AI ALS AANJAGER VAN VERANDERING

KVAB DENKERSRAPPORT 2023

KVAB Press

KVAB STANDPUNTEN
85

Concept cover: Francis Strauven
Ontwerp cover: Charlotte Dua
Afbeelding: Shutterstock

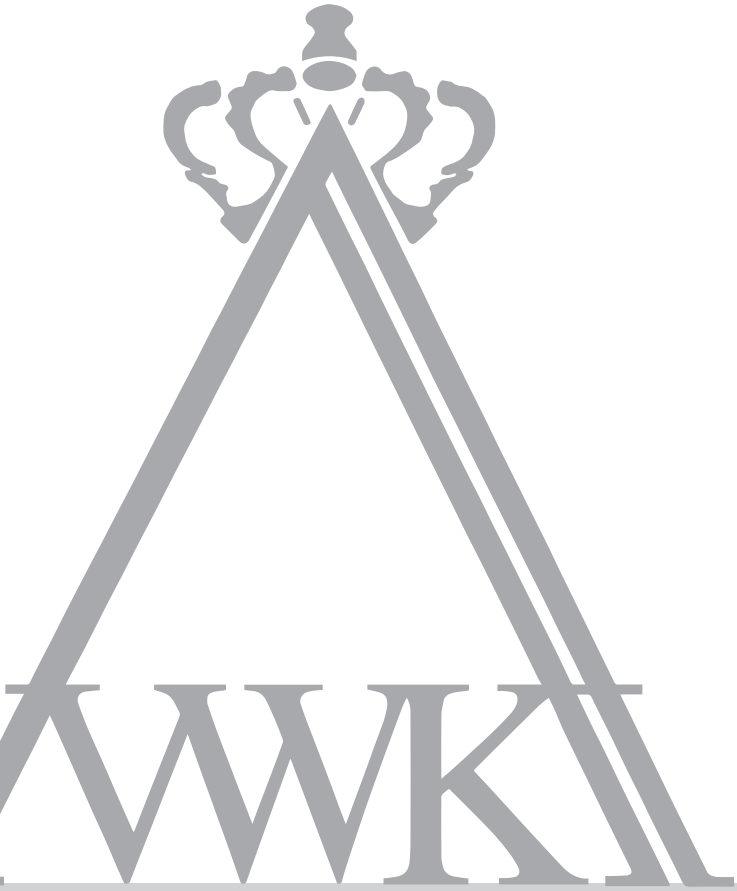
De tekening van het Paleis der Academiën is een reproductie van het originele perspectief van Charles Vander Straeten in 1823. Jozef Cantré ontwierp het logo van de KVAB in 1947.

De KVAB Standpunten worden gepubliceerd door de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Hertogsstraat 1, 1000 Brussel.
Tel. 00 32 2 550 23 23 – info@kvab.be – www.kvab.be

AI ALS AANJAGER VAN VERANDERING

KVAB DENKERSRAPPORT 2023

Helga Nowotny
Ine Van Hoyweghen
Joos Vandewalle



D/2023/0455/09
ISBN 978 90 656 922 76

Aanbevolen citeerwijze: Helga Nowotny, Ine Van Hoyweghen, Joos Vandewalle, *AI als aanjager van verandering, KVAB Denkersrapport 2023, KVAB* Standpunt 85, 2023.

Drukkerij Universa

AI ALS AANJAGER VAN VERANDERING

INHOUDSOPGAVE

Samenvatting

Dit Standpunt kwam tot stand in het kader van het Denkersprogramma van de KVAB. Op basis van uitgebreid overleg met stakeholders heeft de "Thinker-in-residence" inzichten geformuleerd over de huidige praktijk en de mogelijke toekomst van het Vlaamse AI onderzoek en de samenleving als geheel.

Hoewel Artificiële Intelligentie (AI) geen magisch instrument is, maar het simpele resultaat van wiskundige optimalisatie, enorme computerkracht en datasets, zijn haar toepassingen buitengewoon veelbelovend. Zij hebben al geleid tot aanzienlijke voordelen, met name in de vorm van een verbeterde efficiëntie, nauwkeurigheid, tijdigheid en gemak in een breed scala aan onderzoek en diensten, producten en processen, waaronder gezondheidszorg, genomica, fusieonderzoek, productontwerp en -simulatie, autonome systemen, voorspellend onderhoud, kwaliteitscontrole, inspectieprocessen en milieueffectbeoordeling.

Tegelijkertijd gaat de opkomst van AI gepaard met een toenemende bezorgdheid over de potentiële risico's en schadelijke effecten: voor individuen, voor kwetsbare groepen en voor de samenleving in het algemeen. Dit roept vragen op over hoe AI ons dagelijks leven zal beïnvloeden, zowel in de privé- als in de arbeidscontext, en hoe deze technologie onze kijk op wie we zijn als mens zal beïnvloeden. Dit vereist een mensgerichte aanpak bij het ontwerp, het gebruik en de verdere ontwikkeling van AI, wat inhoudt dat AI dient worden afgestemd op menselijke waarden en behoeften. We dienen technologieën te vormen in overeenstemming met menselijke waarden en behoeften, in plaats van toe te staan dat technologieën mensen vormen. Dit wordt met de dag dringender. Nu generatieve AI-systemen (bv. ChatGPT) zich razendsnel ontwikkelen, moet de wetenschappelijke gemeenschap een centrale rol spelen bij het vormgeven van de toekomst van een mensgerichte benadering van AI.

Het doel van de Denkerscyclus van de KVAB was om over de impact van deze nieuwste ontwikkelingen in AI te reflecteren, over de vraag wat de implicaties van AI voor ons mensbeeld zijn – voor onze autonomie, het menselijk handelingsvermogen ("agency") en de noodzaak van een "digitaal humanisme". De titel van de Denkerscyclus, "AI as an Agent of Change" ("AI als aanjager van verandering"), is geïnspireerd op het tweedelige werk van historica Elizabeth Eisenstein, *The Printing Press as an Agent of Change* ("De drukpers als aanjager van verandering", 1980).

Deze reflecties waren gericht op de nieuwste ontwikkelingen in AI door een breed filosofisch, sociologisch en historisch perspectief te bieden op de impact van AI. Op basis van de specifieke achtergronden, perspectieven en expertises van de betrokkenen presenteert dit Standpunt niet alleen een consistente analyse, maar ook waardevolle benaderingen en voorstellen voor verdere stappen. Gezien de

kwaliteit van de discussies van de Denker met de deelnemers en haar constructieve bevindingen, bieden deze inspanningen een solide basis voor de productieve integratie van AI in de wetenschap en de maatschappij in Vlaanderen.

Aanbeveling 1: **We bevelen aan een brede publiekscampagne te lanceren** onder het motto "AI voor burgers – burgers voor AI" om burgers te ondersteunen bij het gebruik van AI in hun dagelijks leven en voor een betere samenleving. Het doel is om het begrip van de werking van AI en digitale systemen te verdiepen en te verspreiden, het potentieel van huidige en toekomstige toepassingen en het gebruik ervan te onderzoeken, en te leren over mogelijke beperkingen.

Daartoe moet een solide institutioneel kader worden opgezet en voorzien van de nodige financiële en personele middelen, in eerste instantie voor een periode van drie jaar, en hernieuwbaar na evaluatie.

Aanbeveling 2: **We bevelen aan om fundamenteel AI-onderzoek een hoge prioriteit te geven** en uit te voeren volgens de lijnen van de Europese Onderzoeksraad (ERC) (bottom-up, uitgaand van een hoofdonderzoeker). Dit dient als tegenwicht voor de dominantie van een ééndimensionaal "technologisch oplossingsdenken", dat alternatieven negeert en/of terzijde schuift bij de keuze van AI onderzoeksproblemen, methoden en technieken. Hierdoor ontstaat bovendien een meer humanistisch begrip van de reikwijdte en de diepte van de menselijke ervaring en wat het betekent om mens te zijn.

Als onderzoeksgebied is AI, inclusief machine leren en generatieve AI, relatief jong, terwijl een historisch perspectief grotendeels ontbreekt, vooral in Europa. Hierdoor bestaat er een grote kans op het verlies van waardevolle technische kennis, wiskundige concepten, technieken en wetenschappelijke inzichten. Veelbelovende onderzoekslijnen werden vaak voortijdig afgesloten. Alleen een sterke focus op fundamenteel onderzoek kan de aanzet geven tot hun herontdekking en de verdere verkenning van historische paden die niet werden bewandeld.

Aanbeveling 3: **We bevelen een krachtige ondersteuning aan van onderzoek naar de maatschappelijke impact van AI wat betreft aspecten en gebieden die naar alle waarschijnlijkheid niet zullen worden opgepakt door de grote internationale bedrijven.**

Omdat we nog maar aan het begin staan van het systematisch volgen en analyseren van de mogelijke toepassingen van AI voor verschillende groepen in de samenleving en het leren vermijden van mogelijke sociale schade, is het cruciaal om de snel evoluerende ervaringen, stemmen en behoeften van burgers mee te nemen.

Studenten AI en aanverwante technische gebieden (en hun docenten) moeten worden aangemoedigd om een perspectief van digitaal humanisme op te nemen in

hun technische opleiding en praktijk. Ook studenten in de geesteswetenschappen en sociale wetenschappen (en hun docenten) moeten meer vertrouwd raken met de technische aspecten. Dit zijn de voorwaarden voor een meer en beter gefundeerde inter- en zelfs transdisciplinariteit, die dringend nodig is.

Voorwoord

Artificial Intelligence
human company
invisible algorithms
future needs wisdom

(Nowotny, 2021)

Elk jaar organiseert de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten (KVAB) binnen het Denkersprogramma twee zogenaamde Denkerscycli op initiatief van een van haar Klassen en/of Reflectiegroepen. In elke Cyclus maakt de uitgenodigde Denker kennis met de specifieke kenmerken van een bepaalde maatschappelijke uitdaging in Vlaanderen. Deze Cycli resulteren in Standpunten waarvan de meningen en bevindingen kunnen worden opgenomen in beleidsaanbevelingen die door de Europese Federatie van Academies worden opgesteld in het kader van SAPEA (Science Advice for Policy by the European Academies; https://www.sapea.info).

In 2022 initieerde de Klasse van de Menswetenschappen (KMW) de Denkerscyclus "AI als aanjager van verandering" om de impact van artificiële intelligentie (AI) op wetenschap en samenleving te onderzoeken. Een stuurgroep van deze Denkerscyclus selecteerde de internationale expert prof. dr. Helga Nowotny als "Thinker-in-residence", vanwege haar uitgebreide expertise over dit onderwerp in relatie tot de academische wereld en het wetenschapsbeleid. In overleg met de stuurgroep ging zij vervolgens in debat met relevante experts, stakeholders en praktijkmensen in Vlaanderen. Op basis van deze interacties heeft ze haar onafhankelijke beoordeling ontwikkeld. Als Thinker-in-residence heeft zij haar bevindingen en aanbevelingen op een afsluitend, openbaar symposium naar buiten gebracht.

Wij willen hier graag iedereen bedanken die heeft bijgedragen aan het succes van deze Cyclus: de stuurgroep, de geraadpleegde experts en stakeholders, het publiek en de KVAB-medewerkers. In het bijzonder gaat onze erkentelijkheid uit naar onze Thinker-in-residence, Helga Nowotny, voor een uitstekend rapport en haar aanbevelingen. Ze heeft haar langetermijnvisie op de impact van AI op een prikkelende manier geformuleerd op basis van haar interacties met Vlaamse experts en stakeholders. Ze heeft aldus een uitstekende basis gelegd voor de Vlaamse wetenschappelijke gemeenschap ter voortzetting en verbreding van de discussie over de integratie van AI in wetenschap en samenleving.

Deze publicatie is een Standpunt gebaseerd op de bevindingen en aanbevelingen van de Denkerscyclus. Dit Standpunt is op 15 december 2023 digitaal goedgekeurd voor publicatie door de Klasse van de Menswetenschappen (KMW) van de KVAB.

Ine Van Hoyweghen
Coördinator van de Denkerscyclus
29 november 2023

1. Inleiding: Positionering, doel en benadering van de Denkers-cyclus

Ine Van Hoyweghen

De mogelijkheden van Artificiële Intelligentie (AI) zijn de afgelopen tien jaar snel gegroeid. Met behulp van algoritmische inzichten, toegang tot enorme gegevensbronnen en rekenkracht hebben AI-onderzoekers systemen gecreëerd die taal kunnen begrijpen, afbeeldingen en videobeelden kunnen herkennen en genereren, computerprogramma's kunnen schrijven en wetenschappelijk kunnen redeneren. Als de huidige trends zich doorzetten, zullen AI-systemen een transformatieve invloed hebben op de wetenschap en onze maatschappij. Krachtige AI-systemen zullen aanzienlijke voordelen en risico's met zich meebrengen. Dit roept vragen op over hoe AI ons dagelijks leven zal beïnvloeden, zowel in de privé- als in de arbeidscontext, en hoe deze technologie onze kijk op wie we zijn als mens zal beïnvloeden. Het is binnen deze huidige en voortdurende transformaties dat de Denkerscyclus werd ontwikkeld, met als hoofddoel een reflectie en debat over "**AI als aanjager van verandering**".

Positionering van de Cyclus

In de afgelopen jaren heeft AI veel aandacht gekregen in de industrie, het onderwijs, het onderzoek, de politiek, de overheid en de samenleving in het algemeen. De snelle vooruitgang op het gebied van artificiële intelligentie (AI) heeft geleid tot de ontwikkeling van generatieve AI (GenAI), waaronder grote taalmodellen (large language models, LLM's) die tekst, afbeeldingen en code kunnen produceren op basis van patronen in hun trainingsgegevens. ChatGPT bijvoorbeeld is zo'n groot taalmodel (LLM). Het is het laatste in een reeks van dergelijke modellen die zijn uitgebracht door OpenAI, grotendeels gefinancierd door Microsoft, terwijl andere techbedrijven in een race zijn verwikkeld om vergelijkbare werktuigen uit te brengen.

Hoewel systemen van Artificiële Intelligentie geen magische instrumenten zijn, maar het simpele resultaat van wiskundige optimalisatie, een enorme computerkracht en enorme datasets, zijn deze buitengewoon veelbelovend. Ze hebben al geleid tot aanzienlijke voordelen, met name in de vorm van een verbeterde efficiëntie, nauwkeurigheid en gemak in een breed scala aan onderzoek en diensten, waaronder de gezondheidszorg, genomica, fusieonderzoek, productontwerp en -simulatie, autonome systemen, voorspellend onderhoud, kwaliteitscontrole, inspectieprocessen en milieueffectbeoordeling. Ook wetenschappers verwachten dat ze binnenkort een centrale plaats zullen innemen in het onderzoek, zoals blijkt uit een recente paper in het tijdschrift *Nature* waarbij meer dan 1.600 onderzoekers van over de hele wereld betrokken waren (Van Noorden & Perkel, 2023). Wetenschappers gebruiken AI als hulp bij het samenvatten en schrijven

van onderzoeksartikelen, het brainstormen van ideeën en het schrijven van code, terwijl anderen het potentieel van generatieve AI hebben onderzocht bij het produceren van nieuwe eiwitstructuren, het verbeteren van weersvoorspellingen en het suggereren van medische diagnoses, naast vele andere ideeën. Volgens de OESO (2023) zou het versnellen van de productiviteit van onderzoek wel eens de meest economisch en sociaal waardevolle van alle toepassingen van artificiële intelligentie (AI) kunnen zijn.

Tegelijkertijd gaat de opkomst van AI gepaard met toenemende bezorgdheid over de potentiële risico's en schadelijke effecten ervan: voor individuen, voor kwetsbare groepen en voor de samenleving in het algemeen. Onlangs luidden insiders uit de techindustrie de alarmbel over de "existentiële risico's" van artificiële intelligentie (*Nature* Editorial, 2023). Er werden open brieven gepubliceerd met duizenden handtekeningen waarin werd gepleit voor een pauze in het trainen van AI-systemen. Ongecontroleerd kan de ontwikkeling van AI "ons uiteindelijk in aantal overtreffen, te slim af zijn, overbodig maken en ons vervangen", of zelfs leiden tot een "verlies van controle over onze beschaving", waarschuwde een van de brieven. Deze focus op de "existentiële risico's" van AI leidt echter de aandacht af van de risico's van AI die zich momenteel al voordoen (*Nature* Editorial, 2023). AI versterkt en verergert veel problemen die zich vandaag al in de wereld voordoen, zoals vooroordelen, ongewenste profilering/ discriminatie, desinformatie, misbruik van gegevens, gesloten toegang en toenemende sociale ongelijkheid.

Deze AI-risico's zijn inmiddels goed gedocumenteerd door sociale wetenschappers (zie bijvoorbeeld Crawford, 2021; Benjamin, 2019). Veel van de modellen op basis van machine leren (ML) zijn 'zwarte dozen' die hun voorspellingen niet kunnen uitleggen op een manier die mensen begrijpen. Deze machine leer-modellen worden tegenwoordig in de maatschappij ingezet voor besluitvorming waarbij veel op het spel staat, wat leidt tot problemen met vooroordelen en discriminatie in de gezondheidszorg, sociaal beleid, verzekeringen en andere domeinen (Obermeyer et al., 2019; Rudin, 2019). Bevooroordeelde AI-systemen zouden ondoorzichtige algoritmes kunnen gebruiken om mensen uitkeringen, medische zorg of asiel te weigeren – toepassingen van de technologie die waarschijnlijk het meest van invloed zijn op mensen in gemarginaliseerde gemeenschappen (Kalluri, 2020). Een van de grootste zorgen rond generatieve AI is de mogelijkheid om desinformatie en "deep fakes" te stimuleren – video's met synthetische gezichten en stemmen die niet te onderscheiden zijn van die van echte mensen. Op termijn kan dergelijke praktijk het vertrouwen tussen mensen, politici, de media en de wetenschap aantasten, vooral als er geen regels zijn voor de productie van de onderliggende modellen en codes (Jones, 2023; van Dis et al., 2023; Van Noorden & Perkel, 2023). De onderliggende trainingssets en LLM's voor ChatGPT zijn niet openbaar en techbedrijven hebben de neiging om de interne werking van hun generatieve AI-systemen te verbergen (Ferrari et al., 2023; Bockting et al., 2023).

Dit alles vraagt om een uitgebalanceerde governance aanpak – een aanpak die de maatschappelijke waarden respecteert en door de bevolking wordt gesteund – voordat de technologie de wetenschap en het publieke vertrouwen ondermijnt. Deze zorg wordt ook uitgesproken op Europees beleidsniveau. In haar *State of the Union* in september 2023 riep Ursula von der Leyen, voorzitter van de Europese Commissie, op tot een wereldwijde aanpak om inzicht te krijgen in de gevolgen van AI, gemodelleerd naar het Intergovernmental Panel on Climate Change (IPCC), met de opdracht om "wereldwijde minimumnormen" vast te stellen voor een veilig en ethisch gebruik van AI (EC, 2023). Dit nieuwe orgaan over de voordelen en risico's van AI voor de mensheid zal bestaan uit wetenschappers, technologiebedrijven en onafhankelijke deskundigen. Deze oproep voor "verantwoorde AI" is in lijn met de baanbrekende wetgeving die de Commissie in april 2021 heeft voorgesteld, de Artificial Intelligence Act (EC, 2021), waarover momenteel wordt onderhandeld in het Europees Parlement en de lidstaten. De wet, die marktregels oplegt aan AI-systemen op basis van hun potentiële risico's voor de samenleving, wordt beschouwd als "nu al een blauwdruk voor de hele wereld" (EC, 2023). Terwijl de EU bezig is met het afronden van haar eerste wet over artificiële intelligentie, moet de wetenschappelijke gemeenschap nog met een eensluidend antwoord komen over hoe generatieve AI kan worden ingezet in het hoger onderwijs en onderzoek. Er zijn plannen om een speciale nieuwe eenheid op te richten binnen het Onderzoeksdirectoraat van de Europese Commissie om richtlijnen op te stellen en om een debat op gang te brengen als onderdeel van de beleidsagenda voor de Europese Onderzoeksruimte (ERA). In juli 2023 publiceerden wetenschapsadviseurs van de Europese Commissie een verkennende nota over de betrokken kwesties, waarin werd gewezen op een gebrek aan "specifiek en systemisch beleid dat de invoering van AI in de wetenschap vergemakkelijkt" (SAM, 2023).

