en Creatieve Industrie (FDMCI), die sterk betrokken is bij het onderwerp verantwoorde digitalisering. Frank, ik waardeer onze samenwerking enorm. Jouw niet-aflatende positieve aanpakkersmentaliteit is een bron van inspiratie, en het vertrouwen dat je in mij stelt maakt dat ik, zoals jij het noemt, een 'serial entrepreneur' binnen de HvA kan zijn.

Het lectoraat Reponsible IT zou niet mogelijk zijn zonder de steun van de gemeente Amsterdam. Mijn dank gaat uit naar Ger Baron en wethouder Alexander Scholtes, die het belang van verantwoorde digitalisering hoog in het vaandel hebben en het praktijkgericht onderzoek op dit gebied een warm hart toedragen en die het lectoraat een lab ter grootte van een stad bieden.

Ook Nanda Piersma wil ik graag bijzonder bedanken. In onze gezamenlijke avonturen om praktijkgerichte AI een plek te geven, heb ik veel van je geleerd. De steun van het Centre of Expertise Applied AI is onontbeerlijk. Geert, Liza en Susanna enorm bedankt voor de manier waarop jullie initiatieven ontplooien en bij de wildste ideeën steun bieden. Of we nu binnen drie weken een ChatGPT-evenement uit de grond willen stampen of een AI-rave op het dak van Nemo willen houden, jullie staan klaar.

Graag bedank ik ook André Neumann – hoofd kenniscentrum van FDMCI – voor onze gesprekken, jouw rust en jouw managementervaring waar ik op kan bouwen. Ook Marleen Stikker en Sander van der Waal wil ik bedanken voor de inspirerende samenwerking in de afgelopen jaren. En mijn dank gaat uit

naar Somaya Ben Allouch, Tamara Witsche, Troy Nachtigall, Martijn de Waal, Sabine Niederer, Geert Lovink en Marten Teitsma, mijn medelectoren binnen de faculteit FDMCI, die laten zien hoe ontwerpgedreven praktijkgericht onderzoek eruit kan zien. Ik wil Ben Kröse en Martijn Vastenburg bedanken, van wie ik veel leerde over het vormgeven en draaiend houden van een onderzoeksgroep.

Maaike Harbers, Marieke Peeters, Huib Aldewereld en Stefan Leijnen, hoewel we voor verschillende instituten werken, is onze samenwerking zo intensief dat jullie als directe collega's voelen. Regelmatig zijn jullie mijn sparringpartners bij het opzetten van een nieuw onderzoek of het doordenken van wat verantwoorde AI is. Heel veel dank voor de afgelopen jaren en op naar de volgende. Ook de rest van het team van het Responsible Applied AI-programma wil ik hartelijk danken voor onze samenwerking en de mooie inzichten en producten die daaruit voortkomen. Natali Helberger, Claes de Vreese, Pablo Cesar, Han La Poutré en natuurlijk Sara Spaargaren wil ik bedanken voor onze inspirerende samenwerking binnen het AI, Media & Democracy ELSA-lab. Mijn dank gaat ook uit naar de vele praktijkpartners waar Responsible IT mee samenwerkt.

Responsible IT zou er niet zijn zonder Bas Pijls-Van Kooten, Coline Pannier, David van Dijk, Hans de Zwart, Leontine Born, Lisanne Henzen, Maarten Groen, Marcio Fuckner, Marzin Magzidak, Merel Noorman, Michiel Bontenbal, Saskia Robben, Sophie Horsman, Suzana Basic, Renushka Madarie, Rick van Kersbergen, Rik van Eijk, Rover Vos en Yuri Westplat. Ook Sonja Rouwhorst, Marise van Noordenne en Stijn Oomes horen in dat rijtje thuis. Jullie geven het onderzoek van Responsible IT en de sfeer in het lectoraat vorm. Ik prijs me gelukkig om met zo'n team te mogen werken. Onze vele gesprekken, zoals de discussies in de boekenclub, waren een belangrijke voedingsbodem voor deze rede. Bedankt!