De impact van AI is een veelzijdig thema met vele invalshoeken en gebieden, evenals meerdere stakeholders en beleidsniveaus. Geconfronteerd met deze laatste technologische verandering wendt men zich vaak instinctief tot technologen voor oplossingen. Maar de gevolgen van AI kunnen niet met louter technologische middelen worden geborgd; oplossingen zonder breder maatschappelijk inzicht zullen de gevaren van AI alleen maar vergroten (Lazar & Nelson, 2023). Dit vereist een mensgerichte aanpak bij het ontwerp, het gebruik en de verdere ontwikkeling van AI, wat inhoudt dat AI moet worden afgestemd op menselijke waarden en behoeften. Dit alles wordt met de dag dringender. Nu generatieve AI-systemen zich razendsnel ontwikkelen dient de wetenschappelijke gemeenschap een centrale rol te spelen bij het vormgeven van de toekomst van een mensgerichte benadering van AI.

Het doel van de Cyclus en de Denker

Het doel van de Denkerscyclus was om te reflecteren over de impact van deze nieuwste ontwikkelingen in AI, over de vraag wat de implicaties zijn voor ons

mensbeeld: autonomie, menselijk handelingsvermogen ("agency") en de noodzaak van een "digitaal humanisme". De titel van de Denkerscyclus, "AI as an Agent of Change" ("AI als aanjager van verandering"), is geïnspireerd op het tweedelige werk van historica Elizabeth Eisenstein, *The Printing Press as an Agent of Change* ("De drukpers als aanjager van verandering", 1980). In haar boek stelt Eisenstein dat de transformatie van een cultuur van manuscripten naar een cultuur van drukwerk een fundamentele invloed had op de renaissance, de protestantse reformatie en de opkomst van de wetenschappelijke revolutie. Het AI-tijdperk dat nu aanbreekt lijkt misschien wel op deze transformatie, omdat het niet alleen talloze voordelen zal opleveren, maar ook onbedoelde effecten die niet werden onderkend op het moment dat de technologie tot ontplooiing kwam.

Om door deze vragen van de Cyclus te navigeren, hadden we de eer om Helga Nowotny als onze Thinker-in-residence uit te nodigen. Helga Nowotny is emeritus hoogleraar wetenschaps- en technologiestudies aan de EHT in Zürich. Ze is stichtend lid en voormalig voorzitter van de Europese Onderzoeksraad (ERC). Ze is buitenlands lid van de Klasse van de Menswetenschappen (KMW) van de KVAB (zie bijlage 1 voor cv). Ze volgt de ontwikkelingen op het gebied van AI al meer dan een halve eeuw op de voet. In haar boek *In AI We Trust* (2021) gaat Nowotny in op de nieuwste ontwikkelingen door een breed filosofisch, sociologisch en historisch perspectief te bieden op de impact van AI. Ze wijst daarbij op een inherente paradox van ons vertrouwen in AI: "We willen AI gebruiken om onze toekomst beter te beheersen, maar door de voorspellende algoritmes vermindert AI onze vrijheid om die toekomst vorm te geven. AI moet daarom worden geflankeerd door onze menselijke capaciteit als 'aanjager van verandering' om een gedeelde, open toekomst te behouden" (Nowotny, 2021).

De interesses van de Denkerscyclus komen overeen met die van het Digital Humanism Initiative (Vienna Manifesto on Digital Humanism, 2019). Het doel van deze internationale samenwerking is een gemeenschap op te bouwen van wetenschappers, beleidsmakers en industriële spelers die ervoor willen zorgen dat deze technologische ontwikkeling gericht blijft op menselijke waarden. Digitaal humanisme observeert en beschrijft de veranderingen in digitale technologie en wil de ontwikkeling van deze technologieën en het beleid vormgeven en beïnvloeden in de richting van de waarden van mensenrechten, democratie, participatie, inclusie en diversiteit. De afgelopen jaren zijn wereldwijd soortgelijke initiatieven opgezet, zoals bijvoorbeeld het Institute for Human-Centered AI (HAI) aan Stanford University in 2019. Publicaties en themabijeenkomsten onderstrepen de urgentie van de betrokken kwesties, waaronder vrijheid, algoritmische transparantie, cognitieve autonomie en een "hybride geest" in de mens-machine-symbiose.

Het onderwerp AI wordt ook veel besproken in Vlaanderen en verschijnt in bijna elke uitgave van de tweewekelijkse mededelingen van de Vlaamse Adviesraad voor Innoveren en Ondernemen (VARIO). AI is ook het onderwerp van Vlaams en nationaal

beleid (Vlaams Beleidsplan AI (VAIA, 2019)) en het nationaal convergentieplan voor de ontwikkeling van artificiële intelligentie (Federale Overheidsdienst Beleid en Ondersteuning (BOSA, 2022)). Verschillende Denkerscycli van de KVAB hebben zich gericht op het thema van AI en digitalisering en zijn besproken in de reeks KVAB Standpunten: "Artificiële Intelligentie: naar een vierde industriële revolutie?" (Steels et al., 2017), "Maatschappelijke waarden bij digitale innovatie: wie, wat en hoe?" (Rabaey et al., 2019) en een recent gezamenlijk Standpunt van de KVAB en de ARB "Een oproep tot een versnelde digitale transformatie voor België" (Vandewalle et al., 2022). Andere academies en internationaal overkoepelende organisaties voeren debatten en organiseren activiteiten die ook belangrijk zijn voor deze Denkerscyclus. ALLEA bijvoorbeeld, waarvan de gedragscode voor onderzoeksintegriteit een van de leidende documenten voor Horizon Europe is, heeft dit kader eerder dit jaar bijgewerkt om de veranderingen als gevolg van AI te belichten (ALLEA, 2023).
In samenwerking met de Thinker-in-residence was het doel van deze Denkerscyclus om op basis van een brede reflectie voorstellen te onderzoeken voor aanvullend beleid dat op lokaal, nationaal en internationaal niveau wordt ontwikkeld om de productieve integratie van AI in de wetenschap en de maatschappij te stimuleren. De Cyclus levert aldus bijdragen en advies met betrekking tot de doelstellingen voorgesteld door de Vlaamse overheid voor 1. strategisch basisonderzoek, 2. opleidingsbehoeften, 3. ethische uitdagingen en 4. publieksgerichtheid.

Aanpak

De Cyclus werd voorgedragen door de Klasse van de Menswetenschappen (KMW), op initiatief van de leden Marc de Mey en Ine Van Hoyweghen. Het voorstel werd op 19 november 2022 door de Klasse van de Menswetenschappen aanvaard, waarna een startnota en een oproep tot deelname aan de stuurgroep werd verspreid onder alle KVAB-leden. In december 2022 werd deze stuurgroep samengesteld, bestaande uit leden uit de verschillende Klassen van de KVAB, de Jonge Academie en andere relevante experten (zie bijlage 2). De rol van de stuurgroep bestond erin de activiteiten van de Denker goed te onderbouwen en de nodige input te leveren. De Thinker-in-residence kreeg alle vrijheid en bleef volledig onafhankelijk bij het schrijven van het rapport en de aanbevelingen. Door samen te werken met de stuurgroep en tal van Vlaamse experts en stakeholders slaagde ze erin een belangrijke bijdrage te leveren aan het thema van de impact van AI door een langetermijnvisie te formuleren en op die manier bij te dragen tot de Vlaamse beleidsvorming. Ze ontwikkelde haar standpunten en aanbevelingen na verschillende rondes van intensieve ontmoetingen en overleg met experts en stakeholders in heel Vlaanderen.

De experts, stakeholders en praktijkmensen die hebben bijgedragen aan deze discussies zijn doorgaans afkomstig van alle relevante instellingen, zoals
- KU Leuven

- Universiteit Gent
- VUB
- Universiteit Antwerpen
- Universiteit Hasselt
- Onderzoekscentra (vertegenwoordigers van imec, Flanders AI, VIB ...)
- Koepelorganisaties (Jonge Academie)
- Financiers en beleidsmakers (vertegenwoordigers van VLAIO, Vlaamse overheid - dep. EWI)
- AI-praktijkmensen

In een eerste fase wisselde de Thinker-in-residence haar visie en ideeën over de Denkerscyclus uit met de initiatiefnemers van de Cyclus en met de stuurgroep. De stuurgroep had een eerste informele bijeenkomst op 7 maart 2023, gevolgd door een startbijeenkomst met de Thinker-in-residence in de Academie op 8 maart 2023. Deze startbijeenkomst werd georganiseerd voor leden van de KVAB, de Jonge Academie en andere experten uit het veld om Nowotny als Thinker-in-residence en het onderwerp van de Denkerscyclus te introduceren. Tijdens deze drukbezochte bijeenkomst gaf ze een presentatie gebaseerd op haar meest recente boek, getiteld *In AI We Trust: Power, Illusion and Control of Predictive Algorithms* (2021, Polity Press), gevolgd door een discussie en uitgebreide debatten.

Programma KVAB-bijeenkomst "AI as an Agent of Change", 8 maart 2023, KVAB, Brussel

10:00-10:15:	Welkomstwoord en introductie door Ine Van Hoyweghen, voorzitter Denkerscyclus
10:15-11:00:	Lezing door Helga Nowotny: "Liberal Democracies at Risk: Algorithmic Communication and the Delegitimation of Truth"
11:10-12:00:	Q&A met moderator Katleen Gabriels (Universiteit Maastricht)

Op basis van deze discussies werd een overzicht van visies en ideeën van de Denkerscyclus gemaakt en besproken tijdens een fysieke bijeenkomst van de Stuurgroep. Er werd besloten om te focussen op de impact van generatieve AI op de wetenschap en de maatschappij. Op basis hiervan heeft de Thinker-in-residence een eerste concept van het rapport opgesteld.

In een tweede fase werd de Thinker-in-residence uitgenodigd om deel te nemen aan debatten met Vlaamse experts, partners en stakeholders. Ze bereidde relevante vragen voor en besprak deze met de stuurgroep in verschillende online bijeenkomsten in het voorjaar van 2023, waarin de verdere planning van de bijeenkomsten van experts en stakeholders werd voorbereid voor haar bezoek in september 2023.

Om de Denker in contact te brengen met Vlaamse relevante experts in het veld, werd op woensdag 13 september 2023 een Expertbijeenkomst georganiseerd in

de Academie. De experts werden uitgenodigd om het concept van haar rapport te lezen en gevraagd om ideeën en opmerkingen te presenteren vanuit hun specifieke domeinen. Ze gebruikte de input van deze bijeenkomst als feedback voor haar rapport en om concrete ideeën voor beleidsaanbevelingen te ontwikkelen. De experts werden ook uitgenodigd om een reflectie te schrijven op het rapport van de Denker voor dit Standpunt (zie hoofdstuk 3).

Programma van de Expertbijeenkomst "AI as an Agent of Change", 13 september 2023, Brussel

10:00-10:15	Welkomstwoord en introductie door Ine Van Hoyweghen, voorzitter Denkerscyclus
10:15-11:00	Presentatie van het concept van het rapport door Helga Nowotny, Denker
11:00-11:30	Presentatie door Tinne Tuytelaers, KU Leuven
13:00-13:30	Presentatie door Ann Dooms, VUB
13:30-14:00	Presentatie door Johan Wagemans, KU Leuven
14:30-15:00	Presentatie door Rosamunde Van Brakel, VUB/UHasselt
15:00-15:30	Afsluitend debat en brainstorm voor beleidsaanbevelingen

Workshops met Vlaamse stakeholders en praktijkmensen

De Denker ging ook in debat met verschillende Vlaamse stakeholders, praktijkmensen en partners. Hoewel de gesprekken met de stakeholders flexibel en open waren, gingen ze gepaard met een lijst van vragen die vóór elke bijeenkomst werd verspreid. Daarnaast werden de deelnemers uitgenodigd om voorafgaand aan de workshops schriftelijke input voor te bereiden. Hierdoor kreeg de Denker geleidelijk meer inzicht in de lokale situatie en kon ze hierop reflecteren vanuit haar internationale perspectief. De input van deze stakeholder workshops werd door de Denker gebruikt om concrete ideeën voor beleidsaanbevelingen te ontwikkelen. Een samenvatting van deze workshops is opgenomen in dit Standpunt (zie hoofdstuk 5).

Stakeholderworkshop 1 over "ChatGPT en onderwijs/onderzoek", 12 september 2023, Brussel

Thema: AI als aanjager van verandering: Hoe worden AI/ChatGPT gebruikt, ervaren en ondersteund in het formele onderwijs en in de informatie aan en opleiding van het bredere publiek in Vlaanderen?

Stakeholderworkshop 2 over "AI-onderzoek en -toepassingen", 15 september 2023, Brussel

Thema: AI als aanjager van verandering: Hoe worden AI/ChatGPT gebruikt, ervaren en ondersteund in het onderzoek en in de opleiding van studenten in Vlaanderen?

Deze intensieve bijeenkomsten, debatten en workshops en hun schriftelijke input leverden samen een realistisch beeld op van de activiteiten die in Vlaanderen plaatsvinden, in combinatie met de benaderingen, uitdagingen, problemen en vooruitzichten. Vanuit het perspectief van haar internationale ervaring stelde de Denker vervolgens een basis voor vergelijking samen voor de Vlaamse context om beleidsaanbevelingen te ontwikkelen. Deze beleidsaanbevelingen werden goedgekeurd door de Stuurgroep op 21 september 2023.

Tot slot werden het rapport en de aanbevelingen van de Thinker-in-residence gepresenteerd op een drukbezocht publiek symposium "AI as an Agent of Change" in het Paleis der Academiën op maandag 20 november 2023.

Tijdens het symposium werd ook een tentoonstelling georganiseerd over de impact van AI op kunst. Recente ontwikkelingen in AI hebben een sterke invloed op de kunsten en de stuurgroep (op initiatief van KVAB-lid Luc Steels) vond het interessant om dit facet te belichten met een tentoonstelling over intrigerende *crossovers* tussen AI en beeldende kunst, poëzie en muziek.

Programma van het publieke symposium "AI as an Agent of Change", 20 november 2023, Brussel

9:00 REGISTRATIE & BEZOEK TENTOONSTELLING
9:45 Opening
Gastheer/moderator: Jan Hautekiet
Welkomst- en openingswoord – Christoffel Waelkens, voorzitter KVAB
Een introductie – Ine Van Hoyweghen, voorzitter van de Denkerscyclus, KU Leuven
Presentatie van het Denkersrapport – Helga Nowotny, Thinker-in-residence
11:10 Impact van AI op kunst
Paneldiscussie met kunstenaars gevolgd door een Q&A met het publiek
Voorzitter: Luc Steels (in het Nederlands)
- Danny Devos, beeldend kunstenaar
- Maarten Inghels, dichter
- Andrew Claes, muzikant
- Kris Stroobants, dirigent Frascati Symphonic

12:10 Inzichten van onderzoekers over de impact van AI
Presentaties gevolgd door een Q&A met het publiek
- Ann Dooms, VUB
- Johan Wagemans, KU Leuven
- Walter Daelemans, UAntwerpen

13:30 Hoe doet Vlaanderen het op het gebied van AI? – Bart De Moor, KU Leuven
De ervaring en vooruitzichten voor toekomstig beleid – Lucilla Sioli, directeur Artificiële Intelligentie en Digitale Industrie van DG CONNECT, Europese Commissie

Tentoonstelling samengesteld door Luc Steels: Impact van AI op kunst

AI & «La révolte des machines ou la pensée déchaînée»

Beeldend kunstenaar **Danny Devos** verkent het gebruik van AI in de kunst in een tentoonstelling van werken gegenereerd door AI-modellen op basis van machine leren. Dit heeft geresulteerd in sculpturale objecten geproduceerd door 3D printen en CNC frezen, met elektromotoren en microcontrollers, gebaseerd op de illustraties van **Frans Masereel** voor "La Révolte des Machines".
Danny Devos (°1959), die in Antwerpen woont, heeft zijn artistieke inspanningen over de hele wereld gepresenteerd. Sinds 1979 heeft hij 160 voorstellingen gegeven in meer dan 40 steden in 12 verschillende landen. Al bijna veertig jaar treedt hij op als performance-, geluids- en "forensisch" kunstenaar, waarbij hij er doelbewust naar streeft om een kritische stem te blijven binnen en tegenover de kunstscène.

Secrets

In 2021 werkten kunstenaar **Luc Tuymans** en AI-wetenschapper **Luc Steels** samen om het proces van kunstcreatie en kunstperceptie en -interpretatie te begrijpen. Het resultaat was te zien in een tentoonstelling in Bozar in april 2021. Hier tonen we video's van Tuymans en Steels die deze resultaten uitleggen en hun bredere implicaties bespreken. Dit project werd geïnitieerd door het EU Starts programma, met de steun van Gluon (Brussel) en Bozar.

Poem Booth

Maarten Inghels (voormalig stadsdichter van Antwerpen) presenteert zijn "poem booth" tijdens het symposium: het gaat om een experiment met generatieve AI, waarbij hij relevante kwesties aan de orde stelt voor het publiek. Taal: Nederlands.
Maarten Inghels debuteerde in 2008 met "Tumult" in de Sandwich-reeks, onder redactie van auteur Gerrit Komrij, en heeft zich sindsdien ontwikkeld tot een origineel kunstenaar, dichter en schrijver. Zijn roman *Het mirakel van België*, over zijn ervaringen met de grootste oplichter ter wereld, verscheen in 2021. Zijn boek *Contact* verbond poëzie, beeldend werk en actie. Van 2016 tot 2018 was hij stadsdichter van Antwerpen.
Een voorproefje:

> *Kus elkaar, verliefden, onder de kruin,*
> *Gegiechel galmt, vanuit de massa tuin.*
> *Plots, een snorvogel schiet voorbij!*
> *Herhaalt dit lied, dit zoete vrij.*

AI Musicking

Andrew Claes en Frascati Symphonic geven een live uitvoering van een nieuwe compositie gegenereerd met AI en gespeeld door klassieke musici.
Andrew Claes is een professionele saxofonist en componist verbonden aan het Koninklijk Conservatorium Antwerpen. Een van zijn belangrijkste projecten, AI Musicking, is gericht op het verkennen van innovatieve benaderingen van muzikale co-creatie door middel van machine leren. **Frascati Symphonic** is een groep muzikanten uit Leuven. Ze staan bekend om hun uitvoeringen van het klassieke repertoire, variërend van symfonische werken en opera's tot kamermuziek. Het orkest staat onder leiding van dirigent Kris Stroobants. De musici die deelnemen aan een kort optreden zijn Hrayr Karapetyan (viool), Delejan Breynaert (viool) en Shuya Tanaka (cello).

Deze verschillende activiteiten in 2023, waaronder het symposium, maakten het mogelijk voor het publiek en andere sprekers om verschillende interessante interacties te hebben met de Thinker-in-residence. Op basis van deze wederzijdse communicatie heeft ze haar eindrapport en de aanbevelingen opgesteld.

Dit Standpunt bestaat verder uit het Denkersrapport, de reflecties van experts, de reacties van beleidsmakers, de verslagen van de stakeholder workshops en een afsluitend hoofdstuk met conclusies en aanbevelingen. Het doel was om een brede reeks aanbevelingen te genereren die relevant zijn over de disciplines heen om bij te dragen aan toekomstig Vlaams beleid op dit vlak. Op basis van de specifieke achtergronden, perspectieven en expertises presenteert dit Standpunt niet alleen een consistente analyse, maar ook waardevolle benaderingen en voorstellen voor verdere stappen. Gezien de kwaliteit van de discussies van de Denker met de deelnemers en haar constructieve bevindingen, bieden deze inspanningen een solide basis voor de productieve integratie van AI in de wetenschap en de maatschappij in Vlaanderen. Het kan zelfs inzichten, reflecties en aanbevelingen opleveren die verder reiken dan Vlaanderen, in het bijzonder voor het SAM-SAPEA-programma voor een versnelde invoering van AI in de wetenschap (SAM, 2023) en andere acties in Europa en wereldwijd.

Referenties

ALLEA (2023) The European Code of Conduct for Research Integrity – Revised Edition 2023. Berlin. DOI 10.26356/ECOC

Benjamin, R. (2019) *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Bockting, C.L., van Dis, E.A.M., van Rooij, R., Zuidema, W. Bollen, J. (2023) Living guidelines for generative AI – why scientists must oversee its use. *Nature*. 622, 7984, 693-696. doi: 10.1038/d41586-023-03266-1

Crawford, K. (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, Conn.: Yale University Press.

Eisenstein, E.L. (1980) *The Printing Press as an Agent of Change*. Cambridge: Cambridge University Press.

European Commission (EC). (2021) Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

European Commission (EC) (2023) 2023 State of the Union Address by President von der Leyen Strasbourg, 13 September 2023, State of the Union Address by President von der Leyen (europa.eu)

Federal Public Service, Policy and Support (BOSA) (2022) National Convergence Plan for the development of Artificial Intelligence (2022) Plan AI (NL)-compressed.pdf (belgium.be)

Ferrari, F., van Dijck, J. & van den Bosch, A. (2023) Foundation models and the privatization of public knowledge. Nat Mach Intell 5, 818–820 https://doi.org/10.1038/s42256-023-00695-5

Human-Centred Artificial Intelligence (HAI) Stanford University, Home | Stanford HAI

Jones, N. (2023) How to stop AI deepfakes from sinking society - and science. *Nature*. 621, 7980, 676-679. doi: 10.1038/d41586-023-02990-y

Kalluri, P. (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*. 583, 7815, 169. doi: 10.1038/d41586-020-02003-2

Lazar, S. and Nelson, A. (2023) AI safety on whose terms? Science. 381, 6654, 138. doi: 10.1126/science.adi8982.

Nature Editorial, Stop talking about tomorrow's AI doomsday when AI poses risks today, *Nature* 618, 885-886 (2023) doi: https://doi.org/10.1038/d41586-023-02094-7

Nowotny, H. (2021) *In AI We Trust. Power, Illusion and Control of Predictive Algorithms*. Cambridge, UK: Polity Press.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 366, 6464, 447-453. doi: 10.1126/science.aax2342.

OECD (2023) Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, OECD Publishing, Paris, https://doi.org/10.1787/a8d820bd-en

Rabaey, J., van Est, R., Verbeek, P.P., Vandewalle, J. (2020) Societal values in digital innovation: who, what and how?- KVAB Thinkers' Programme 2019, KVAB Position paper 66 b, 2020.

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 206–215 https://doi.org/10.1038/s42256-019-0048-x

Scientific Advice Mechanism to the European Commission (SAM) (2023) Artificial intelligence in science. Scoping paper, 4 July 2023. https://scientificadvice.eu/advice/artificial-intelligence-in-science/

Steels, L. e.a. (2017) Artificiële intelligentie. Naar een vierde industriële revolutie?, KVAB Standpunt 53.

VAIA (2019) Flemish Policy Plan for Artificial Intelligence. Flemish Policy Plan AI - VAIA - Flanders AI Academy

Vandewalle, J., Acheroy, M. e.a. (2022) A call for an accelerated digital transformation for Belgium, ARB/KVAB Position paper 77, 2022.

van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L. (2023) ChatGPT: five priorities for research. *Nature*. 614, 7947, 224-226. doi: 10.1038/d41586-023-00288-7.

Van Noorden, R., Perkel, J.M. (2023) AI and science: what 1,600 researchers think. *Nature*. 621, 7980, 672-675. doi: 10.1038/d41586-023-02980-0

Vienna Manifesto on Digital Humanism (2019) https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto/

2. Denkersrapport

AI als aanjager van verandering

Helga Nowotny, Thinker-in-residence

Elizabeth Eisensteins invloedrijke klassieker *The Printing Press as an Agent of Change*, oorspronkelijk gepubliceerd in twee delen in 1980, was de aanleiding voor het thema van de Denkerscyclus van de KVAB van 2023. Het is een uitnodiging om AI in een breder historisch kader te plaatsen – inclusief de hype en de geweldige efficiëntie die zelfs experten verbaast, maar ook de dreigende zorgen waartoe ze aanleiding geeft. Technologieën komen niet uit de lucht vallen en het is zinvol om na te denken over de langdurige, vaak niet-lineaire en onvoorspelbare gevolgen van menselijke uitvindingen voor de samenlevingen waarin ze zijn ontstaan. In de relatie tussen technologie en maatschappij is nooit sprake van eenrichtingsverkeer. Technologieën geven vorm aan samenlevingen en hun economieën, maar worden er ook door gevormd. Samenlevingen passen zich aan, veelal op onvoorziene manieren, aan de technologieën die er een invloed op hebben. Ze eigenen ze zich ook toe en vinden nieuwe en ongeplande toepassingen uit, die bestaande machtsstructuren kunnen versterken of ondermijnen, latente behoeften kunnen vervullen of, meer in het algemeen, de weg kunnen vrijmaken voor het verkennen en benutten van nieuwe mogelijkheden.

Een historische terugblik: overeenkomsten en verschillen

"AI als aanjager van verandering" is een intrigerende metafoor. Het situeert technologische verandering als verweven met sociale verandering en plaatst deze in een bredere historische context wat minstens drie vragen oproept die de leidraad zullen vormen voor dit rapport. De eerste voor de hand liggende vraag gaat over de overeenkomsten en verschillen, de continuïteiten en discontinuïteiten die kunnen worden gevonden bij het vergelijken van de maatschappelijke impact van technologische vooruitgang in AI/ML met die van voorgaande technologieën. De drukpers is ongetwijfeld een goed begin. De impact ervan was enorm, voor Europa en door de koloniale expansie ver daarbuiten. Ze veroorzaakte veranderingen die varieerden van de proliferatie van drukkerijen in Europese steden vanaf 1500 tot de ontwrichtende effecten ervan op bestaande sociale en politieke structuren. Ideeën werden verspreid via nieuw gecreëerde sociale netwerken, wat leidde tot mentaliteitsveranderingen die op hun beurt in grote mate bijdroegen aan de opkomst van de moderne wetenschap, de reformatie en de Europese verlichting.

Er ontstonden nieuwe markten voor kapitalistische ondernemingen, waardoor de financiering hiervan verder veranderde. De opvallende toename van het aantal beschikbare boeken zorgde voor een nieuw publiek en nieuwe lezers. De boekproductie ging gepaard met de opkomst van een divers lezerspubliek,

waardoor een langdurige, zij het ongelijke, verspreiding van geletterdheid op gang kwam. In de woorden van Elizabeth Eisenstein: "Het feit dat identieke afbeeldingen, kaarten en diagrammen tegelijkertijd bekeken konden worden door verspreide lezers vormde een soort van communicatierevolutie op zich" (Eisenstein, 1980, p. 53). Het resultaat was een ware kennisexplosie in de 16de eeuw. Hoewel dit vaak wordt geassocieerd met de ontdekking van de Nieuwe Wereld, heeft de toegang tot boeken en de ideeën die door deze boeken werden overgebracht minstens zoveel bijgedragen. Galileo Galilei beweerde dat het boek van de natuur is geschreven in de taal van de wiskunde. Het feit dat de natuur – en al het andere dat de moderne wetenschap ons laat ontdekken – toegankelijk is voor de mensheid, is voor een groot deel te danken aan de drukpers en de maatschappelijke veranderingen die daardoor in gang werden gezet.

Het is daarom verleidelijk om parallellen te trekken tussen de kennisexplosie van de 16de eeuw en de "informatie-explosie" die ons sinds enige tijd in haar greep houdt. De recente publicatie van generatieve AI op basis van LLM's (large language models) heeft verder bijgedragen aan de overweldigende hoeveelheid mogelijkheden van AI/ML. De convergentie van rekenkracht, de prestaties van neurale netwerken en de toegang tot een enorme en groeiende hoeveelheid gegevens heeft de aanzet gegeven tot de versnelling van de meeste recente technologische ontwikkelingen. Het is een andere "Nieuwe Wereld" die we op het punt staan te ontdekken, het nog grotendeels onbekende gebied van "digi-land" en wat het voor ons in petto heeft. Veel mensen vrezen dat ze niet meer in staat zijn om te gaan met de snelheid en de stortvloed aan informatie en wat er van hen wordt gevraagd.

Sociale media – toen en nu

Zoals zo vaak het geval is, zijn ogenschijnlijk duidelijke overeenkomsten met historische precedenten bij nadere beschouwing al gauw een stuk minder eenduidig. Ongetwijfeld opende de drukpers nieuwe horizonten voor het lezerspubliek, dat gretig alle nieuwe kennis en informatie verslond. Dit stimuleerde de verspreiding van ideeën, wat leidde tot verhitte discussies en controverses die de verspreiding ervan verder bevorderden. Vandaag is het verlangen naar het nieuwe steeds groter. Sociale media zijn geprogrammeerd om dit verlangen te versterken door zich te richten op individuen of groepen, waardoor mensen zich verder terugtrekken in zogenaamde bubbels van gelijkgestemden. Onze samenlevingen lijken steeds meer gefragmenteerd en velen geven de sociale media de schuld hiervan. Hun aanbevelingsalgoritmes en rangorde versterken reeds bestaande tendensen, maar ze doen dit via specifieke manieren van interactie met gebruikers. Zo speelden zowel de algoritmes van Facebook als de keuzes die gebruikers maakten, of waartoe ze werden aangezet, een niet te verwaarlozen rol in de Amerikaanse presidentsverkiezingen van 2022. De polariserende invloed van FB-algoritmes is ingebouwd in de inhoud die gebruikers te zien krijgen, maar dat geldt ook

voor wat ze kiezen om te zien. Dat de feed van gebruikers bij elke stap van het aanbevelingsalgoritme meer polarisering in de hand werkt, waardoor gebruikers meer betrokken raken bij de polariserende inhoud, heeft een sterk ondermijnend effect (Uzogara, 2023).

Dit is slechts een van de vele gedetailleerde, maar belangrijke mechanismen waarmee machines gedrag beïnvloeden die ons eraan herinneren dat machines gebouwd zijn om bepaalde functies te vervullen. Er zijn menselijke bedoelingen in vastgelegd. Het is een feit dat propaganda ook welig tierde in de tijd van de drukpers, toen pamfletten en lasteraanvallen relatief goedkoop gedrukt en snel verspreid konden worden. Maar het verschil met het bereik, de snelheid en de onomkeerbaarheid van de huidige verspreiding via sociale media is even duidelijk als zorgwekkend. Nepnieuws, zo wordt ons herhaaldelijk verteld, is niets nieuws, maar op geen enkel moment in de geschiedenis konden "deep fakes" worden geproduceerd die het vrijwel onmogelijk maken om te onderscheiden of het gezicht dat we zien of de stem die we horen echt is of niet. De grenzen tussen "waar" en "onwaar" worden steeds vager, niet alleen als het gaat om uitspraken over de echte wereld, maar ook over de vele digitale representaties ervan. Als de drukpers werd gezien als een bedreiging voor de religieuze en seculiere autoriteiten van die tijd, dan vormen de digitale technologieën van vandaag een enorme bedreiging voor de instellingen en principes waarop liberale democratieën zijn gebouwd. Zodra het legitimerende onderscheid tussen "waar" en "onwaar" teniet is gedaan, lijken we te zijn overgeleverd aan de willekeur van anomie of de onderwerping aan autoritair gezag.

Machtsverschuivingen – de staat versus bedrijven

Technologieën geven aanleiding tot verschuivingen in de machtsstructuren. De drukpers versterkte de centralisatie van de macht in de natiestaat. De boekdrukkunst hielp bij het codificeren en standaardiseren van de taal en droeg zo bij aan de opkomst van nationale identiteiten. Vandaag is er daarentegen sprake van een sterke concentratie van economische macht in de handen van een paar grote internationale bedrijven en hebben regeringen en staten moeite om deze in toom te houden. Ze weten niet hoe ze de rechten van burgers moeten beschermen en hoe ze collectieve schade moeten aanpakken zonder het potentieel van technologische innovatie te belemmeren. De uitdagingen waarmee overheden worden geconfronteerd variëren van de bescherming van de privacy die burgers eisen tot de vraag of er op tijd genoeg nieuwe banen zullen worden gecreëerd om de banen die zullen verdwijnen te vervangen. Het is ook niet bekend hoe een geherstructureerde arbeidsmarkt, - een van de belangrijkste pijlers van de natiestaat -, het belastingsysteem, zal beïnvloeden. Een ander urgent probleem dat moet worden aangepakt, heeft te maken met de betekenis van een snelle verspreiding van AI voor het beheer van openbare diensten, met name de gezondheidszorg en het onderwijs. Vooral in de gezondheidszorg leiden de

data-intensivering en de integratie van AI-ondersteunde datapraktijken tot een verschuiving van de controle naar meer standaardisatie en een grotere efficiëntie, maar ook naar de privé-sector die veel diensten overneemt die tot nu toe tot het publieke domein hebben behoord.

Eisenstein herinnert ons eraan dat het drukwerk de functie had om bestaande normen, waarden, overtuigingen en ideologieën te versterken. Vandaag maken we ons zorgen dat de schijnbaar oncontroleerbare verspreiding van nepnieuws en samenzweringstheorieën onze gemeenschappelijke normen, waarden en overtuigingen verder zal ondermijnen, waardoor er een gevaarlijke 'publieke leegte' ontstaat die door eender wat gevuld kan worden. Hannah Arendt waarschuwde enige tijd geleden al voor de opkomst van het totalitarisme: als de wereld eenmaal onbegrijpelijk is geworden, hebben mensen "het punt bereikt waarop ze tegelijkertijd alles en niets geloven, denken dat alles mogelijk is en dat niets waar is" (Arendt, 1951). Een dergelijke situatie leent zich, zoals we tijdens de pandemie hebben gezien, voor een regelrechte aanval op het sociale gezag van op wetenschap gebaseerde expertise, die uiteindelijk leidt tot de afschaffing van het onderscheid tussen "waar" en "onwaar".

Het identificeren van historische overeenkomsten en verschillen is daarom nooit eenvoudig. We benaderen de geschiedenis door de lens van de meest dringende zorgen die ons in het heden bezighouden. De vragen die we stellen zijn geworteld in, en omkaderd door, wat ons het meest bezighoudt. De geschiedenis wordt steeds weer geherinterpreteerd, deels omdat er nieuwe bronnen en materialen blijven verschijnen, maar vooral omdat we nieuwe vragen stellen. Sommige komen voort uit praktische overwegingen en kunnen ons behoeden voor de illusie dat de nieuwste technologische golf altijd "revolutionair" is. Geschiedenis is het beste tegengif tegen hypes dat ooit is uitgevonden. Wat wij als ongekend beschouwen, blijkt toch precedenten te hebben, ook al zijn die slechts gedeeltelijk en zeer selectief. Toch proberen we te leren hoe samenlevingen in het verleden zijn omgegaan met de uitdagingen van nieuwe technologieën. Wat heeft gewerkt, voor wie, en wat waren de positieve en negatieve effecten, achteraf gezien?

AI – een systeemtechnologie met brede toepassingsmogelijkheden

Eén zo'n benadering wordt geboden door innovatiehistorici. Algemeen is men het er over eens dat AI een "General Purpose Technology" (GPT) is. Dit is een geheel van technologieën met een breed scala aan toepassingen in verschillende economische sectoren en industrieën. Hun alomtegenwoordigheid biedt innovatieve complementariteiten en hun doorsijpelende effecten hebben de neiging om zich te verspreiden naar lagere niveaus. De langetermijneffecten zijn daarom moeilijk te voorspellen, omdat het tijd kost voordat een systemische verandering is bereikt die alle sectoren en niveaus van de economie omvat. Dit kan ook verklaren waarom we veranderingen op korte termijn meestal overschatten en op lange

termijn onderschatten. Het meest prominente historische voorbeeld van een GPT zijn elektriciteit en elektrificatie, inclusief de rol van de kleinere elektrische motor in de industriële productie. De economisch historica Carlotta Perez heeft de korte- en langetermijneffecten geanalyseerd vanuit het perspectief van techno-economische paradigmaveranderingen. Ze laat zien dat elk van de voorgaande grote paradigmaverschuivingen heeft geleid tot een snelle concentratie van rijkdom in de handen van een paar ondernemers en van meedogenloze investeerders en roekeloze speculanten. Er ontstaan grote inkomensverschillen tussen winnaars en verliezers en er heerst een mentaliteit van "alles voor de winnaar". Uiteindelijk moesten regeringen ingrijpen om sociale onrust te voorkomen en/of een meer solidaire en progressieve politieke visie na te streven (Perez, 2018).

Zodra een technologie mainstream wordt, zoals in het geval van AI-toepassingen op veel gebieden en de snelle verspreiding van generatieve AI, slaat de verandering over en verandert het economische ecosysteem en zijn complexe dynamiek. Onderwijs, gezondheid, werk en bedrijfsleven zullen te maken krijgen met "revolutionaire veranderingen", in de oorspronkelijke betekenis van "omwentelingen". Dergelijke overwegingen liggen ten grondslag aan een conceptuele benadering waarbij AI wordt gezien als een "systeemtechnologie" die het bredere technologische en sociale ecosysteem omvat. Het kan dan vergeleken worden met de maatschappelijke effecten van eerdere systeemtechnologieën, zoals de stoommachine, elektriciteit, de verbrandingsmotor en de computer. Op een meer pragmatisch niveau kunnen dan aanbevelingen aan de overheid over hoe AI in te bedden in de samenleving worden afgeleid uit de geschiedenis van eerdere systeemtechnologieën (Prins et al., 2021).

De historische terugblik maakt het mogelijk om overeenkomsten en verschillen te identificeren, waaruit hopelijk een aantal lessen kunnen worden getrokken als leidraad voor de toekomst. Omdat "lessen" uit het verleden altijd samengaan met een groot voorbehoud, is een van de belangrijkste boodschappen voor vandaag waarschijnlijk het aanscherpen van de kritische blik op wat er deze keer anders is. Het is duidelijk dat dit niet alleen de technologie is die verbazingwekkende en significante vooruitgang brengt in vergelijking met wat voorheen mogelijk was. Zo'n benadering stelt ons eerder in staat om het grotere plaatje te zien waarin technologie nauw verweven is met de maatschappij die de technologie die ze heeft voortgebracht op veel verschillende manieren absorbeert, integreert, vormt, aanpast en zich toe-eigent. Dit gebeurt via uiterst selectieve mechanismen, afhankelijk van bestaande sociale structuren en praktijken die worden bemiddeld door complexe processen. Op basis van gedeelde praktijken hebben mensen het vermogen om nieuwe performantere relaties, structuren en netwerken te creëren. De performantie komt voort uit het gebruik van symbolen, het opnieuw uitvinden van sociale relaties en het zich verbeelden van collectieve toekomsten. We noemen het cultuur – en wij zijn actieve deelnemers aan een AI-cultuur in wording.

Waar zijn de burgers?

Een ander belangrijk aspect is het feit dat technologie nooit los kan worden gezien van de macht die ze verleent. Ze kan bestaande machtsstructuren versterken of deze verzwakken door nieuwkomers in staat te stellen macht te verwerven. Gevestigde belangen van de zittende machtshebbers spelen daarbij altijd een rol. Ondanks de retoriek van innovatie die een groot deel van het officiële politieke discours domineert, is het nieuwe niet altijd welkom en zeker niet bij degenen wiens gevestigde belangen worden bedreigd. In de begindagen van het internet was er een korte periode die doordrenkt was van een emancipatorische impuls. Veel technologische pioniers geloofden dat het internet een "democratiserende" invloed kon uitoefenen, waardoor iedereen kon deelnemen en delen in de voordelen. Helaas werden dergelijke idealistische impulsen al snel opgegeven, gulzig geabsorbeerd door wat de "Tech Bro"-cultuur van Silicon Valley is geworden, gevoed door het succes en het geloof in – of de illusie van – de eigen onmetelijke macht.

Onlangs, toen ChatGPT op de markt werd gebracht zonder iemand toestemming te vragen, laat staan rekening te houden met de stemmen en behoeften van burgers, werden we onderdeel van een groot experiment uitgevoerd door OpenAI en zijn concurrenten. De strijd om de macht van de grote internationale bedrijven te reguleren is nog maar net begonnen en de pogingen van ingewijden in de technologie om open source te introduceren staan nog in de kinderschoenen. De participatie van burgers wordt gereduceerd tot de rol van gebruikers op sterk vooraf gedefinieerde en gestructureerde manieren, gebaseerd op de handelingen van algoritmes die zijn ontworpen om het aantal "kliks" en de winst te maximaliseren. De disbalans in de financiering van AI-onderzoek en -ontwikkeling is schrijnend: slechts een tiende van de investeringen in de VS en de EU komt uit publieke bronnen, terwijl de overige 90% uit private bronnen afkomstig is. Dit bepaalt in grote mate ook de richting van toekomstig onderzoek. Het doel om van AI een publiek goed te maken is nog ver weg.