Ik wil mijn ouders Paul en Diny bedanken voor hun onvoorwaardelijke steun bij de keuzes die ik maak, en mijn zussen Corine en Danielle voor de nodige relativering. Bedankt dat jullie de afgelopen maanden samen zorgden dat ik met enige regelmaat de rust had om te denken en te schijven. Ik wil mijn schoonvader Arie-Jan bedanken voor alles wat hij voor onze jongens en ons gezin doet.

Tot slot, het onderwerp verantwoorde technologie gaat mij nauw aan het hart en ik ben trots dat ik de positie van lector Responsible IT mag vervullen, maar

dat zou niet mogelijk zijn zonder de steun en liefde van mijn thuisfront, die mij steeds weer doen beseffen wat echt belangrijk is. Eline, jij bent mijn anker en houvast. Olivier, Tobias en Florian – het meest AI-kritische publiek dat ik ken – jullie zijn de reden dat het zo belangrijk is om de toekomst verantwoord vorm te geven.

Referenties

Hans Achterhuis (ed.). 2001. *American Philosophy of Technology: The Emperical Turn*. Indiana University Press.

Laurens van Aggelen. 2025. Meta haalt jonge CEO van Scala AI binnen met miljardeninvestering. *ICT/Magazine.* Geraadpleegd op: 13 juni 2025, van: https://www.ictmagazine.nl/nieuws/meta-haalt-jonge-ceo-van-scale-ai-binnen-na-miljardeninvestering

Dario Amodei. 2024. *Machines of Loving Grace, How AI Could Transform the World for the Better*. Geraadpleegd op: 20 januari 2025, van https://www.darioamodei.com/essay/machines-of-loving-grace

Sam Altman. 2025. *The Gentle Singularity*. Geraadpleegd op 24 augustus 2025, van: https://blog.samaltman.com/the-gentle-singularity

Hannah Arendt. 2018. *The Human Condition* (2nd edition). The University of Chicago Press.

Soe-Fa Bakboord, Debbie Dekkers, Oumaima Nedoui, Bart Burkhardt, Wibold Jongsma, Charlie Helmijr, Lianne Nelemans. Z.d. *Digital Urban Planning*. https://openresearch.amsterdam/nl/page/41686/digital-urban-planning

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, ... Jared Kaplan. 2022. *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. https://arxiv.org/abs/2204.05862

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 03, 2021. ACM, Virtual Event Canada, 610–623.

Sam Bogerd & Daan Juijn. 2025. De overheid moet zich voorbereiden op AI die superintelligent is. *NRC*. Geraadpleegd op: 12 augustus 2025, van: https://www.nrc.nl/nieuws/2025/08/12/de-overheid-moet-zich-voorbereiden-op-ai-die-superintelligent-is-a4902765

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama & Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in neural information processing systems*, *29*.

James Bridle. 2018. *New Dark Age, Technology and the End of the Future*. Verso Books.

James Bridle. 2023. The Stupidity of AI. *The Guardian*. Geraadpleegd op: 22 augustus 2025, van: https://www.theguardian.com/technology/2023/mar/16/the-stupidity-of-ai-artificial-intelligence-dall-e-chatgpt

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, ... Dario Amodei. 2020. Language Models are Few-Shot Learners. In H. Larochelle,

M. Ranzato, R. Hadsell, M.F. Balcan & H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Joy Buolamwini. 2023. *Unmasking AI*. Penguin Random House.

Nicholas Carlini. 2023. Playing Chess with large language models. Geraadpleegd op 24 augustus 2024, van: https://nicholas.carlini.com/writing/2023/chess-llm.html

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm & Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 10 augustus 2015. ACM, Sydney NSW Australia, 1721–1730.

Lyndie Chiou & Clara Moskowitz. 2025. *At Secret Math Meeting, Researchers Struggle to Outsmart AI*. Geraadpleegd op 25 augustus 2025, van https://www.scientificamerican.com/article/inside-the-secret-meeting-where-mathematicians-struggled-to-outsmart-ai/

Cynthia Cook, Laurie Cuffney, John Howard, Yevgeniy Sirotin, Jerry Tipton, Arun Vermury. 2025. Performance Differentials in Deployed Biometric Systems Cause by Open-Source Face Detectors, Proceedings of FAccT '25'.