Het werk van Elizabeth Eisenstein is indrukwekkend omdat ze een bredere blik werpt op hoe de maatschappij zich de mogelijkheden die de drukpers bood actief en selectief toe-eigende. Ze laat zien hoe deze uitvinding werd gebruikt door kerk en staat, kapitalisten, handelaren en geleerden om hun belangen en overtuigingen te behartigen en te bevorderen. Technologie kan in verschillende culturen voor verschillende doeleinden worden gebruikt; de machtshebbers kunnen technologie inzetten voor onderdrukking, en dat is ook geprobeerd. De belangen van de elites, of ze nu materieel zijn of op het gebied van ideeën, doen er altijd toe. Vandaag worden we opnieuw volledig blootgesteld aan de verschillende krachten die aan het werk zijn. De concurrentie tussen de grote bedrijven om marktaandeel manifesteert zich in de verbijsterende verscheidenheid aan ChatGPT-modellen die nog steeds worden uitgebracht, samen met de inspanningen van kleine start-ups

die inzetten op open source in de hoop het groeiende oligopolie van de grote techbedrijven te ondermijnen. Het is duidelijk dat de consolidatiefase nog moet beginnen. Nog zorgwekkender zijn de geopolitieke spanningen tussen de VS en China. Ze komen onder andere tot uiting in de felle concurrentiestrijd om de onmisbare zeldzame materialen en de productie van chips, die weerklinkt in de roep van Europa om "technologische soevereiniteit". De strijd om de regelgeving, waarin de EU de wetgevende voorloper is met implementatie als moeilijkste taak, is nog maar net begonnen. Het blijft nog altijd lastig om zelfs maar te komen tot minimumnormen voor wereldwijde regulering.

De vergelijking met de veranderingen die in gang werden gezet door de drukpers verscherpt aldus de kritische blik op de huidige situatie. Ondanks enkele overeenkomsten zijn de verschillen groot. Tegelijkertijd, zoals ik zal laten zien, is er sprake van continuïteit in de co-evolutie tussen mens en technologie. Het betreft de culturele co-evolutie tussen mensen en de machines die zij ontwikkelen met een open einde, zoals er ook in de biologische co-evolutie altijd sprake is van een open einde.

Wie is een aanjager van verandering en wat is handelingsvermogen? De functie van communicatie

Het thema van de Denkerscyclus van 2023 stelt ook de vraag wie een aanjager is en wat handelingsvermogen is. De antwoorden liggen allesbehalve voor de hand. Gedeeltelijk omdat de definitie van "aanjager" binnen academische disciplines veel variatie vertoont, gaande van technische specificaties in "agent-based modeling" tot grote filosofische vragen over vrije wil. Om pragmatische redenen geef ik de voorkeur aan de middenweg en definieer ik "handelingsvermogen" als iemands vermogen om actief te interageren met zijn of haar omgeving. Technologie als aanjager van verandering is natuurlijk een metafoor.

We kunnen een lang debat beginnen over wie de "echte" aanjager van verandering was: was het de drukpers, zoals de krachtige titel van Eisensteins boek suggereert, of waren er vele aanjagers, namelijk de talrijke drukkers die hun werkplaatsen opzetten in verschillende Europese steden en degenen die hen financierden? En wat met de fervente lezers en de allianties of tegenstellingen die zich vormden tussen hen en de ideeën die ze trachtten te verspreiden? Bovendien kon de drukpers alleen slagen onder specifieke institutionele en culturele omstandigheden om de beoogde veranderingen teweeg te brengen. Blokdrukken in China dateert uit de 9de eeuw en drukken met beweegbare metalen letters werd in Korea al lang voor Gutenberg uitgevonden. Het is duidelijk dat technologieën geen "aanjager" kunnen zijn zonder de mensen die ze uitvinden, financieren, bedienen, verspreiden en blijven verbeteren. Een toevallige combinatie van verschillende actoren en van culturele en institutionele krachten moet samengaan met een technologische innovatie om een impact te genereren zoals die van de drukpers.

Wat de drukpers van andere technologieën onderscheidt, is de functie die ze aannam als katalysator van communicatie. Deze functie diende als kanaal voor de verspreiding van ideeën, waarvan vele nieuw en subversief waren voor de bestaande orde. Ze waren aantrekkelijk genoeg voor de elites, en voor degenen die ernaar streefden om deel uit te maken van de elite, om ze te aanvaarden en in te zetten voor de behartiging van hun belangen. De technologie bood de mogelijkheid om het brein van mensen in verafgelegen plaatsen te bereiken, zodat ze gemotiveerd en gemobiliseerd konden worden. Ze waren allemaal aanjagers van verandering, met verschillende belangen en doelen, maar verenigd in het zo goed mogelijk gebruiken van de technologie volgens hun intenties. Communicatie werd tegelijkertijd het middel en het doel, maar – zoals altijd – bleef de uitkomst onvoorspelbaar omdat ze open was.

Communicatie als katalysator voor vele doelen is ook een kenmerk van "AI als aanjager van verandering". Sinds de uitvinding van de drukpers zijn er veel nieuwe lagen toegevoegd aan de functie van communicatie. Op AI gebaseerde algoritmes voorspellen en worden steeds vaker ingezet bij het nemen van beslissingen. Maar het basisidee om door middel van specifieke inhoud het brein van de anderen te bereiken, wie en waar ze ook zijn, is blijven bestaan. AI/ML is in staat om dieper door te dringen in de cognitieve en emotionele toestand van gebruikers, van wie de gegevens nodig zijn om hen en alle anderen met wie ze verbonden zijn te *targeten*. Met voldoende gegevens kunnen zelfs degenen die geen sociale media gebruiken om te communiceren, worden geïdentificeerd. Al deze functies worden bereikt door informatie over iemands verleden, die blijkt uit de digitale sporen die de gebruiker heeft achtergelaten (dit betreft vandaag de dag bijna ieder van ons), op te halen, op te slaan, te verbinden en te verwerken. AI/ML heeft een indrukwekkend voorspellend vermogen verkregen via de extrapolatie van deze sporen uit het verleden, en die vervolgens kan worden gecombineerd met informatie over iedereen met wie we in het verleden hebben gecommuniceerd, waardoor een krachtig hulpmiddel ontstaat om de toekomst vorm te geven.

De hoeveelheid gegevens die beschikbaar is om algoritmes te trainen is duizelingwekkend. Om te voorkomen dat de beschikbare gegevens uitgeput raken, wordt nu al gebruik gemaakt van synthetische gegevens. AI/ML maakt het mogelijk om netwerken van netwerken te bouwen, opgebouwd uit verschillende soorten verbindingen en interacties. Zo wordt een enorme hoeveelheid informatie verzameld over wie we zijn, wat we doen, met wie, wanneer, en hoe we met elkaar omgaan en hoe we ons voelen. Dankzij sensoren in camera's en satellieten, die boven en onder de grond zijn geïnstalleerd, is AI/ML in staat om een spiegelwereld te bouwen van de fysieke en sociale wereld waarin we leven, en om interactie ermee mogelijk te maken. Bijna elk fenomeen en bestaand voorwerp is inmiddels digitaal gedocumenteerd of heeft een digitale handtekening die kan worden gevolgd, waardoor nieuwe verbanden kunnen worden gelegd via iteraties en bijna oneindige combinaties.

We kunnen concluderen dat AI een aanjager van verandering is, maar net als bij de drukpers is ze alleen een "aanjager" in de zin dat wij mensen er handelingsvermogen aan delegeren en toekennen. We laten ze voor ons presteren om door ons gestelde doelen te bereiken. We gebruiken ze om samen te komen en ons te onderscheiden. We delegeren er bepaalde taken aan, ons vaak niet bewust van de gevolgen die dit kan hebben. AI wordt aldus een uitbreiding van de menselijke mogelijkheden, maar daarmee gaan we tegelijk een ambivalente relatie met een open einde aan met een machine waarover we geen volledige controle hebben. We hebben het over "complementariteit" bij het uitvoeren van bepaalde taken, maar voelen ons ongemakkelijk over de toekomst, wanneer de machines door hun verbazingwekkend efficiënte prestaties steeds meer zouden kunnen overnemen van wat mensen voorheen deden.

De automatisering zal doorgaan, dit keer niet langer ter vervanging van fysieke arbeid, maar steeds meer van cognitieve taken. De autonomie die aan de machines wordt gegeven is nog steeds relatief. Ze zijn afhankelijk van mensen voor de enorme hoeveelheden energie die ze nodig hebben en voor onderhoud en reparaties. Ze hebben infrastructuur nodig, inclusief de organisatie om de onderneming te leiden, investeringsstrategieën en juridische en financiële afdelingen – de ingewikkelde hiërarchieën van de bedrijfswereld. Hun verdere ontwikkeling vereist nog steeds menselijke denkkracht en de vele toepassingen vereisen geschoolde arbeidskrachten die zich voortdurend bijscholen en goed zijn in multitasking. Maar de algemene trend wijst duidelijk in de richting van steeds meer terrein prijsgeven aan digitale machines.

Een machine is dus niets zonder de mensen erachter. Een machine is het door mensen gemaakte artefact dat het dichtst in de buurt komt van wat de natuur door de evolutie heen heeft gedaan – virussen produceren die zich niet alleen kunnen vermenigvuldigen. Een virus moet een cel infecteren om kopieën van zichzelf te maken. Een machine heeft mensen nodig om te kunnen blijven functioneren. Tegelijkertijd, zoals we met verbazing waarnemen, kan een digitale machine zichzelf trainen en zelf leren. Het handelingsvermogen dat we eraan hebben gedelegeerd, lijkt zich steeds verder uit te strekken, en dit roept de vraag op of we niet te veel hebben gedelegeerd, op welke gebieden we dat hebben gedaan en wat er moet gebeuren om een soort meta-controle te behouden.

Het concept "handelingsvermogen" onderzoeken is daarom een lastige taak. Het wordt meestal gedefinieerd als het vermogen van individuen om hun eigen beslissingen te nemen en verantwoordelijkheid te nemen voor hun acties. De sociologische definitie omvat het vermogen en de middelen van individuen om hun potentieel te vervullen. Maar kunnen deze of vergelijkbare definities van menselijk handelingsvermogen worden uitgebreid naar machines? En wat bedoelen we

als we handelingsvermogen overdragen aan AI? In technische termen worden machines ontworpen met verschillende niveaus van autonomie, wat betekent dat ze in staat zijn om gedurende bepaalde periodes en soms op afstand complexe taken uit te voeren met een aanzienlijk verminderde menselijke tussenkomst.

Met andere woorden, een autonoom systeem is een instrument of systeem (een machine of verzameling machines) dat wordt aangestuurd om te presteren in overeenstemming met het niveau van autonomie dat eraan is gegeven. In de praktijk kan dit behoorlijk angstaanjagende vormen aannemen, zoals nu gebeurt met de ingrijpende verschuiving die plaatsvindt in legers over de hele wereld – een verschuiving naar AI, robotica en autonome oorlogsvoering (*The Economist*, 6 juli 2023). Het is geen toeval dat er onlangs een discussie losbarstte over de vraag of de VN-Veiligheidsraad zou moeten overwegen om grenzen te stellen aan het delegeren van "commando- en besturingssystemen" aan autonome wapens, vergelijkbaar met de non-proliferatieverdragen die werden gesloten om de verspreiding van kernwapens tegen te gaan.

De angst dat mensen de controle kunnen verliezen over de machines die ze hebben ontworpen en gebouwd is niet nieuw en bestaat al sinds mensenheugenis. Homerus gebruikte reeds het woord "automaton" ("uit eigen wil handelend") om de automatische beweging van drievoeten op wielen te beschrijven. Geautomatiseerde poppen die op mensen of dieren lijken, werden gebruikt om de menselijke vindingrijkheid te demonstreren, om te vermaken en om te misleiden. De mythe van Frankenstein leeft voort in ontelbare verschijningsvormen. Ze is nieuw leven ingeblazen in beschaafdere, maar ook verraderlijkere vormen, in de "deep fakes" die door AI worden geproduceerd. Onder het mom dat ze "objectiever" is dan mensen, wordt ze nog steeds gevoed door de ondoorzichtige handelingen van AI, de welbekende "zwarte doos"-algoritmes. Er zijn technisch en wetenschappelijk onderbouwde argumenten naar voren gebracht om aan te tonen dat de "uitlegbaarheid" van AI niet mogelijk is (Lee, 2022). Zelfs de beste experts in de voorhoede van de ontwikkelingen op het gebied van generatieve AI geven publiekelijk toe dat ze de verbazingwekkende prestaties van LLM's (nog) niet volledig begrijpen en dat de vraag of ze "emergentie" produceren vooralsnog openblijft.

Of AI in de toekomst in staat zal zijn om volledig aan de menselijke controle te ontsnappen en volledig zelfstandig te handelen, is een van de vele speculaties om het publiek te waarschuwen voor een veelheid aan "existentiële risico's". Deze risico's, die zich in een verre en hypothetische toekomst bevinden, verbleken echter in vergelijking met de AI-gestuurde gevechtsschepen zonder bemanning of de zelfsturende dronezwermen die slechts twee voorbeelden zijn van de snel evoluerende technologieën die de toekomst van oorlog op dit moment vormgeven. Het zien van "vonken van kunstmatige algemene intelligentie" bij GPT-4 (Bubeck et al., 2023) of stellen dat generatieve AI op het punt staat "steeds krachtigere

digitale breinen te ontwikkelen en in te zetten die niemand – zelfs hun makers niet – kan begrijpen, voorspellen of betrouwbaar controleren", zoals werd geschreven in de open brief "Pause Giant AI Experiments" van 29 maart 2023, is een onverantwoord gebruik van een hype die alleen dient om de publieke discussie af te leiden van de serieuze zorgen en problemen die op dit moment moeten worden aangepakt.

Onze antropomorfische neigingen

Op een meer alledaags en praktisch niveau hebben mensen in hun interactie met artefacten altijd handelingsvermogen hieraan toegekend. Dit is diepgeworteld in onze antropomorfische neiging om het gedrag van een andere entiteit of voorwerp te zien in termen van mentale eigenschappen. Daniel Dennett heeft ons verteld hoe het werkt: "Eerst besluit je om het voorwerp waarvan het gedrag voorspeld moet worden te behandelen als een rationele entiteit; dan zoek je uit welke overtuigingen die entiteit zou moeten hebben, gezien haar plaats in de wereld en haar doel. Vervolgens bepaal je welke verlangens ze zou moeten hebben, op basis van dezelfde overwegingen, en tenslotte voorspel je dat deze rationele entiteit zal handelen om haar doelen na te streven in het licht van haar overtuigingen. Een beetje praktisch redeneren vanuit de gekozen set van overtuigingen en verlangens zal in de meeste gevallen een beslissing opleveren over wat de entiteit zou moeten doen; dit is wat je voorspelt dat de entiteit zal doen" (Dennett, 1989, p. 17).

Afgezien van de bewoordingen van de filosoof, is dit inderdaad de manier waarop we tegen het koffiezetapparaat of de computer praten als die "weigert" te doen wat we willen. We gebruiken elke dag antropomorfische taal in onze interacties met machines. Het is daarom niet verwonderlijk dat ChatGPT en zijn tegenhangers die zijn ontworpen om met mensen te communiceren, ons ertoe brengen om te zeggen dat ze "denken", "geloven" of "weten" – ook al begrijpen we dat het gaat om niet-denkende en niet-gelovende, en zeker niet-bewuste digitale artefacten die "slechts" zijn gemaakt om te doen alsof ze denken, begrijpen en geloven. Het ondoordachte gebruik van dergelijke woorden in het dagelijks taalgebruik blijft relatief onschuldig zodra dit verwijst naar vertrouwde technologieën die we al in onze wereld hebben opgenomen en waarmee we dus hebben leren leven. Tegelijkertijd beïnvloedt zo'n gebruik wel de manier waarop we de wereld waarnemen. Als het echter om AI gaat, kan dit de waarneming veranderen in een gevaarlijk dwingende illusie dat we in de aanwezigheid zijn van een denkend wezen zoals wijzelf.

Als we dit niet in de hand houden en er niet kritisch naar kijken, kunnen onze antropomorfische neigingen zich tegen ons keren en ernstige schade aanrichten. Dit werd op tragische wijze benadrukt door de zelfmoord in België van een man die een week lang gesprekken voerde met een "therapeutische" AI-toepassing (toegegeven, een oudere generatie dan ChatGPT). Het toeschrijven van hande-

lingsvermogen aan een AI-programma heeft blijkbaar bijgedragen aan de fatale beslissing van de gebruiker. In mijn boek *In AI We Trust* heb ik gewezen op een paradox die ontstaat wanneer we handelingsvermogen toekennen aan voorspellende algoritmes en beginnen te geloven dat hun voorspellingen zullen uitkomen. We gebruiken AI om onze controle over de toekomst en onzekerheid te vergroten, maar tegelijkertijd vermindert de performantie van AI, de macht die het heeft om ons te laten handelen op manieren die het voorspelt, ons handelingsvermogen met betrekking tot de toekomst. Dit gebeurt wanneer we vergeten dat wij mensen de digitale technologieën hebben gecreëerd waaraan we handelingsvermogen toekennen. Als hier niets aan wordt gedaan, zou het zelfs kunnen leiden tot de terugkeer van een deterministisch wereldbeeld waarin de meeste mensen geloven dat AI hen beter kent dan zijzelf, inclusief hun toekomst (Nowotny, 2021).

Sociale verandering: overgangen en omslagpunten

Het thema "AI als aanjager van verandering" bevat nog een andere vraag: hoe sociale verandering moet worden begrepen. Tegenwoordig horen we veel over de verschillende overgangen waarin we ons bevinden of die we zouden moeten nastreven. De EU heeft de "dubbele overgang" naar "groen" en "digitaal" als programmatisch doel vooropgesteld. Veel overheden hebben strategische programma's opgesteld om een grotere duurzaamheid te bereiken en over hoe de overgang te beheren om daartoe te komen. Toch is onze kennis van de processen die ten grondslag liggen aan maatschappelijke veranderingen en die tot een overgang kunnen leiden, eerder gebrekkig. We kunnen ze achteraf analyseren en bijvoorbeeld een aantal processen identificeren die tot omslagpunten leiden. Talrijke gevallenstudies over sociale verandering en over succesvolle of mislukte innovatie bieden interessante bevindingen, maar het empirisch bewijsmateriaal blijft meestal beperkt tot lokale gevallen. Deze gevallenstudies zijn vaak te klein in omvang, geografisch te wijdverspreid en institutioneel te verschillend, waardoor vergelijkbaarheid en generalisatie nauwelijks mogelijk zijn. Op macroschaal daarentegen kunnen simulaties van complexe adaptieve systemen, gebaseerd op wiskundige hulpmiddelen en voorzien van voldoende empirische gegevens, voorspellen wanneer en waar in een complex netwerk of systeem dergelijke omslagpunten waarschijnlijk zullen optreden. Ze worden gevolgd door een overgang of zelfs een ineenstorting van het systeem. De kloof tussen micro en macro blijft bestaan en als het aankomt op het begrijpen van maatschappelijke verandering lijkt het alsof we klem zitten.

Nochtans bevinden we ons te midden van maatschappelijke veranderingsprocessen die enorme gevolgen zullen hebben voor individuele levens en de toekomst van onze samenlevingen. Maatschappelijke verandering heeft vele dimensies en de impact ervan is ongelijk verdeeld over verschillende lagen en sectoren van een samenleving. Ze levert onvermijdelijk winnaars en verliezers op. Verandering gaat

gepaard met beloften en verwachtingen, waarvan sommige opzettelijk worden overdreven en andere impliciet inspelen op latente behoeften of onverzadigbare menselijke verlangens. Beloften zijn meestal moeilijk na te komen en eindigen vaak in een teleurstelling. Verwachtingen moeten zorgvuldig worden beheerd – een moeilijke taak, omdat nieuwe technologieën meestal worden omgeven door een hype en de neiging hebben om te veel te beloven. In het meer recente verleden zijn er al talloze voorbeelden: te beginnen met zelfrijdende auto's die "voor de deur stonden"; MOOC's (Massively Open Online Courses) die een "revolutie" in het hoger onderwijs zouden teweegbrengen; de metaverse die binnenkort ons leven in de fysieke wereld zou overnemen en de beloften van cryptomunten die menigeen tot roekeloze investeringen heeft verleid; om nog maar te zwijgen over de fantasieën van het transhumanisme die het eeuwige leven beloven. Het teken aan de horizon van een betere toekomst blijft hetzelfde: "Deze keer is het anders, geloof me."

De lange weg voor ons: AI als publiek goed

Maar, zoals de historische terugblik laat zien, deze keer is het anders – alleen begrijpen we de precieze betekenis daarvan nog niet. De ervaring van ingrijpende veranderingen in onze samenlevingen is alomtegenwoordig en de turbulentie die gepaard gaat met AI als aanjager en drager van verandering is even verontrustend als het vooruitzicht van een verdere versnelling van de verandering. De overheersende reactie tot nu toe is de tweedeling tussen degenen die techno-utopische visies aannemen en degenen die opgaan in hun dystopische opvattingen. Op een perverse manier voedt deze tweedeling de reeds bestaande polarisatie in onze samenlevingen, nog verergerd door de Covid-ervaring met de opkomst van de anti-vaxbeweging en het groeiende wantrouwen van burgers in hun regering en in experts. Bovendien zitten we gevangen in een somber vooruitzicht wat betreft klimaatverandering dat niet langer kan worden ontkend en worden we omringd door een economische recessie die op het punt staat te beginnen. Geopolitieke spanningen blijven toenemen terwijl de oorlog in Oekraïne voortduurt zonder vooruitzicht op een snel en goed einde. Wat staat ons te doen?

Een eerste stap is om af te stappen van het simplistische binaire utopisch-dystopische denkschema en een nuchtere beoordeling te maken van de risico's en de kansen. Dit zijn geen vaste categorieën. Ze vereisen eerder een waakzaam, flexibel en op wetenschap gebaseerd begrip van wat er op het spel staat, voor wie en onder welke omstandigheden. Misschien moet het concept van risico worden aangepast voor AI, omdat het niet langer voldoet aan de eenvoudige definitie uit het industriële tijdperk: de waarschijnlijkheid van een gebeurtenis vermenigvuldigd met de omvang van de schade. AI-risicobeheer en verantwoorde AI-praktijken zullen waarschijnlijk een belangrijk onderdeel worden van de toekomstige ontwikkeling van AI-systemen. Goede controles en rekening houden met de context zullen van cruciaal belang zijn (National Institute of Standards and Technology, AI 100/1, 2023).

AI/ML is een krachtige motor voor verandering, maar het is geen natuurkracht waaraan samenlevingen en burgers hulpeloos zijn blootgesteld. Ondanks vele institutionele gebreken en het slecht functioneren van bestaande instellingen, beschikken onze samenlevingen over voldoende middelen om risico's te "beheersen", mits de politieke wil er is. Ze kunnen en moeten de kansen grijpen die AI blijft bieden, zelfs als dat betekent dat ze uitdagingen moeten aangaan die de bestaande orde zullen verstoren of gevestigde belangengroepen omver zullen werpen. In de gezondheidszorg bijvoorbeeld biedt AI/ML enorme mogelijkheden voor gepersonaliseerde voorspellende geneeskunde (Hood & Price, 2023). Nu al biedt AI een grotere diagnostische nauwkeurigheid en behandelingsopties, en er komen nog snellere efficiëntievoordelen. Als dit niet zorgvuldig wordt gecontroleerd, krijgen de grote techbedrijven toegang tot gegevens in ruil voor AI-ondersteunde diensten op basis van contracten die nadelig kunnen zijn voor het publieke gezondheidssysteem, waardoor dit wordt onderworpen aan langdurige afhankelijkheden onder oneerlijke voorwaarden.

Toekomstige historici zullen de voor ons onvoorspelbare uitkomsten kunnen reconstrueren. Wat wij – als wetenschappers en als burgers – kunnen doen, is de kansen grijpen door de verschillende processen van maatschappelijke verandering in wording te observeren en analyseren. We kunnen nagaan welke wegen er zijn om schade te voorkomen en een redelijk evenwicht te vinden tussen risico's en kansen. Bovenal moet AI stevig worden geïnstitutionaliseerd als een publiek goed waarvan de voordelen voor iedereen beschikbaar moeten zijn (Boulton, 2021). We kunnen interventiepunten identificeren in de complexe assemblage van AI/ML als systeemtechnologie en in de fijnere technische en sociale details van de werking ervan en aanbevelingen doen voor te nemen acties. De kansen op succes hiervan zullen toenemen als we het belang kunnen laten zien van het samenbrengen van overheden, inclusief de wetgevende tak, beleidsmakers, de industrie, gemeenten, de media en de kunsten. Samen moeten we een vernieuwde publieke ruimte creëren, een soort 21ste-eeuwse agora die hersteld is van de bezetting, zo niet vernietiging, door sociale media, en de openstelling ervan bevorderen voor een publiek discours waaraan gewone burgers graag willen deelnemen.

De maatschappij kan alleen profiteren van AI als de voorwaarden voor het verzamelen, verwerken en bezitten van gegevens en het leveren van diensten niet worden gedicteerd door de grote internationale bedrijven en hun economische macht. In plaats daarvan moet AI worden gereguleerd door overheden en moeten de participatie en stemmen van burgers worden meegewogen. AI moet een publiek goed worden. De scheve verhouding tussen private en publieke financiering van AI-onderzoek moet worden aangepakt, omdat het universitair onderzoek momenteel wordt benadeeld bij de toegang tot de benodigde rekenkracht, gegevens voor het trainen van de algoritmes, het aantrekken van talent en het bepalen van de richting van toekomstig onderzoek.

Tot slot heeft de roep om een digitaal humanisme met een mensgerichte focus in alle AI-gerelateerde technologische ontwikkelingen alleen een kans om gerealiseerd te worden als er een robuust, geïnstitutionaliseerd kader bestaat om dit te ondersteunen (Vienna Manifesto on Digital Humanism, 2019). De bestaande instellingen werden in een andere tijd opgericht om met een andere reeks problemen om te gaan. De tijd is gekomen om serieus na te denken over een nieuw institutioneel kader dat beter is toegerust en in staat is om te gaan met de vele uitdagingen die AI/ML met zich meebrengt, en tegelijkertijd de basis te leggen voor het verder verkennen en op een meer rechtvaardige manier benutten van de mogelijkheden van deze technologie.

AI en de uitbesteding van kennisactiviteiten

"Het is onmogelijk om niet te communiceren", verklaarde Paul Watzlawick, de bekende communicatietheoreticus, en we communiceren inderdaad voortdurend en in veel verschillende vormen. Sommige zijn analoog (met verwijzing naar een object) en andere digitaal (logische en statistische verbindingen). We communiceren verbaal, maar ook via lichaamstaal. We dragen informatie over en wisselen gegevens uit, over onszelf, anderen en de wereld. Dit kunnen ideeën, praktijken en kennis zijn op verschillende niveaus van abstractie en complexiteit. Communicatie is een sociale praktijk die plaatsvindt in sociale omgevingen. Ze kunnen symmetrisch zijn, op ooghoogte en op gelijke voet, of juist sociale hiërarchieën benadrukken. Mensen hebben uitgebreide codes ontwikkeld die alle aspecten van het sociale leven doordringen om zichzelf van anderen te onderscheiden. Communicatie ligt aan de basis van de sociale organisatie van samenlevingen, die in de loop der tijd complexer is geworden.

Bovenal heeft communicatie de enorme groei van de menselijke kennis gestimuleerd als gevolg van de selectieve accumulatie van informatie die op verschillende manieren en voor verschillende doeleinden wordt overgedragen, verbeterd en doorgegeven. Nieuwe ideeën, kennis of praktijken worden gecombineerd en vervolgens opnieuw op nieuwe manieren gecombineerd, waarbij de inhoud in het proces van overdracht en uitwisseling door selectieve filters gaat. Deze filters zijn sociaal en cultureel. Ze volgen de normen en waarden in een samenleving die bepalen welk soort uitwisseling en inhoud cultureel en sociaal gewaardeerd en erkend worden. Samenlevingen vertrouwen op een expliciete of impliciete "kennishiërarchie". Voor AI laat de bekende "DIKW"-kennispiramide verschillende niveaus zien en tracht ze het verschil te verklaren tussen AI als kennisgedreven technologie en IT, die datagedreven is. De lagen van de piramide lopen op van gegevens naar informatie, gevolgd door kennis en wijsheid aan de top. In mijn boek *In AI We Trust* is een heel hoofdstuk gewijd aan de wijsheid die nodig is in de toekomst.

De technologieën die in deze kennishiërarchieën zijn ingebed controleren welke kennis en informatie circuleert. AI-algoritmes, zoals aanbevelingssystemen en

rangordes van prioriteiten, verfijnen deze filtermechanismen nog verder. Ze lijken technisch, maar zijn ontworpen om te voldoen aan de voorkeuren, waarden en belangen van de bedrijven die ze bezitten. De voortdurende controverses tussen grote techbedrijven en overheden over de vraag of de eerste wel genoeg doen om haat zaaiende uitingen in te dammen of te verwijderen, illustreert dat wie de media controleert ook de boodschap controleert, des te meer omdat de media de boodschap zijn geworden, zoals McLuhan terecht vaststelde. De katholieke kerk behield zich het recht voor om boeken op de index te zetten waarvan men vond dat de inhoud tegen de leer indruiste. Totalitaire regimes passen censuur toe, terwijl liberale democratieën in meer of mindere mate vasthouden aan het recht op "vrije meningsuiting". Toch classificeren ook zij bepaalde soorten informatie als "geheime" informatie waarvan de verspreiding de nationale veiligheidsbelangen in gevaar zou kunnen brengen.

De groeiende productie van menselijke kennis

Evolutie verloopt via variatie en selectie, en een soortgelijk mechanisme is aan het werk in de groei van de menselijke kennis. Selectieve filters werken niet alleen om uit te sluiten, door te controleren wat niet mag worden gecommuniceerd, maar werken actief aan het opnemen, absorberen en verbeteren van die communicatie die nieuwe kennis zal opleveren. Het egaliserende effect van gedrukte edities, in plaats van fluctuerende en instabiele producten van kopiisten, was essentieel voor de cumulatieve cognitieve vooruitgang en incrementele verandering waardoor echte wetenschappelijke groei wordt gekenmerkt (Eisenstein, 1980, p. 412).

De groei van de menselijke kennis wordt sterk bevorderd door technologieën die het uitbesteden, of outsourcen, van kennisactiviteiten mogelijk maken: het verwerken en toepassen van kennis op andere domeinen; de opslag en het beheer van gegevens; de verspreiding van bevindingen; nieuwe combinaties en het hergebruiken van kennis. Deze en andere activiteiten, evenals de infrastructuur en processen die eraan ten grondslag liggen, zijn essentieel voor de selectieve opname en de verdere bewerking van kennis via communicatiepraktijken. Kennis-activiteiten verspreiden wat bekend is in de tijd en de ruimte, wat niet mogelijk zou zijn zonder outsourcingtechnologieën. De geschiedenis van de mensheid en wat ze tot nu toe heeft kunnen bereiken is ook een geschiedenis van de outsourcingtechnologieën die werden ingezet voor de groei van de kennis.

Nergens is dit duidelijker dan in de moderne wetenschap. Een van de kenmerken daarvan was het openbaar maken en delen van kennis, wat een radicale breuk betekende met de traditie van geheimhouding van kennisbezitters in vroegere tijden. Door de wetenschappelijke bevindingen en de manier waarop ze tot stand kwamen zichtbaar en voor iedereen toegankelijk te maken, werden nieuwe communicatiekanalen geopend die enorm bijdroegen aan de verspreiding van kennis en het wetenschappelijke wereldbeeld. De wetenschap volgde daarbij haar

eigen epistemische waarden, terwijl ze zorgvuldig de grenzen afbakende waarover ze cognitieve en maatschappelijke autoriteit claimde. Een van de epistemische waarden voor het ontwikkelen en toegankelijk maken van wetenschappelijk onderzoek ligt aan de basis van de praktijken van reproduceerbaarheid, het thema van de Denkerscyclus van 2022 (Leonelli & Lewandowsky, 2022). De wetenschap heeft uitgeblonken in het optimaliseren van haar uitbestedingspraktijken. Dit is de reden waarom de wetenschappelijke gemeenschap er zeer waarschijnlijk snel in zal slagen om gebruik te maken van de mogelijkheden die AI/ML bieden, of het nu gaat om het ontdekken van medicijnen of op literatuur gebaseerde ontdekkingen, numerieke weersvoorspelling, het zoeken naar nieuwe materialen voor batterijen, het ontwerpen van nieuwe experimenten of het verder automatiseren van laboratoria.

De uitvinding van het schrift als uitbesteding van een kennisactiviteit

AI als aanjager van sociale verandering kan daarom worden gezien als een integraal onderdeel van het lange traject van het uitbesteden van kennisactiviteiten met behulp van technologieën. Het begon allemaal met de uitvinding van het schrift, dat de overgang markeerde van orale culturen naar schriftculturen. Het schrift is meerdere keren uitgevonden, onafhankelijk van elkaar, op verschillende plaatsen en momenten. Het is een verzameling van elementen die de uitvinding en beheersing omvat van symbolen zoals hiërogliefen, spijkerschrifttekens en alfabetten; de gedetailleerde uitwerking van de fysieke substraten en infrastructuur die nodig waren voor de productie, logistiek, levering en het gebruik van geschikte materialen, zoals klei, steen, papyrus, dierenhuiden en andere; de sociale competentie en vaardigheden voor samenwerking en arbeidsverdeling, zoals de specialisatie van kopiisten, de overdracht van vaardigheden en van interpretatieve capaciteiten.

Samen vormen deze elementen een geheel dat communicatie in tijd en ruimte efficiënter heeft gemaakt. Kennis die voorheen alleen in het geheugen van individuen en hun mondelinge communicatievaardigheden aanwezig was (zelfs met de hulp van mnemotechnische middelen) en mondeling van generatie op generatie werd overgedragen, kon nu worden uitbesteed en op een fysiek medium worden vastgelegd. Een redenaar had de vrijheid (en hij werd geacht die te nemen) om de inhoud aan te passen aan de gelegenheid en het publiek tot wie hij zich richtte, terwijl de woorden die waren gegraveerd in steen, op papyrusrollen of op palmbladeren een afstand in de tijd creëerden tussen het moment waarop ze waren geschreven en het moment waarop ze werden gelezen en geïnterpreteerd. Het is aannemelijk dat de nieuwe uitbestedingspraktijken ook hebben bijgedragen aan het vermogen van onze voorouders om abstracte symbolen te bedenken en te gebruiken, wat aanleiding heeft gegeven tot de wiskunde. Het zwarte (of witte) schoolbord dat nog steeds door wiskundigen wordt gebruikt als het belangrijkste medium om met elkaar te communiceren, ondersteunt deze hypothese.

De sociale en epistemische implicaties van het schrift waren enorm. Voor het eerst werd taal gecodeerd in symbolen die niet alleen op nieuwe manieren gelezen, geïnterpreteerd, begrepen, doorgegeven en gedeeld konden worden, maar ook ingezet konden worden voor een reeks nieuwe doeleinden. Metingen en cijfers floreerden en wonnen aan belang. In de oudheid hadden goden hun standbeelden en tempels die aan hen waren gewijd, terwijl het schrift vooral werd gebruikt voor belastingen en handel. Pas met de opkomst van de monotheïstische religies werd het geschreven woord de basis van heilige geschriften. Woorden konden zich verspreiden zonder dat een mens ze uitsprak. Er ontstonden nieuwe netwerken voor de overdracht, de handel werd geografisch uitgebreid en de meting van de graanoogst die moest worden belast kreeg een aanzienlijke impuls. Voor het eerst ontstond er een directe confrontatie met het verleden zoals het op schrift was vastgelegd. Dit beperkte de mondelinge interpretatieve flexibiliteit, maar versterkte het gewicht dat aan het geschreven woord werd toegekend. Schriftelijke contracten bleken betrouwbaarder dan mondelinge, met verdere gevolgen voor de handel, maar ook voor vredesonderhandelingen.

Omdat de bronnen schaars en het materiaal kostbaar waren, versterkte de controle erover de centralisatie van de interpretatieve autoriteiten, wat leidde tot een machtsconcentratie in de handen van een kleine elite van priesters, kopiisten en heersers. Bibliotheken werden de opslagplaatsen van alle beschikbare kennis en hun achteruitgang of vernietiging betekende een aanzienlijk verlies van kennis. Misschien ook voor het eerst werd duidelijk dat een nieuwe technologie gepaard ging met het verlies van bepaalde cognitieve vaardigheden die mensen voorheen bezaten. Zoals bekend betreurde Plato de uitvinding van het schrift omdat dit het vermogen van mensen verminderde om een enorme hoeveelheid kennis uit het hoofd te leren.

Wat kunnen de mechanismen en patronen die in deze eerste fase van de uitbesteding van kennisactiviteiten naar voren komen ons vertellen? Hoe wordt een sociale technologie – het schrift – een aanjager van verandering? Er is geen centraal, coördinerend mechanisme. Zoals blijkt uit de herhaalde keren dat het schrift onafhankelijk van elkaar werd uitgevonden, is de menselijke vindingrijkheid actief en produceert ze symbolen om mee te communiceren en te handelen. Wiskunde zoals wij ze kennen is ondenkbaar zonder het schrijven van symbolen. Uitbesteding betekent dat er nieuwe ruimtes voor communicatie en actie worden gecreëerd, die nieuwe mogelijkheden bieden en andere beperken. Sommige van deze ruimtes zullen veranderen in "creatieve niches", waar de technologie wordt ingezet voor nog uit te vinden doeleinden. Net als bij elke andere technologie worden het gebruik en de voordelen van het uitbesteden van kennisactiviteiten bepaald door de bestaande sociale en economische machtsstructuren. In een zeer scheve, ongelijke samenleving zullen de voordelen onevenredig ten goede komen aan degenen die macht hebben. Zij zullen proberen zich de technologie toe te eigenen en zullen deze niet gebruiken om verandering teweeg te brengen, maar om hun machtsbasis te consolideren.

En toch is het algehele effect een uitbreiding van de kennisbasis. Bibliotheken werden de fysieke opslagplaatsen, die in eerste instantie alleen toegankelijk waren voor de elite, maar ze blijven de bewakers van een belangrijk deel van het menselijk verleden en vertellen ons wat eerdere samenlevingen waardeerden en hoe ze de wereld zagen en begrepen. Het schrift vormt tot op de dag van vandaag de basis voor de heilige geschriften van de monotheïstische religies en het is moeilijk om je hun invloed voor te stellen zonder het schrift. Door het woord uit te besteden aan een materieel substraat, konden woorden zich dus losmaken van de lokale context waarin ze waren ontstaan en ze konden dienen om kennis over te dragen en uit te wisselen met mensen in verre oorden en met geesten die ze gretig ontvingen, betwistten of zich toe-eigenden. De richtingen die de uitbestede kennisactiviteiten insloegen en de effecten die ze teweegbrachten, konden echter niet nauwkeurig worden voorspeld.

Van de boekdrukkunst als uitbesteding tot sociale media

De tweede fase van het uitbesteden van kennisactiviteiten werd ingeluid door de boekdrukkunst, die de uitwisseling en verspreiding van nieuwe ideeën met een ongekende snelheid en reikwijdte vergemakkelijkte. Er ontstonden nieuwe doelgroepen en industrieën rond het publiceren. Door de uitbesteding op grote schaal aan boeken die in grote aantallen werden geproduceerd, konden oudere teksten worden herzien en bijgewerkt om er nieuwe kennis in op te nemen, konden banden worden gesmeed tussen een lezerspubliek dat wijdverspreid was over Europa en konden sociale bewegingen worden gevormd en gemobiliseerd. Dit droeg ook bij aan het uitbreiden van de geletterdheid als sleutel tot toegang tot de buitenwereld en veranderde de houding ten opzichte van leren. Het gaf aanleiding tot een positieve spiraal die de weg vrijmaakte voor meer inclusie en participatie. De komst van de druktechnologie viel samen met de Europese ontdekkingsreizen over de hele wereld en bevorderde een grotere openheid naar een meer kosmopolitische blik, die het stellen van vragen en de verspreiding van nieuwe ideeën stimuleerde.