Kate Crawford. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

DeepSeek-AI et al. 2025. *Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* https://arxiv.org/abs/2501.12948

Tim Demedts. 2025. Drie argumenten om te weigeren AI te gebruiken. *Trouw*. Geraadpleegd op: 24 agustus 2025, van https://www.trouw.nl/opinie/opinie-drie-argumenten-om-te-weigeren-ai-te-gebruiken~b373b437/

Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 4171–4186.

Thomas Dohmke. 2025. *Developers, Reinvented*. Geraadpleegd 4 augustus 2025, 10, van: https://ashtom.github.io/developers-reinvented

Thijs Doorenbosch. 2023. ChatGPT is de snelstgroeiende app ooit. *AGConnect*. Geraadpleegd op 24 augustus 2025, https://www.agconnect.nl/tech-en-toekomst/artificial-intelligence/chatgpt-is-de-snelstgroeiende-app-ooit

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit & Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Arxiv. http://arxiv.org/abs/2010.11929

Hubert L. Dreyfus. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason* (revised Edition). The MIT Press.

Lisa Feldman Barrett, et al. 2019, Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1).

Jason Fiawo, Maarten Sukel, Lianne Nelemans. 2022. AI wordt ingezet voor verbetering van toegankelijkheid in Amsterdam. Geraadpleegd op 22 augustus 2025, https://openresearch.amsterdam/nl/page/81876/ai-wordt-ingezet-voor-de-verbetering-van-toegankelijkheid-in-amsterdam

Gemini Team. 2025. *Gemini: A Family of Highly Capable Multimodal Models.* https://doi.org/10.48550/arXiv.2312.11805

Donna Haraway. 2016. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press.

Maaike Harbers, Sophie Horsman, Pascal Wiggers, Huib Aldewereld, Marcio Fuckner, Coert Van Gemeren, Oumaima Hajri, Floor Schukking, Nathalie Stembert. 2025. Responsible AI in Media Organizations: Four Case Studies Implementing Ethical Tools in Practice. *In proceedings of AIES 2025*

Martin Heidegger. 2006. *Sein und Zeit* (neunzehnte Auflage). Max Niemeyer Verlag Tübingen.

Martin Heidegger. 2014. *De vraag naar de techniek* (vertaling Mark Wildschut). Uitgeverij VanTilt.

Sophie Horsman, Martijn Veerman, Jeroen Silvis, Pascal Wiggers. 2025. Empowering Civil Servants: Fostering Ethical Reflection on GenAI Use through Experimentation and a Serious Game. *In proceedings of the Extended Abstacts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25

Marco Iacoboni. 2010. *Het spiegelende brein, over inlevingsvermogen, imitatiegedrag en spiegelneuronen* (derde oplage). Uitgeverij Nieuwezijds.

Steven Johnson. 2024. *You Exist in The Long Context*. Geraadpleegd op 24 augustus 2025 van https://thelongcontext.com

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, ... Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 7873 (August 2021), 583–589.

Daniel Jurafsky & James H. Martin. 2014. *Speech and Language Processing*, 2nd Edition. Pearson.

Kaggle blog. 2025. *Introducing Kaggle Game Arena*. Geraadpleegd op 24 augustus 2025, van https://www.kaggle.com/blog/introducing-game-arena

KNMI. 2023. *Betere lange termijnverwachtingen voor warm weer dankzij AI*. Geraadpleegd op 24 augustus 2025, van: https://www.knmi.nl/over-het-knmi/nieuws/betere-lange-termijnverwachtingen-voor-warm-weer-dankzij-ai

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo & Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in neural information processing systems*, 35: 22199-22213.

Golden Krishna. 2015. *The Best Interface is No Interface: The Simple Path to Brilliant Technology*. New Riders Pub.

Olya Kudina, Peter-Paul Verbeek. 2019. Ethics from Within: Google Glass, the Collingride Dilemma and the Mediated Value of Privacy, Science, *Technology & Human Values*, Vol 44(2), 291-314

Joris van Laarhoven & Eva van Vugt. 2025. Lever de rechtstaat niet uit aan ChatGPT en verbied AI aan de juridische faculteit. *NRC*. Geraadpleegd op: 2 juli 2025, van https://www.nrc.nl/nieuws/2025/07/02/lever-de-rechtsstaat-niet-uit-aan-chatgpt-verbied-ai-aan-de-juridische-faculteit-a4899056

George Lakoff & Mark Johnson. 1980. *The Metaphors we Live By*. University of Chicago Press.