Zoals Eisenstein in detail beschrijft, bracht de boekdrukkunst een diepgaande culturele mentaliteitsverandering teweeg, die deze periode uiteindelijk markeert als een cruciaal keerpunt in de Europese geschiedenis. De uitbesteding van kennis in boeken, kranten, pamfletten en illustraties betekende dat kennis niet langer kon worden gemonopoliseerd door de elite, maar een (relatief) massapubliek zou bereiken van geletterden die in aantal toenamen. Dit had een grote invloed op de renaissance, met de heropleving van de klassieke literatuur; op de protestantse reformatie, omdat het de interpretatie van de Bijbel door elke lezer mogelijk maakte en zo vorm gaf aan religieuze debatten; op de wetenschappelijke revolutie, omdat het drukken de kritische vergelijking van teksten en illustraties mogelijk maakte; en door de snelle uitwisseling van nieuwe ontdekkingen en experimenten aan te moedigen, waardoor de Republiek der Letteren ontstond (Eisenstein, 1980).

Er dient te worden opgemerkt dat sommige van de huidige zorgen ook al bestonden tijdens de culturele omwenteling die de drukpers een paar eeuwen geleden teweegbracht. Religieuze en politieke pamfletten stonden vol haat en gemene aanvallen op tegenstanders (Darnton, 1984); nepnieuws circuleerde op grote schaal, hoewel veel langzamer en meer lokaal beperkt dan tegenwoordig. De Europese verlichting had haar schaduwzijde als het ging om het uitbreiden van haar geclaimde universalisme naar de koloniën buiten het metropolitaanse gebied. Het recht op "vrije meningsuiting" moest nog grondwettelijk worden vastgelegd, terwijl het vandaag de dag in de VS, in een perverse wending, wordt gebruikt om te pleiten voor een bijna onbeperkte vrijheid om racistische en haatdragende meningen te uiten op sociale media. Het is veelzeggend dat in de emblematische confrontatie tussen kerk en wetenschap het proces tegen Galileo Galilei niet ging over de vraag of de wetenschap juist of fout was. Hij moest zijn leer afzweren omdat hij ervan beschuldigd werd dat hij de voorwaarden had geschonden die de kerk had opgelegd voordat ze de publicatie van de *Dialoog over de twee voornaamste wereldsystemen* in 1632 toestond.

De diepgaande transitie die we nu meemaken, veroorzaakt door de verbazingwekkende vooruitgang in AI/ML, komt overeen met de evolutie van het uitbesteden van kennisactiviteiten in eerdere fasen. De effecten ervan zullen echter in verschillende mate groter zijn. De uitbesteding is niet langer beperkt tot woorden op materiaal neerschrijven en ze door de tijd te laten reizen, noch tot het verspreiden van ideeën via goedkoop papier naar een nieuw gecreëerd publiek. Gezien de tijdschaal van de vorige fasen fungeerden de informatie- en communicatietechnologieën van het einde van de 19de en de 20ste eeuw – telefoon en telegraaf, radio en tv – slechts als voorspel voor vandaag. Ze luidden het verkleinen van de afstand over de hele wereld in, terwijl ze het bewustzijn van wat er elders gebeurde vergrootten. De massamedia introduceerden de communicatie van één naar velen, gevolgd door de communicatie van velen naar velen, individuele targeting en door gebruikers gegenereerde inhoud toen het internet het overnam, gevolgd door de alomtegenwoordige verspreiding van sociale media.

Generatieve AI: de uitbesteding van de kennisproductie

De grote sprong in de uitbesteding van kennisactiviteiten op basis van LLM's ligt in het feit dat de productie van kennis zelf wordt uitbesteed. Door het trainen en aanleren van zelftraining aan steeds geavanceerdere algoritmes aan de hand van triljoenen tokens, bestaande uit alle teksten, beelden en geluiden die beschikbaar zijn op het internet, heeft de mens de productie van nieuwe kennis gedelegeerd aan de machines die door hem zijn ontworpen en gebouwd. Hoewel het "slechts" gaat om extrapolaties uit het verleden en het gebruik van waarschijnlijkheden, resulteert de combinatie in het genereren van iets nieuws. Of de antwoorden juist zijn of verzonnen, op feiten berusten of hallucinaties zijn, is een andere kwestie

die kritisch moet worden beoordeeld. Als automatisering door AI bestaat uit het uitbesteden van zware of vervelende fysieke taken van mensen aan machines, dan neemt generatieve AI een toenemend aantal en soorten cognitieve taken over die aan AI worden uitbesteed. ChatGPT is ontworpen als dialoog met een digitale Andere en het is door de dialoog – de vragen die gesteld worden, de sturing door ingenieurs – dat er nieuwe kennis ontstaat. In het licht van het feit dat het uitbestedingsproces begon met een verschuiving van een orale naar een geschreven cultuur, is het een ironische wending in de geschiedenis dat generatieve AI een gedeeltelijke terugkeer naar een orale cultuur betekent. Het wordt opnieuw belangrijk om te weten hoe je een dialoog en een gesprek voert, dit keer met een machine.

Het uitbesteden van de kennisproductie aan digitale machines brengt een reeks uitdagingen met zich mee; enkele van de meest dringende zullen later in dit rapport worden behandeld. De voordelen van deze laatste en meest radicale stap in het uitbestedingsproces zijn enorm en de integratie ervan in onze individuele levens en in het functioneren van onze samenlevingen heeft een explosief potentieel. AI/ML wordt bijvoorbeeld al gebruikt om de meest veelbelovende "cocktail" van medicijnen te vinden voor de precieze behandeling van specifieke, zeldzame soorten kanker. Daarbij presteert de technologie beter dan de meest ervaren arts, door de toegang tot een schat aan medische literatuur, inclusief de meest recente. Dit werpt de fundamentele vraag op hoe artsen in de toekomst zullen worden opgeleid. Worden ze supervisors van de AI? Misschien. Gelijkaardige vragen duiken op in veel andere toepassingsgebieden waar de voordelen duidelijk zijn, maar de rol van de mens steeds ongrijpbaarder wordt en dringend opnieuw gedefinieerd moet worden.

Misschien is het grootste, onbedoelde en onderschatte geschenk van generatieve AI wel dat hierdoor een reeks fascinerende nieuwe onderzoeksvragen wordt geopend. Ze variëren van diepgaande verkenningen van hoe het menselijk brein werkt bij het oplossen van taken in vergelijking met dat van AI, over vragen over de toekomstige evolutie van taal eens LLM's alomtegenwoordig zijn geworden in het dagelijks leven, de impact van steeds intiemere en intensere interacties met AI (met name op de jongere generatie en de vorming van de identiteit), tot vragen over de impact van AI op liberale democratieën en wat er gedaan kan worden om verdere erosie ervan te stoppen.

Naast dergelijke onderzoeksvragen en het lanceren van nieuwe onderzoeksgebieden, heeft de wetenschap een belangrijke rol te spelen in het uitleggen van de werking ervan aan het publiek. De natuurkundige Richard Feynman zei ooit: "Wetenschap is wat we hebben geleerd over hoe we kunnen vermijden onszelf voor de gek te houden". Gezien het ontwerp van ChatGPT om te doen geloven dat je met een mens communiceert en gezien onze antropomorfische neigingen, is het nog belangrijker voor de wetenschap om Feynmans inzicht onder de aandacht te

brengen. De pandemie maakte pijnlijk duidelijk hoe weinig politici en het publiek begrijpen dat wetenschap georganiseerd scepticisme is en dat het in twijfel trekken van beweringen over wetenschappelijke bevindingen in een uitgebreid proces van verificatie en validatie een essentiële epistemische deugd van de wetenschap is en geen fout.

Daarom is het in begrijpelijke en toegankelijke termen uitleggen hoe je op een kritische maar constructieve manier kunt denken in de context van AI/ML een van de belangrijkste verantwoordelijkheden van wetenschappers. Hoe reageert een wetenschapper als mensen haar vertellen dat "AI mij beter kent dan ikzelf" en als ze beginnen te geloven dat AI een entiteit is wiens voorspellingen onvermijdelijk waar zullen blijken te zijn? De stilzwijgende aanname in wetenschapscommunicatie is nog steeds dat zodra een bepaald niveau van digitale geletterdheid is bereikt, burgers rationeel zullen handelen en digitale oplossingen en de bijbehorende gedragsaanbevelingen zullen overnemen. Maar zo werkt het niet. Empirisch onderzoek heeft aangetoond dat we moeten afstappen van het "deficit model" van wetenschapscommunicatie, dat de weigering om te accepteren wat wetenschappers zeggen, toeschrijft aan onbegrip (Wynne, 1993). In plaats daarvan moeten we, voor een toegankelijke interactie met het publiek, tonen en uitleggen hoe de wetenschap AI als ondersteuning van wetenschappelijke arbeid kan inzetten. Wetenschappers bevinden zich in een unieke positie omdat ze AI al op grote schaal gebruiken om te helpen bij hun onderzoek. Ze kunnen concrete voorbeelden laten zien en de voordelen die daaruit voortvloeien, of het nu gaat om medisch, milieu- of ander onderzoek. Tegelijkertijd moeten ze communiceren over hoe het werkt, zodat "de wetenschap ervoor zorgt dat we onszelf niet voor de gek houden".

In dit rapport heb ik een beeld geschetst, op basis van mijn observaties en analyse, van "AI als aanjager van maatschappelijke verandering". Na inspirerende en intensieve discussies met groepen van stakeholders en de stuurgroep van de KVAB zijn we het eens geworden over de volgende uitvoerbare aanbevelingen.

Aanbeveling 1:

We bevelen aan een brede publiekscampagne te lanceren onder het motto "AI voor burgers – burgers voor AI" om burgers te ondersteunen bij het gebruiken van AI in hun dagelijkse leven en voor een betere samenleving.

Het doel is om het begrip van de werking van AI en digitale systemen te verdiepen en te verspreiden, het potentieel van huidige en toekomstige toepassingen en het gebruik ervan te onderzoeken, en te leren over hun beperkingen.

De vele reeds bestaande en nieuwe initiatieven moeten een officieel mandaat krijgen om

1. de educatieve inspanningen gericht op deze doelen onderling te coördineren;

2. hun respectievelijke doelgroepen (leeftijdsgroepen, formele en informele settings); de middelen en materialen die ze gebruiken, testen en ontwikkelen (bv. voor leraren in het basis- en secundair onderwijs); en vormen van samenwerking met universiteiten, media, de kunsten en het bedrijfsleven te specificeren en in kaart te brengen;
3. voldoende ruimte te creëren voor een voortdurende uitwisseling van ervaringen en wederzijds leren, over academische disciplines en generaties heen;
4. ervoor te zorgen dat alle onderwijsinspanningen een digitaal humanistisch perspectief bevatten (en dus veel verder gaan dan 'digitale geletterdheid') https://informatics.tuwien.ac.at/digital-humanism/

Daartoe moet een solide institutioneel kader worden opgezet en voorzien van de nodige financiële en personele middelen, in eerste instantie voor een periode van drie jaar, en hernieuwbaar na evaluatie.

Aanbeveling 2:

We bevelen aan om fundamenteel AI-onderzoek een hoge prioriteit te geven en uit te voeren volgens de lijnen van de Europese Onderzoeksraad (ERC) (bottom-up, uitgaand van een hoofdonderzoeker). Dit dient als tegenwicht voor de dominantie van een eendimensionaal "technologisch oplossingsdenken", dat alternatieven negeert en/of terzijde schuift bij de keuze van onderzoeksproblemen, methoden en technieken. Hierdoor ontstaat bovendien een meer humanistisch begrip van de reikwijdte en de diepte van de menselijke ervaring en wat het betekent om mens te zijn.

De huidige overconcentratie van de financiering van AI-gerelateerd O&O in de private sector leidt tot een zorgwekkend onevenwicht voor (voornamelijk) universitair onafhankelijk onderzoek met betrekking tot de toegang tot rekenkracht, trainingsgegevens, het aantrekken van talent en het pionieren in nieuwe onderzoeksrichtingen. In het belang van AI als een publiek goed moeten deze nadelen worden aangepakt.

Als onderzoeksgebied is AI, inclusief machine leren en generatieve AI, relatief jong, terwijl een historisch perspectief grotendeels ontbreekt, vooral in Europa. Hierdoor bestaat er een grote kans op het verlies van waardevolle technische kennis, wiskundige concepten, technieken en wetenschappelijke inzichten. Veelbelovende onderzoekslijnen werden vaak voortijdig afgesloten. Alleen een sterke focus op fundamenteel onderzoek kan de aanzet geven tot hun herontdekking en de verdere verkenning van historische paden die niet werden bewandeld.

Aanbeveling 3:

We bevelen een krachtige ondersteuning aan van onderzoek naar de maatschappelijke impact van AI wat betreft aspecten en gebieden die

naar alle waarschijnlijkheid niet zullen worden opgepakt door de grote internationale bedrijven.

Omdat we nog maar aan het begin staan van het systematisch volgen en analyseren van de mogelijke nuttige toepassingen van AI voor verschillende groepen in de samenleving en het leren over het vermijden van sociale schade, is het cruciaal om de snel evoluerende ervaringen, stemmen en behoeften van burgers mee te nemen.

Studenten AI en aanverwante technische gebieden (en hun docenten) moeten worden aangemoedigd om een perspectief van digitaal humanisme op te nemen in hun technische opleiding en praktijk. Ook studenten in de geesteswetenschappen en sociale wetenschappen (en hun docenten) moeten meer vertrouwd raken met de technische aspecten.

Dit zijn de voorwaarden voor een meer en beter gefundeerde inter- en zelfs transdisciplinariteit, die dringend nodig is.

Referenties:

Arendt, H. (1951) *The Origins of Totalitarianism*. Berlin: Schocken Books.

Boulton, G.S. (2021) Science as a Global Public Good. International Science Council Position Paper, https://council.science/wp-content/uploads/2020/06/Science-as-a-global-public-good_v041021.pdf

Bubeck, S. et al. (2023) Sparks of Artificial General Intelligence: Early Experiments with GPT-4, (24.3.2023) https://arxiv.org/abs/2303.12712

Darnton, R. (1984) *The Great Cat Massacre and Other Episodes in French Cultural History*. New York City: Basic Books.

Dennett, D.C. (1989) *The Intentional Stance*. Cambridge: The MIT Press.

Eisenstein, E.L. (1980) *The Printing Press as an Agent of Change*. Cambridge: Cambridge University Press.

Hood, L. and Price, N. (2023) *The Age of Scientific Wellness. Why the Future of Medicine is Personalized, Predictive, Data-Rich, and in Your Hands*. Harvard University Press: Cambridge, Mass.

Lee, E. A. (2022) Limits of Machines, Limits of Humans. DigHum Lecture, https://caiml.dbai.tuwien.ac.at/dighum/dighum-lectures/edward-lee-limits-of-machines-limits-of-humans-2022-05-24/

Leonelli, S. and Lewandowsky, S. (2022) The Reproducibility of research in Flanders: Fact finding and recommendations. KVAB Thinkers' report 2022.

National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0): https:doi.org/10.6028/NIST.AI.100-1.

Nowotny, H. (2021) *In AI We Trust. Power, Illusion and Control of Predictive Algorithms*. Cambridge, UK: Polity Press.

Perez, C. (2018) Second Machine Age or Fifth Technological Revolution? (Part 4) The Historical Patterns of Bounty and Spread. (21.11.2018). https://medium.com/iipp-blog/second-machine-age-or-fifth-technological-revolution-part-4-4420c29ceed

Prins, C., Sheikh, H., Schrijvers, E., de Jong, E. and Steijns, M. (2021), Mission AI. The New System Technology. Summary of the Dutch report Opgave ai. De nieuwe systeemtechnologie published by the Netherlands Scientific Council for Government Policy www.wrr.nl

The Economist, (2023) A new era of high-tech war has begun, The Future of War. (06.07. 2023). https://www.economist.com/leaders/2023/07/06/a-new-era-of-high-tech-war-has-begun

Vienna Manifesto on Digital Humanism, (2019) https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto/

Wynne, B. (1993) Public uptake of science: a case for institutional reflexivity. *Public Understanding of Science*, 2 (4), pp.321-337.

Uzogara, E. (2023) Democracy Intercepted. Did platform feeds sow the seeds of deep divisions during the 2020 US presidential election? *Science* Vol. 381 Issue 6656, 28 July 2023, pp.386-387.

3. Reflecties van experts

Grote taalmodellen: De opkomst van de dagdromende zombies

Walter Daelemans, Universiteit Antwerpen

Helga Nowotny's tekst "AI as an Agent of Change" is een welkome herinnering aan de onvoorspelbare gevolgen die een nieuwe technologie kan hebben op de samenleving. Het beste wat we kunnen doen is scenario's voor onbedoelde gevolgen bespreken en ons voorbereiden om ons daaraan aan te passen. Maar dit doen voor AI is niet zonder problemen omdat het geen nieuwe technologie is en het label wordt gebruikt voor een breed scala aan toepassingen. Ik zal beargumenteren dat we, om de negatieve gevolgen te verzachten, meer moeten investeren in publiek AI-onderzoek in plaats van te proberen om dit soort onderzoek tegen te houden.

Er zijn al veel voorbeelden van goede en slechte AI. Er zijn al autonome "*rovers*" en helikopters actief op Mars, maar we hebben ook de eerste autonome wapens in gebruik en de VN is er nog niet in geslaagd om ze te laten verbieden. We hebben *deep fakes*, maar we hebben ook meer dan tien jaar ervaring met bruikbare machinevertalingen. Dit heeft de werkgelegenheid voor vertalers niet negatief beïnvloed, maar juist nieuwe markten en toepassingen geopend. Deze systemen bieden een enorme stimulans voor communicatie en begrip over taalbarrières heen, van internationale handel en wetenschap en entertainment tot hulp aan vluchtelingen en migranten. Spraaksynthese (tekst naar spraak) en spraakherkenning (transcriptie) zijn ook een belangrijke hulp geweest voor mensen met perceptuele beperkingen en zullen voornamelijk positieve effecten hebben in veel nieuwe contexten dankzij verdere kwaliteitsverbetering.

De plotselinge beschikbaarheid van ChatGPT voor het grote publiek aan het eind van 2022 veroorzaakte grote bezorgdheid over de gevaren van AI, en ik zal me hier richten op die grote taalmodellen (LLM's), en meer in het algemeen op generatieve AI (GenAI) en op scenario's waarin ze een rol spelen, ten goede of ten kwade.

Ondergang door AI is niet nabij

LLM's zijn gebaseerd op een lange onderzoekstraditie in Natural Language Processing (NLP). Statistische taalmodellen, de voorlopers van ChatGPT, waren verantwoordelijk voor de eerste bruikbare automatische vertaal- en spraakherkenningstoepassingen vanaf de jaren negentig. Een statistisch taalmodel kent een waarschijnlijkheid toe aan een reeks woorden. Dit is handig wanneer de beste vertaling of de beste spraaktranscriptie moet worden gekozen uit vele mogelijkheden. Het exponentieel vergroten van de schaal ervan, zowel wat betreft de omvang van het taalmodel (het aantal parameters) als de hoeveelheid

gegevens waarmee het getraind wordt, maakte echter een enorm verschil. LLM's vertonen nu "emergent" gedrag dat veel verder gaat dan het toewijzen van waarschijnlijkheden aan brokken tekst. De modellen begrijpen, genereren, vertalen en transformeren lange teksten, veranderen hun stijl of complexiteit, vatten ze samen en beantwoorden er vragen over. Ze kunnen goed programmeren, zijn empathisch en creatief en redeneren op basis van gezond verstand. Men kan zeggen dat ze tekst "begrijpen" en "intelligent" zijn, hoewel dat natuurlijk vooral een (filosofische) terminologische kwestie is. Als begrip en intelligentie functionele concepten zijn (zoals vliegen), dan is de simulatie ervan in alle opzichten gelijkwaardig aan de werkelijkheid. Net zoals een vliegtuig de functie van vliegen anders kan uitvoeren dan een vogel, kunnen LLM's functioneel tekst "begrijpen" zonder menselijk te zijn. Maar LLM's zien de wereld alleen door de filter van teksten die door mensen zijn geschreven, zowel fictie als non-fictie. Ze hebben duidelijk geen handelingsvermogen, emoties, bewustzijn, meningen of doelen (afgezien van het doel om een prompt optimaal te voltooien), ook al zeggen ze van wel, en ze leven in een droomwereld zonder basis in de realiteit (vandaar "dagdromende zombies", nauw verwant aan de filosofische zombies). Het is ook moeilijk te zien hoe een dergelijke basis en zelfbewustzijn zouden kunnen voortkomen uit de objectieve functies waarmee taalmodellen momenteel worden getraind, zelfs met meer of multimodale gegevens. Al lijkt de dreiging van "ondergang van de mensheid" door bovenmenselijke intelligentie nog ver weg, er is wel degelijk reden tot zorg.

Verdwijnende vaardigheden

Op LLM gebaseerde AI zal prominent aanwezig zijn in de maatschappij in de vorm van assistenten. Onderwijsassistenten, studieassistenten, programmeursassistenten, doktersassistenten enz. Op middellange termijn zal dit niet noodzakelijk negatieve gevolgen hebben voor de werkgelegenheid of voor onze kijk op wat de moeite waard is om te bestuderen. Net als bij machinevertaling zullen de productiviteit en de kwaliteit toenemen, wat de weg vrijmaakt voor nieuwe mogelijkheden. AI kan zelfs een democratiserend effect hebben. De deskundige programmeur of copywriter heeft de grootste kans om veel te verliezen omdat in die vakgebieden iedereen in staat zal zijn om een hoger kwaliteitsniveau te bereiken. Maar op de lange termijn, met nog geavanceerdere AI-modellen, zal de maatschappij zich misschien moeten aanpassen aan vooralsnog ongeziene automatiseringsniveaus van cognitieve taken, en men zal terughoudend moeten zijn ten aanzien van het opdoen van expertise in vaardigheden waarin AI-systemen toch beter zijn. Het is nu nog moeilijk voor te stellen, maar menselijk schrijven en programmeren kan op een dag net zo achterhaald zijn als hoofdrekenen.

De behoefte aan geavanceerde spamfilters

Onmiddellijke aandacht verdient de manier waarop LLM's kunnen worden gebruikt om mensen te beïnvloeden, over te halen, aan te vallen en te manipuleren

door tekst- en beeldgebaseerde GenAI te combineren. Het gaat hier om nepmensen die zijn gemaakt door echte mensen. De maatschappij zal moeten beslissen of men dit wil verbieden en zo ja, hoe. In ieder geval hebben we AI-onderzoek nodig, uitgevoerd binnen de publieke sector, om systemen te ontwikkelen die ongewenste AI-gegenereerde inhoud kunnen detecteren en bestrijden. Dit is geen onmogelijke taak. Er moet ook worden geïnvesteerd in de ontwikkeling van betere benaderingen voor validatie, kwaliteitscontrole, de ontwikkeling van vangrails en verklaringsmogelijkheden toegepast op LLM's. Om dit te bereiken is het noodzakelijk dat de AI-onderzoeksgemeenschap toegang heeft tot open source LLM's die qua omvang en mogelijkheden vergelijkbaar zijn met de commerciële modellen. Er is een rol weggelegd voor Europa om dit te realiseren.

Vóór ChatGPT bekeek menigeen de doelen van AI met scepticisme of men beschouwde ze als verre beloften. Nu sommige van deze doelen reeds zijn bereikt, zij het op onverwachte wijze, moeten we het onderzoek niet opgeven uit angst voor misbruik. LLM's en GenAI vertellen niet het hele verhaal en mogelijk zijn ze niet zo krachtig als sommigen denken. Tegelijkertijd zijn ze interessant genoeg om onderzocht en geanalyseerd te worden en om ermee te experimenteren. Alleen dan wordt het mogelijk om waarborgen in te bouwen en misbruik te bestrijden.

Een gedragswetenschappelijk perspectief op AI

Jan De Houwer, Universiteit Gent

Hoewel gedragswetenschappen zich meestal bezighouden met het gedrag van individuele organismen zoals een mens of een ander dier, zou je vanuit een gedragswetenschappelijk perspectief ook naar andere systemen kunnen kijken. Je zou bijvoorbeeld kunnen stellen dat AI (bv. een computeralgoritme of een artificieel neuraal netwerk) een systeem is dat zich gedraagt: het verandert zijn toestand (bv. zijn output en/of de sterkte van de verbindingen in zijn netwerk) als gevolg van gebeurtenissen in zijn omgeving (bv. de input die het ontvangt van gebruikers; De Houwer & Hughes, 2023). Op basis van de vooronderstelling dat AI een gedragssysteem is, kan men methodes, concepten en inzichten uit de gedragswetenschap inzetten voor de studie van AI (Rahwan et al., 2019). Dit stelt ons in staat om op een systematische manier de overeenkomsten en verschillen tussen het gedrag van AI en andere systemen te onderzoeken, zelfs als de mechanismen die ten grondslag liggen aan het gedrag van de verschillende systemen fundamenteel verschillend of nog onbekend zijn (zoals bijvoorbeeld het geval is bij deep learning-modellen).

Een gedragsperspectief op AI benadrukt de vele parallellen tussen AI en andere gedragssystemen zoals individuele organismen. Net als zulke organismen kunnen AI-systemen niet alleen reageren op gebeurtenissen in hun omgeving, maar op

basis van ervaring kunnen ze ook de manier veranderen waarop ze op een gebeurtenis reageren (d.w.z. AI-systemen kunnen leren; De Houwer & Hughes, 2023). Via hun gedrag kunnen AI-systemen ook de omgeving van andere systemen zodanig veranderen dat die andere systemen ook hun gedrag veranderen (bv. wanneer een AI-systeem tekst presenteert als antwoord op een vraag van een persoon, kan dit het gedrag van die persoon veranderen). Bovendien kunnen AI-systemen dit doen op een geïndividualiseerde (d.w.z. op basis van informatie over het andere systeem) en onmiddellijke manier (d.w.z. online reagerend op het gedrag van het andere systeem). Geleid door bekende gedragsprincipes (bv. bekrachtiging; zie Catania, 2013) kunnen AI-systemen aldus worden geprogrammeerd of getraind als hulpmiddel voor gedragsverandering. Deze capaciteiten staan in schril contrast met die van oudere technologieën zoals de drukpers. Dergelijke oudere technologieën bieden een manier om het gedrag van mensen te veranderen door hun omgeving te veranderen (bijvoorbeeld door mensen de mogelijkheid te geven om boeken te lezen), maar ze missen de capaciteit om te reageren op de omgeving, om de manier waarop ze reageren op hun omgeving te veranderen en dus om de omgeving en het gedrag van andere systemen dynamisch te beïnvloeden. Vanuit een gedragsperspectief is het logischer om AI niet te vergelijken met een technologie zoals de drukpers, maar met een gedragssysteem dat technologie gebruikt. Vóór de komst van AI werden technologieën gebruikt door systemen die bestonden uit één persoon of een groep personen. De drukpers werd en wordt bijvoorbeeld gebruikt door individuele of groepen romanschrijvers, filosofen, wetenschappers, marketeers enzovoort met als doel het gedrag van mensen op bepaalde manieren te beïnvloeden (bv. ideeën overnemen, producten kopen). Als gedragssysteem kan AI technologieën gebruiken zoals een uit mensen bestaand systeem dat zou doen (bv. tekst genereren, beslissen wie welke tekst te zien krijgt, mensen aanmoedigen om zich op bepaalde manieren te gedragen). AI blijft een technologie in de zin dat ze door mensen wordt ontworpen en onderhouden om bepaalde taken uit te voeren, maar zelfs het ontwerp en onderhoud van AI kan, tenminste in principe, door AI worden uitgevoerd. In het laatste geval zouden AI-systemen als autonome gedragssystemen beschouwd kunnen worden, net als individuele organismen.

Een gedragsperspectief op AI onthult echter niet alleen overeenkomsten tussen AI en individuele organismen. Een gedetailleerde analyse van het gedrag van huidige AI-systemen onthult cruciale verschillen met het gedrag van mensen. Een cruciaal verschil ligt in het vermogen om symbolisch gedrag te vertonen. Mensen hebben een uniek vermogen om op iets te reageren alsof het op een bepaalde manier gerelateerd is aan iets anders. Ze kunnen bijvoorbeeld doen alsof het woord "GLAS" verwijst naar een fysiek glas, ook al is de relatie tussen beide arbitrair en gedefinieerd door sociale conventies, of doen alsof een muntstuk van 10 eurocent meer is dan een muntstuk van 5 eurocent in termen van monetaire waarde, ook al is een muntstuk van 5 eurocent meer in termen van fysieke grootte. Veel wetenschappers hebben betoogd dat dit soort symbolisch gedrag de kern vormt van de

menselijke cognitie (zie McLoughlin et al., 2020, voor een overzicht). Sommigen hebben ook (in grote lijnen) de uitgebreide leergeschiedenis in kaart gebracht die mensen moeten doorlopen voordat ze dit soort gedrag kunnen vertonen (bv. Hayes et al., 2001). Er zijn goede redenen om te stellen dat de huidige AI-systemen geen symbolisch gedrag vertonen. ChatGPT kan bijvoorbeeld geen zinnig antwoord geven op de volgende vraag: "Stel dat geel meer is dan blauw en dat rood minder is dan blauw. Is rood meer dan geel?" Een verbaal vaardig mens kan uitleggen dat rood niet meer kan zijn dan geel omdat hij of zij kan doen alsof geel meer is dan blauw en rood minder dan blauw. ChatGPT heeft nooit de training gekregen die nodig is om symbolisch gedrag te vertonen. Deze technologie is gevoed met enorme hoeveelheden gegevens en getraind om antwoorden te construeren die redelijk klinken op basis van deze gegevens, zonder het symbolische stadium te bereiken. Het is natuurlijk mogelijk dat toekomstige AI-systemen symbolisch gedrag vertonen als ze getraind worden op een manier die vergelijkbaar is met de leergeschiedenis die symbolisch gedrag bij mensen produceert. In deze context kan een gedragsperspectief op AI niet alleen licht werpen op de huidige aard van AI, maar ook helpen om de toekomst ervan vorm te geven.

Referenties

Catania, A. C. (2013). *Learning*. Sloan.

De Houwer, J., & Hughes, S. (2023) Learning in individual organisms, genes, machines, and groups: A new way of defining and relating learning in different systems. *Perspectives on Psychological Science*, 18, 649-663. https://doi.org/10.1177/17456916221114886

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001) *Relational Frame Theory: A Post-Skinnerian account of human language and cognition*. New York, NY: Plenum Press.

McLoughlin, S., Tyndall, I., & Pereira, A. (2020) Convergence of multiple fields on a relational reasoning approach to cognition. *Intelligence*, 83, 101491. https://doi.org/10.1016/j.intell.2020.101491

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., . . . Wellman, M. (2019). Machine behaviour. *Nature*, 568, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Wie zal de beschermengel zijn in de voetsporen van Erasmus?

Marc De Mey, Universiteit Gent

Het boek *The Printing Press as an Agent of Change*, Eisenstein's (1979) behandeling van de revolutie van de boekdrukkunst in de vijftiende eeuw, is interessant omdat het onderscheid maakt tussen de invloed van de nieuwe technologie op cultuur, religie en wetenschap. De impact van de introductie van de drukpers in de samenleving kan niet worden gereduceerd tot een eenvoudig schaaleffect. De wetenschap wordt anders beïnvloed dan de religie en de literatuur.

Het rapport van Nowotny verwijst naar de wetenschap met de woorden van Elizabeth Eisenstein: "Het feit dat identieke afbeeldingen, kaarten en diagrammen tegelijkertijd bekeken konden worden door verspreide lezers vormde een soort communicatierevolutie op zich" (Eisenstein, 1980, p. 53). "Een communicatierevolutie op zich" lijkt in het citaat misschien terloops te worden vermeld, maar Eisenstein bedoelt een echte kwalitatieve verandering die het aangroeiend karakter van wetenschappelijke vooruitgang beïnvloedt. Gepubliceerde wetenschappelijke boeken met slechts één editie zijn zeker identiek en geven iedere deelnemer die het vakgebied betreedt een identieke weergave van de stand van zaken in dat vakgebied. Door plaatsing van alle deelnemers binnen een gelijkwaardig speelveld ("uniform grid"; Eisenstein, 1980, p. 517) kan de enkele potentiële vernieuwer uiteindelijk het bereikte niveau overstijgen en het veld een stap verder brengen. Het egaliserende effect van gedrukte edities – in tegenstelling tot de fluctuerende en instabiele producten van kopiisten – is "essentieel voor de *cumulatieve cognitieve vooruitgang* en incrementele verandering" (Eisenstein, 1980, p. 412, nadruk toegevoegd), waardoor echte wetenschappelijke groei wordt gekenmerkt.

Door er in het rapport, net na het Eisenstein-citaat, op te wijzen dat de boekdrukkunst "resulteerde in een ware kennisexplosie in de 16de eeuw" wordt vooral de schaal benadrukt zonder het mechanisme van inhoudelijke stabilisatie op disciplinair niveau te specificeren. In de volgende alinea: "Het is daarom verleidelijk om parallellen te trekken tussen de kennisexplosie van de 16de eeuw en de 'informatie-explosie' die ons sinds enige tijd in haar greep houdt." In de volgende zin wordt eveneens de nadruk gelegd op de schaal: "De recente publicatie van generatieve AI op basis van LLM's (large language models) *heeft alleen maar bijgedragen* aan de overweldigende hoeveelheid mogelijkheden die AI/ML heeft geopend" (nadruk toegevoegd). Het is misschien voorbarig om nu al te speculeren over de verschillende kwalitatieve effecten die AI zou kunnen hebben op de wetenschap, literatuur, kunst en cultuur in het algemeen. Als we echter specifieke maatregelen overwegen om de potentiële bijdragen te optimaliseren, is het belangrijk dat deze maatregelen worden aangepast aan de specifieke aard van de betrokken domeinen. Zien we dat AI verschillende effecten heeft die

vergelijkbaar zijn met de effecten die Eisenstein documenteert voor religie en wetenschap?

Neem eerst religie. Als leidmotief voor het vierde hoofdstuk beroept Eisenstein zich op een passage uit *Reformation and Society in Sixteenth-Century Europe* van A.G. Dickens, waarin wordt gewezen op de belangrijke rol van de drukpers: "Tussen 1517 en 1520 werden er van de dertig publicaties van Luther waarschijnlijk meer dan 300.000 exemplaren verkocht ... Al met al lijkt het moeilijk om de betekenis van de drukpers, zonder welke een revolutie van deze omvang nauwelijks had kunnen plaatsvinden, voor de verspreiding van religieuze ideeën te overdrijven. In tegenstelling tot de Wycliffitische en Waldenzische ketterijen was het Lutheranisme vanaf het begin het kind van het gedrukte boek en via dit medium was Luther in staat om een exacte, gestandaardiseerde en onuitwisbare indruk te maken op Europese geest. Voor het eerst in de geschiedenis van de mensheid beoordeelde een groot lezerspubliek de geldigheid van revolutionaire ideeën via een massamedium dat gebruik maakte van de volkstalen in combinatie met de kunst van de journalist en de cartoonist" (geciteerd uit Eisenstein, 1980, p. 303). In de volgende pagina's verwijst Eisenstein naar gegevens die aangeven dat Luther's vijfennegentig stellingen, vertaald in het Duits en gedrukt in grote aantallen, al tegen het einde van 1517 circuleerden in Neurenberg, Leipzig en Bazel. Dit is slechts drie maanden na Wittenberg! De drukkers stonden blijkbaar te popelen om elk kort stukje tekst waarvan ze dachten dat het commercieel aantrekkelijk was, te produceren en te verkopen. Eisenstein citeert Zwingly, die in 1519 de tactiek van Luther aanbeveelt: bied bij huis-aan-huisverkoop slechts één enkel artikel aan, zodat de potentiële koper geen keuzeprobleem heeft en alleen maar hoeft te beslissen "ja, ik koop" of "nee, dank u". Dit is niet anders dan de zakelijke ijver waarmee ChatGPT nu door de strot wordt geduwd van pc- en internetgebruikers over de hele wereld: "Stel me om het even welke vraag". Luther voelde het potentieel van de drukpers en in 1522 kwam hij prompt met een vertaling van het *Nieuwe Testament* in gewone taal en in een grote oplage, die snel uitverkocht was. In 2023 publiceerde Codi Byte een *ChatGPT Bible* met als ondertitel *Everything You Need to Know about AI and Its Applications to Improve Your Life, Boost Productivity, Earn Money, Advance Your Career, and Develop New Skills* ("Alles wat je moet weten over AI en de toepassingen ervan om je leven te verbeteren, je productiviteit te verhogen, geld te verdienen, je carrière te bevorderen en nieuwe vaardigheden te ontwikkelen"). Hij beschouwt ChatGPT als nuttig voor creativiteit, voor ondernemers, voor onderzoekers, voor leerkrachten, voor schrijvers, voor programmeurs, voor professionals, voor social media-managers, voor journalisten en voor taalkundigen. Zo gepresenteerd is ChatGPT een schaamteloze commerciële onderneming bedoeld om het dagelijks leven van internetgebruikers binnen te dringen via een breed scala aan activiteiten, van het vinden van een recept om eieren te koken tot het schrijven van fictie of het debuggen van computercode. In sommige browsers heeft deze ChatGPT zich aan de gebruiker gepresenteerd als "jouw co-piloot", die je kan helpen bij alles wat je

onderneemt, je echte digitale tweelingbroer. Is er nog een ander ChatGPT?

Eisenstein vergelijkt de historische figuren Luther en Erasmus. Beiden wilden disfunctionele situaties in de kerk verhelpen en beiden zagen het potentieel van drukwerk als een goed instrument. Maar waar Luther het gebruikte om zijn religieuze boodschap in algemeen begrijpelijke vorm opnieuw aan het grote publiek te presenteren, zag Erasmus hierin een kans om de kwaliteit te verbeteren door de wetenschappelijke studie van de teksten die in het nieuwe medium zouden worden vastgelegd, te verdiepen. Nu men de mogelijkheid kreeg om het Woord van God in een superieure materiële vorm (druk) te verspreiden, moest dit op de juiste manier gebeuren met een serieuze wetenschappelijke studie van de oorspronkelijk gebruikte talen. Daarom inspireerde en ondersteunde hij aan de Universiteit van Leuven de oprichting van het *Collegium Trilingue* (1517!) voor de studie van Latijn, Grieks en Hebreeuws. Op dit moment zijn er vergelijkbare zorgen over de voortijdige verspreiding van AI-producten zoals ChatGPT en zijn er vergelijkbare initiatieven, zoals *International Institutions for Advanced AI*, mede opgericht door onder andere een van de pioniers van deep learning, Joshua Bengio, voor de diepgaande studie van AI, waaronder bewustzijn (https://arxiv.org/pdf/2307.04699v1.pdf). De aanbevelingen van Helga Nowotny specificeren onze lokale voorwaarden om samen te werken in dergelijke globale acties.

AI als aanjager van verandering – gezien door de ogen van een wiskundige

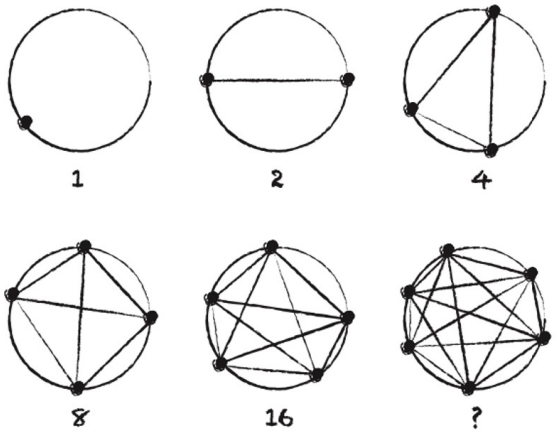
Ann Dooms, VUB

Om te begrijpen waar AI nu staat en hoe het zich in de toekomst zou kunnen ontwikkelen, moeten we terugkeren naar de krochten van de tijd.

Als we 5.000 jaar teruggaan in de tijd, komen we uit bij de Babyloniërs, die gedreven werden door een aangeboren nieuwsgierigheid om de wereld te begrijpen door middel van getallen. Via hun kleitabletten weten we dat ze enorme hoeveelheden gegevens verzamelden door nauwgezette en langdurige observaties. Hoewel ze niet beschikten over de wiskundige formalismen van latere tijdperken, legden ze de basis voor het herkennen van numerieke patronen in gegevens waaruit ze nauwkeurige voorspellingen konden doen, variërend van astronomische verschijnselen tot eigenschappen van wiskundige objecten. Ze begrepen al het concept van rechthoekige driehoeken en het verband tussen de lengtes van hun zijden, dat nu de stelling van Pythagoras wordt genoemd. Op het Plimpton 322-tablet verzamelden ze een lijst van zogenaamde Pythagorese drietallen, gehele getallen die oplossingen zijn van de gelijkheid $x^2+y^2=z^2$. Hebben we de beroemde stelling mogelijk verkeerd toegeschreven?