Mark Lawrence. 2025. *The AI vs. authors results! (part 2)*. Geraadpleegd op 20 augustus 2025, van: https://mark---lawrence.blogspot.com/2025/08/the-ai-vs-authors-results-part-2.html

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, … Joshua Batson. 2025. *On the Biology of a Large Language Model*. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Nelson Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters & Noah Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North*, 2019. Association for Computational Linguistics, Minneapolis, Minnesota, 1073–1094.

Rebecca Lowe. 2024. *Why you should be talking with gpt about philosophy*. Geraadpleegd op 24 augustus 2025, van https://theendsdontjustifythemeans.substack.com/p/why-you-should-be-talking-with-gpt

Alasdair MacIntyre. 2024. *Na de Deugd – Een studie naar de theorie van ethiek.* Ten Have.

Gary Marcus. 2025. *GPT-5: Overdue, overhyped and underwhelming. And that's not the worst of it.* Geraadpleegd op 11 augustus 2025, van: https://garymarcus.substack.com/p/gpt-5-overdue-overhyped-and-underwhelming

Maurits Martijn. 2024. Raken we door kunstmatige intelligentie allemaal onze baan kwijt? *de Correspondent*. Geraadpleegd op 4 juli 2025, van: https://decorrespondent.nl/15211/raken-we-door-kunstmatige-intelligentie-allemaal-onze-baan-kwijt/058fc7a1-bb82-08ba-3952-59c6d83e3d65

Corijn van Mazijk. 2021. *De wereld als verschijning – Fenomenologie en de twintigste eeuw*. Boom, 2021.

Eveline Meijer. 2025. Microsoft: 'Tot 30 procent van onze nieuwe code is door AI gegenereerd'. *Tweakers*. Geraadpleegd op: 10 mei 2025, van: https://tweakers.net/nieuws/234508/microsoft-tot-30-procent-van-onze-nieuwe-code-is-door-ai-gegenereerd.html

Meta AI. 2025. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

Melanie Mitchell. 2019. *Artificial Intelligence, A Guide for Thinking Humans*. Penguin Random House UK.

Ethan Mollick. 2024. *Co-Intelligentie, Slimmer werken met AI*. POM Press.

Ethan Mollick. 2025a. *The End of Search, The Beginning of Research*. Geraadpleegd op 24 augustus 2025, van: https://www.oneusefulthing.org/p/the-end-of-search-the-beginning-of

Ethan Mollick. 2025b. *Centaurs and Cyborgs on the Jagged Frontier*. Geraadpleegd op: 3 juli 2025, van: https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged

Timothy Morton. 2018. *Being Ecological*. The MIT Press.

Farida ben Moussa. 2025. AI kan juist bijdragen aan eerlijke(re) kansen. *de Volkskrant*. Geraadpleegd op: 1 juli 2025, van https://www.volkskrant.nl/columns-opinie/opinie-ai-kan-juist-bijdragen-aan-eerlijke-re-kansen~b783d2ab/

Madhumita Murgia. 2024. *In de schaduw van AI – Leven in een algoritmische tijd*. Business Contact.

Ilyaz Nazrullah. 2023. Val niet voor de hype, ChatGPT is echt heel dom. *Trouw*. Geraadpleegd op: 24 augustus 2025, van https://www.trouw.nl/opinie/val-niet-voor-de-hype-chatgpt-is-echt-heel-dom~b37aeedd/

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, … Chris Olah. 2022. *In-context Learning and Induction Heads*. https://arxiv.org/abs/2209.11895

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threathens Democracy*. Penguin books.

Rutger Otto. 2025. Schaakgrootmeesters hebben niets te vrezen van ChatGPT en Grok. *nu.nl*. Geraadpleegd op 9 augustus 2025, van: https://www.nu.nl/tech/6365061/schaakgrootmeesters-hebben-niets-te-vrezen-van-chatgpt-en-grok.html

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, … Ryan Lowe. 2022. *Training language models to follow instructions with human feedback.* https://doi.org/10.48550/arXiv.2203.02155

Dwarkesh Patel. 2025. *The Scaling Era, An Oral History of AI, 2019-2025.* Stripe Press.

Ernst-Jan Pfauth. 2025. Waarom ik niet meer zonder mijn altijd aanwezige AI-gesprekspartner wil. *de Correspondent*. Geraadpleegd op 20 mei 2025, van https://decorrespondent.nl/16096/waarom-ik-niet-meer-zonder-mijn-altijd-aanwezige-ai-gesprekspartner-wil/bbf6ad68-1251-0774-17d2-62519fce88f9

Steven Pinker. 2007. *The Stuff of Thought*. Allen Lane.

Ibo van de Poel. 2011. Nuclear Energy as a Social Experiment. *Ethics, Policy & Environment* 14 (3): 285-90.

Ibo van de Poel. 2016. "An Ethical Framework for Evaluating Experimental Technology." *Science and Engineering Ethics* 22 (3): 667-86.

Jelle Prins, Michiel Bakker, Max Welling, Durk Kingma, Rick Pastoor, Stan van Baarsen & Renée Frissen. 2025. Voorkom dat Nederland een digitale kolonie wordt en investeer meer in AI. *de Volkskrant.* Geraadpleegd op: 8 juli 2025, van https://www.volkskrant.nl/columns-opinie/opinie-voorkom-dat-nederland-een-digitale-kolonie-wordt-en-investeer-meer-in-ai~bcdaf33a/

Jelle Prins, Michiel Bakker, Renée Frissen & Stan van Baarsen. 2025. AI vraagt om politieke keuzes – nu. AI Plan, https://aiplan.nl/openbrief

Mirjam Rasch. 2020. *Frictie: Ethiek in tijden van dataïsme*. De Bezige Bij.

Max van Rest. 2023. *Stedelijke boomanalyse met behulp van 3D-puntenwolk*. Geraadpleegd op: 12 september 2025, van: https://openresearch.amsterdam/nl/page/125084/stedelijke-boomanalyse-met-behulp-van-3d-puntwolken

Adi Robertson. 2024. Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis. *The Verge*. Geraadpleegd op: 24 augustus 2025, van: https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

Michael Rosen & Helen Oxenbury. 2019. *Wij gaan op berenjacht* (46ste druk). Gottmer.

Joshua Rothman. 2025. Are we taking A.I. Seriously enough? *The New Yorker*. Geraadpleegd op 24 augustus 2025, van https://www.newyorker.com/culture/open-questions/are-we-taking-ai-seriously-enough

Erdinç Saçan & Barend Last. 2025. Wie alle AI op een hoop gooit, mist de potentie ervan. *Trouw*, Geraadpleegd op: 24 augustus 2025, van: https://www.trouw.nl/opinie/opinie-wie-alle-ai-op-een-hoop-gooit-mist-de-potentie-ervan~b72332ef

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda & Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems,* 36: 68539-68551.

Douwe Schmidt & Milou Jansen. 2023. *De Ethische Bijsluiter*. Geraadpleegd op 24 augustus 2025, van: https://openresearch.amsterdam/nl/page/99984/de-ethische-bijsluiter

Douwe Schmidt & Diederik Roijers. 2025. *De Amsterdamse Visie op AI*. Geraadpleegd op 24 augustus 2025, van: https://openresearch.amsterdam/nl/page/118198/de-amsterdamse-visie-op-ai

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu & Deyi Xiong. 2023. *Large Language Model Alignment: A Survey*. Arxiv, https://doi.org/10.48550/arXiv.2309.15025

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio & Mehrdad Farajtabar. 2025. *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. https://arxiv.org/abs/2506.06941

Tomas Ptacek. 2025. *My AI Skeptic Friends Are All Nuts*. Geraadpleegd op 20 augustus 2025, van: https://fly.io/blog/youre-all-nuts/

Richard Sutton. 2019. *The Bitter Lesson*. Geraadpleegd op 24 augustus 2025 van http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Ronit van Tijn Palache. 2025. Mijn nachtelijke gesprek met ChatGPT bood verlichting. *Trouw*. Geraadpleegd op 3 augustus 2025, van https://www.trouw.nl/columnisten/mijn-nachtelijke-gesprek-met-chatgpt-bood-verlichting~b20d9c82/

Anna Lowenhaupt Tsing. 2015. *The Mushroom at the End of the World – On the Possibility of Life in Capitalist Ruins*. Princeton University Press.

Alan Turing. 1950. Computing Machinery and Intelligence. *Mind*, New Series. 59(236), 433-460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2023. *Attention is All you Need.* https://doi.org/10.48550/arXiv.1706.03762

Shannon Vallor. 2024. *The AI Mirror: How to Reclaim our Humanity in an Age of Machine Thinking*, Oxford University Press.

Peter-Paul Verbeek. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.

Peter-Paul Verbeek. 2014. *Op de vleugels van Icarus*. Lemniscaat.

Johannes Visser. 2025. Hoe een docent worstelt met ChatGPT. *de Correspondent.* Geraadpleegd op 12 september 2025, van: https://decorrespondent.nl/16048/hoe-een-docent-worstelt-met-chatgpt/503d1488-c7e7-0315-27b8-7bba1a357bfb

Sarah Volinsky-Fremond, Nanda Horeweg, Sonali Andani, Jurriaan Barkey Wolf, Maxime W. Lafarge, ... Tjalling Bosse. 2024. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nat Med* 30, 7 (July 2024), 1962–1973.

Angelina Wang, Sayash Kapoor, Solon Barocas, Arvind Narayanan. 2023. Against predictive optimization – On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *ACM Journal on Responsible Computing*, 1(1): 1-45, https://doi.org/10.1145/3636509

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean & William Fedus. 2022a. *Emergent Abilities of Large Language Models*. https://doi.org/10.48550/arXiv.2206.07682

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le & Denny Zhou. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35: 24824-24837.

Pascal Wiggers. 2008. *Modelling Context in Automatic Speech Recognition*. Proefschrift TU Delft.

Arnoud Wokke. 2025. Aantal Zoekopdrachten via Google in Safari daalt door gebruik van AI-chatbots. *Tweakers*. Geraadpleegd op 7 mei 2025, van https://tweakers.net/nieuws/234724/aantal-zoekopdrachten-via-google-in-safari-daalt-door-gebruik-van-ai-chatbots.html

Arnoud Wokke. 2025. Groot probleem met nieuwe AI-modellen: ze vertellen je meer onzin. *Tweakers*. Geraadpleegd op 11 mei 2025, van: https://tweakers.net/reviews/13236/groot-probleem-met-nieuwe-ai-modellen-ze-vertellen-je-meer-onzin.html

Adam Yedidia. 2023. *Chess as a case study in hidden capabilities in ChatGPT*. Geraadpleegd op 22 september 2025, van https://www.lesswrong.com/posts/F6vH6fr8ngo7csDdf/chess-as-a-case-study-in-hidden-capabilities-in-chatgpt

Hans de Zwart & Pascal Wiggers. 2022. *Als de machine kiest: diversiteit en inclusie in kunstmatige intelligentie voor werving en selectie*. Lectoraat Responsible IT, Hogeschool van Amsterdam.

Generatieve AI ontwikkelt zich in een razend tempo en dringt tot steeds meer domeinen van ons dagelijks leven door. Over de gevolgen verschillen de meningen. Waar de een AI ziet als een onmisbare sparringpartner, maken anderen zich zorgen over de impact op werk, menselijke creativiteit, denken en leren.

Dit roept de vraag op:
Hoe willen we omgaan met generatieve AI?

Een antwoord op die vraag begint met het doorgronden van wat AI is – en wat het níet is.

Pascal Wiggers betoogt dat AI een krachtige, maar niet-neutrale en fundamenteel onbetrouwbare technologie is: het maakt fouten, bevat vooroordelen en effecten zijn slechts beperkt voorspelbaar.

Dat is geen diskwalificatie, maar juist een uitgangspunt voor verantwoorde ontwikkeling en inzet van AI.
Het vraagt om een bewuste omgang met AI vanuit zorg en aandacht, als tegenhangers voor de snelheid van de technologie; hierbij staan de relaties tussen mensen, niet-mensen en technologieën in specifieke praktijken centraal.

Beeld: AdobeStock/bewerkt door textcetera

ISBN 9789463015813