Om die vraag te beantwoorden, gaan we ons even verdiepen in patroonherkenning. Kies twee punten op de rand van een cirkel en verbind ze met een lijnstuk. Dit

verdeelt de cirkel in twee vlakken. Als je nu een derde punt kiest en dat verbindt met de eerste twee, krijg je vier vlakken. Elke keer dat je een nieuw punt kiest en dat verbindt met alle vorige, krijg je extra vlakken. In de tekeningen hieronder kun je zien dat het aantal afhangt van de positie van de punten op de rand. Het Cirkelprobleem van Moser uit 1949 vraagt hoeveel vlakken je maximaal kunt verkrijgen voor een gegeven aantal gekozen punten. Op het eerste gezicht zien we een patroon dat het aantal punten relateert aan een macht van 2 vlakken. Kunnen we dit gebruiken om het aantal vlakken voor 6 punten te voorspellen?



Number of points	Number of regions
1	$1 = 2^0$
2	$2 = 2^1$
3	$4 = 2^2$
4	$8 = 2^3$
5	$16 = 2^4$
6	$32 = 2^5$?

Hoewel het verleidelijk is om te denken dat het probleem kan worden opgelost door dit mooie numerieke verband, is het helaas verkeerd, want je kunt bewijzen dat er voor 6 punten maximaal 31 vlakken zijn. Dit voorbeeld toont duidelijk het gevaar van generalisatie. De loutere aanwezigheid van een ogenschijnlijk patroon betekent niet noodzakelijkerwijs dat het klopt. De Babyloniërs observeerden het verband tussen de lengtes van de zijden in een rechthoekige driehoek, maar Pythagoras en zijn leerlingen waren de eersten die een bewijs formuleerden voor de universele juistheid, waarmee ze de wiskunde als wetenschap in het leven riepen – de enige wetenschap waarin je onbetwiste waarheden kunt verkrijgen door logische bewijzen van beweringen te geven.

De Grieken hebben de toon gezet voor de vooruitgang van de wiskunde. Zij maakten er een krachtig instrumentarium van met instrumenten die voor altijd hun waarde behouden. In de 17de eeuw dook de kosmos op als het canvas voor nieuwe wiskunde gebaseerd op de observaties van de Babyloniërs. De wetten van Johannes Kepler over de beweging van de planeten en de telescopische waarnemingen van Galileo Galilei verbrijzelden het geocentrische wereldbeeld en onthulden een hemels ballet dat beheerst wordt door universele wetten, die prachtig zijn vastgelegd in geometrische formules. Toen verscheen Isaac Newton, wiens wetten van de beweging en de universele zwaartekracht een verklaring boden

voor de waargenomen bewegingen. Om ze te formuleren maakte hij een sprong in abstractie, samen met Gottfried Wilhelm Leibniz, door het wiskundige concept van de afgeleide te introduceren als hulpmiddel om beweging te bestuderen. Als zodanig werd de *afgeleide* letterlijk een *middel om verandering te meten*.

Dit was op zijn beurt het begin van een tak van de wiskunde die calculus wordt genoemd en die zich bezighoudt met continue verandering om te helpen bij het begrijpen en analyseren van variërende grootheden. Hij helpt in het bijzonder om functies te benaderen met prototypes die gemakkelijker te begrijpen en te berekenen zijn. Hierdoor konden bijvoorbeeld sinus- en cosinustabellen worden gemaakt tot de gewenste precisie, wat niet alleen praktische toepassingen zoals navigatie vergemakkelijkte. Om de eenvoudige maar saaie berekeningen uit te voeren legde Leonardo Da Vinci de basis voor een wiskundige machine: de mechanische rekenmachine, later gecommercialiseerd door Blaise Pascal en bediend door mensen, *computers* genoemd.

Naarmate de klok doortikt naar de 19de eeuw, ontmoeten we de astronoom Charles Babbage, die zich ergerde aan de menselijke maar gevaarlijke fouten in de tabellen met functiebenaderingen. Menselijke fouten bij het bedienen van de rekenmachines of het opschrijven van de resultaten konden leiden tot drastisch verkeerd berekende navigatieroutes. De industriële revolutie, aangedreven door stoom, leidde ertoe dat hij een rekenmachine uitvond, de *Difference Engine*, die zelfstandig berekeningen uitvoerde. Om geavanceerdere, met name gecombineerde berekeningen mogelijk te maken, waarbij de menselijke tussenkomst werd geëlimineerd, werkte hij samen met Ada Lovelace, de dochter van Lord Byron. Samen bedachten ze de eerste programmeerbare machine, geïnspireerd op het door stoom aangedreven Jacquard-weefgetouw dat fantastisch uitziende ingewikkelde patronen kon produceren door gebruik te maken van ponskaarten om de machine aan te sturen. Helaas zagen zij hun *Analytical Engine* nooit in actie.

Pas toen de fakkel een eeuw later werd doorgegeven aan de visionair Alan Turing, werd het digitale tijdperk echt geboren. In 1938 bewees hij wiskundig dat men een machine kon ontwikkelen die alles kan berekenen wat een mens met de hand kan doen – een theoretische constructie die de aanzet gaf tot de ontwikkeling van de elektronische computer zoals we die vandaag de dag kennen. Na zijn inspanningen op het gebied van cryptografie tijdens de Tweede Wereldoorlog gaf hij in 1950 verder vorm aan de geschiedenis van de computer met zijn baanbrekende paper *Computing Machinery and Intelligence*. Met het *Imitatiespel*, nu bekend als de *Turingtest*, creëerde hij een kader voor artificiële intelligentie, waarin we proberen machines te maken die kunnen leren problemen op te lossen zoals wij mensen dat doen – van trial-and-error tot leren op basis van voorbeelden.

De wiskundige principes van Turing gaven aanleiding tot de nu razend populaire neurale netwerken, geïnspireerd op het ingewikkelde netwerk van biologische

neuronen in onze hersenen. Een kunstmatig neuron zal informatie leveren (output) als het voldoende wordt gestimuleerd via zijn input. Frank Rosenblatt bewees in 1958 dat het kan leren wanneer het output moet "afvuren" door voorbeelden te zien waarin dat zou moeten. Om complexere problemen op te lossen werden gestaag verbeteringen aangebracht door kunstmatige neuronen te combineren in netwerken, met als hoogtepunt een wiskundig hoogstandje in 1975 dat bewijst dat dit model nog steeds kan leren van voorbeelden. De sleutel is de afgeleide van Newton om de beweging in de parameters van het netwerk tijdens de training te controleren. Het zogenaamde *backpropagation*-algoritme corrigeert tussentijdse onnauwkeurige voorspellingen door terug te reizen door het netwerk en cijfers op de juiste plaatsen aan te passen. Pas dankzij de enorme toename van digitale gegevens en rekenkracht kon men echter dergelijke neurale netwerken implementeren om ingewikkelde patronen in gegevens te onderscheiden, die veel verder gingen dan wat een mens in een mensenleven met de hand zou kunnen berekenen en die tot fascinerende generatieve hulpmiddelen hebben geleid, zoals DALL-E en ChatGPT.

Maar ... we zijn nog ver verwijderd van machines die kunnen leren en denken zoals mensen dat doen. Het is heel makkelijk om de huidige producten voor de gek houden en ze te ontmaskeren als probabilistische papegaaien die niet kunnen redeneren over de geleerde inhoud. Zullen ze ooit kunnen bewijzen dat voorspelde wiskundige patronen altijd opgaan? In deze zoektocht zullen we opnieuw getuige zijn van de fascinerende wisselwerking tussen wiskunde en technologie op onbekend terrein. AI zal de wiskunde zeker veranderen, en omgekeerd.

Moet er een recht op weigering zijn?

Katleen Gabriels, Universiteit Maastricht

Op 20 oktober 2023 meldden Belgische kranten dat netbeheerder Fluvius voor het eerst een weigeraar van een slimme elektriciteitsmeter voor de rechter zal dagen. Uiteraard is het delicaat om te vertrekken van een voorbeeld waarvan ik de onderliggende feiten niet ken, maar toch roept deze zaak dwingende vragen op over hoeveel handelingsvermogen en ruimte voor weigering gebruikers nog hebben in een samenleving die doordrongen is van digitalisering en AI. Als burgers een geldige reden hebben om een "slimme" technologie niet te gebruiken, bijvoorbeeld omdat het in strijd is met hun privacy, in hoeverre hebben ze dan het recht om te weigeren?

Gebruikers zijn natuurlijk niet machteloos of passief. In 2011 vroeg Maximilian Schrems alle gegevens op die Facebook over hem bewaarde. Facebook was op grond van de Europese privacywetgeving verplicht deze aan hem te verstrekken; iedere Europese burger heeft immers recht op toegang tot de gegevens die over

hen zijn verzameld. Facebook bleek meer dan 1.200 pagina's over Schrems te bewaren. Om de privacy van burgers te beschermen, mogen Amerikaanse bedrijven geen persoonlijke gegevens van Europeanen doorgeven voor commerciële doeleinden zonder hun toestemming. De positieve kant van dit verhaal is dat één persoon succesvol de praktijken van grote en machtige wereldwijde spelers kan blootleggen. Schrems werd vervolgens advocaat en privacy-activist.

Naast activisme zijn er nog andere manieren om technologie te 'weigeren', en dat hoeft allerminst te betekenen dat de technologie helemaal niet gebruikt wordt. Brunton en Nissenbaum (2015) moedigen bijvoorbeeld het verduisteren van gegevens aan om gebruikers meer handelingsvermogen te geven om hun privacy online te beschermen. Op die manier kunnen mensen de technologie nog steeds gebruiken, maar zijn hun gegevens beter beschermd, bijvoorbeeld door ze te versleutelen. Gebruikers passen technologieën ook vaak aan hun eigen standaarden aan. Kamphof (2015) observeerde hoe professionele zorgverleners actief proberen om slimme monitoringtechnologieën in te passen in hun dagelijkse zorgpraktijken en zelfs strategieën implementeren om de privacy van hun patiënten beter te respecteren. Ze gebruiken de technologie regelmatig doelbewust anders dan dat ontwikkelaars initieel voor ogen hadden; onderzoek toont aan dat ontwikkelaars eerder geneigd zijn om te focussen op veiligheid en autonomieverbetering en minder op fysieke privacy (Birchley et al., 2020).

Eerder in 2023 lanceerde de Italiaanse start-up Ca_pable hun Manifesto-collectie: de gebreide kledingstukken zijn opzettelijk ontworpen om gezichtsherkenningssoftware, zoals camera's op straat, te verwarren. In plaats van een persoon te herkennen, "ziet" de camera bijvoorbeeld een dier dat in het patroon zit verweven. Zo geeft het bedrijf de burger via kleding meer opties om zich anoniem op straat voort te bewegen. Cap_able streeft ernaar "een voorbeeld te zijn in de bewustmaking van het belang van burgerrechten: kleding als een middel om zichzelf, iemands identiteit en de waarden die gedeeld worden binnen een gemeenschap uit te drukken". Het bedrijf wil ook het bewustzijn van misbruik van gezichtsherkenningstechnologie vergroten.

Bedrijven zoals OpenAI respecteren het auteursrecht niet van alle teksten en afbeeldingen waarmee ze hun taalmodellen, waaronder ChatGPT, trainen. Nightshade is een tool voor het "vergiftigen" van gegevens om kunstenaars meer handelingsvermogen te geven: de tool voegt bewust ruis toe aan hun digitale kunst om te voorkomen dat grote bedrijven ze vervolgens gebruiken om generatieve AI-technologieën zoals Midjourney en Stable Diffusion mee te trainen.

'Weigering' van technologie is een veelzijdig concept dat verschillende niveaus van betrokkenheid omvat. Burgers kunnen op diverse manieren weerstand bieden: weigeren, afwijzen, gegevens verduisteren of "vergiftigen", aanpassen, valsspelen (bv. de gegevens van je activity tracker bewust in de war sturen door die aan

je huisdier te hangen), onderhandelen/lobbyen, protesteren, of gewoon door publiekelijk bezorgdheid te uiten. Het publieke debat en de publieke ruimte spelen een cruciale rol bij het vormgeven van AI-technologie. Als gebruikers hun zorgen over AI en slimme technologieën uiten, leidt dit tot belangrijke discussies over de ethische implicaties van de toepassing van AI. Deze gesprekken hebben al geleid tot veranderingen in de regelgeving (bijvoorbeeld in het geval van Schrems), meer transparantie en een betere verantwoording door de industrie. Gebruikers kunnen dus fungeren als katalysator voor maatschappelijke reflectie en verandering.

In een tijdperk waarin AI-technologie diep verweven is met ons dagelijks leven en de openbare infrastructuur, is de mogelijkheid om AI te weigeren een fundamentele uiting van het handelingsvermogen en de autonomie van de gebruiker. Om die reden moet het recht om te weigeren meer aandacht krijgen. Dit recht bestaat al in verschillende vormen, bijvoorbeeld in het recht om onveilig werk te weigeren of het recht om te staken, wat ook een vorm van weigering is. Toch zou het recht om bepaalde vormen van AI-technologieën in publieke infrastructuren te weigeren een rol kunnen spelen bij het waarborgen dat AI-technologieën de autonomie, privacy en ethische overwegingen van gebruikers respecteren, en uiteindelijk richting kunnen geven aan de toekomst van AI in de samenleving. Pertinente vragen daarbij zijn: Hoe kan zo'n recht eruitzien? En welke vorm moet het aannemen?

Referenties

Birchley, G., Huxtable, Murtagh, M., ter Meulen, R., Flach, P., & Gooberman-Hill, R. (2020). Smart Homes, Private Homes? An Empirical Study of Technology Researchers' Perceptions of Ethical Issues in Developing Smart-Home Health Technologies. *BMC Medical Ethics* 18(23).

Brunton, F. & Nissenbaum, H. (2015). Obfuscation: A User's Guide for Privacy and Protest. MIT Press.

Kamphof, I. (2015). A Modest Art: Securing Privacy in Technologically Mediated Homecare. *Foundations of Science* 22, pp. 411-419.

De Borg-gemeenschap

Yves Moreau, KU Leuven

Wij zijn de Borg. Laat uw schilden neer en geef uw schepen over. Wij zullen uw biologische en technologische kenmerken aan de onze toevoegen. Uw samenleving wordt aangepast om de onze te dienen. Verzet is zinloos.

De Borg. Star Trek: First Contact.

Om te onderzoeken hoe AI werkt als aanjager van verandering in de menselijke samenleving en in welke mate het de menselijke autonomie zal versterken of uithollen, moeten we de interacties met de samenleving bekijken. AI-systemen kunnen worden gezien als "cognitieve machines" die tekst, spraak en afbeeldingen op grote schaal op een mensachtige manier verwerken. Ze interageren met mensen en beïnvloeden hen dus onvermijdelijk.

Laten we beginnen met enkele fundamentele concepten om de basis te leggen voor onze bespreking. Een (mechanische) machine is een ontworpen systeem van onderdelen dat vermogen gebruikt om krachten, beweging en energie zodanig over te brengen dat een voorspelbare en gewenste output wordt geproduceerd op een manier die wordt bepaald door een specifieke input. Het eerste belangrijke aspect van een machine is de modulatie van krachten en beweging om een gewenst effect in de fysieke wereld te bereiken. Als we het biologische domein en "moleculaire machines" buiten beschouwing laten, is het tweede sleutelelement van een machine dat ze ontworpen en gebouwd is om een specifiek doel te bereiken. Het derde aspect is dat de meeste machines werktuigen zijn, wat betekent dat ze een productieve taak uitvoeren (Rube Goldberg-machines en raceauto's zijn daarentegen wel machines maar geen werktuigen). Machines versterken de menselijke capaciteiten, vooral bij repetitieve arbeid. Machines vormden het kloppende hart van de industriële revolutie.

Op dezelfde manier zijn computers "informatiemachines". Het belangrijkste verschil is dat ze in plaats van krachten en beweging over te brengen, informatie verwerken (in de vorm van elektromagnetische signalen). Ze doen dit door middel van algoritmes, wat opeenvolgende stappen zijn om een gewenste output te bereiken, op een manier die niet al te veel verschilt van wat mechanische machines doen. De vroege computers, van de Difference Engine en de Analytical Engine van Charles Babbage en Ada Lovelace tot de Z3 van Konrad Zuse, waren eigenlijk (elektro)mechanische machines. Hoewel de komst van de transistor het ontwerp en de relevantie van computers radicaal veranderde, is de analogie met mechanische machines zodanig dat digitale computers inmiddels ook machines worden genoemd. Computers staan centraal in onze informatierevolutie. In het essay "The Geek Reformation"[1], onderzocht ik de parallel tussen de drukpers als drijvende kracht achter de protestantse reformatie en het internet als drijvende kracht achter de informatierevolutie. Deze parallel is een nuttige manier om plausibele patronen te identificeren in de huidige stand van zaken – ook al moeten we de beperkingen van historische analogieën erkennen.

Met de term "cognitieve machine" willen we enkele belangrijke kenmerken van grootschalige AI-systemen benadrukken: ze verwerken grote hoeveelheden

[1] Yves Moreau (2016) The Geek Reformation, in *A Truly Golden Handbook: The Scholarly Quest for Utopia*. Leuven University Press, 464-477.

gegevens (Big Data), met name ongestructureerde en slecht gestructureerde gegevens, zoals natuurlijke taal, spraak en afbeeldingen, en communiceren op die manier met gebruikers, waardoor de perceptie van mensachtige capaciteiten ontstaat. Bovendien zijn ze zeer schaalbaar en hoeven voor het verwerken van meer gegevens alleen maar meer computers of meer cloudservers te worden toegevoegd, wat gunstige schaalvoordelen oplevert. Een belangrijk aspect is dat het aantal mensen dat nodig is om dergelijke systemen te beheren veel langzamer groeit dan de hoeveelheid verwerkte gegevens of het aantal klanten. Toch zijn cognitieve machines in zekere zin "gewoon" computers. Hoewel we zouden kunnen proberen het verschil tussen AI-systemen en "klassieke" computersystemen af te bakenen, concentreren we ons op de manier waarop grootschalige computersystemen de maatschappij beïnvloeden[2]. Of een bepaalde cognitieve machine al dan niet voldoet aan de drempel om als "artificiële intelligentie" te worden beschouwd, is niet de kern van de zaak.

Het is de buitengewone schaalbaarheid van cognitieve machines die de verbindende netwerken – de structuren die ontstaan uit het netwerken van cognitieve machines met hun gebruikers, leveranciers en ontwerpers – mogelijk maakt. Cognitieve machines hebben vaak interactie met en dus invloed op een groot aantal gebruikers. Omdat ze in een concurrerende economische omgeving bestaan, zijn twee van hun belangrijkste kenmerken dat ze zijn ontworpen om (1) netwerkeffecten te benutten om gebruikers en klanten te werven en te behouden of (2) te concurreren om de aandacht van gebruikers en het gedrag van gebruikers te beïnvloeden om zo het bereik en de inkomsten te maximaliseren. Sociale netwerken, zoals Twitter/X, Facebook of TikTok, zijn verbindende netwerken die beide aspecten benutten. Marktplatformen, zoals Amazon of Uber, maken meestal gebruik van netwerkeffecten tussen aanbieders en klanten. Google Search maakt gebruik van netwerkeffecten tussen adverteerders om de inkomsten te genereren die nodig zijn om een service te bieden die die van zijn rivalen overtreft. OpenAI wil de motor worden voor het genereren van AI-tekst in een groot aantal bedrijven. Een belangrijk aspect van verbindende netwerken is dat ze extreem compacte organisaties met een enorm bereik mogelijk maken. OpenAI bereikte binnen twee maanden na de publieke lancering van ChatGPT eind 2022 honderd miljoen actieve gebruikers met slechts zo'n driehonderd werknemers. Het is onlangs gewaardeerd op meer dan $ 80 miljard. In het verbindende netwerk vormen de cognitieve machine, het topmanagement en een paar ingenieurs het brein van een gigantische octopus waarvan de tentakels honderden miljoenen bereiken. Naarmate deze verbindende netwerken verder worden versterkt door steeds geraffineerdere AI die elke menselijke behoefte en wens ontcijfert, anticipeert en creëert, kun je verwachten dat ze zich verder zullen nestelen in elk hoekje en gaatje van de menselijke ervaring. Hoewel deze metafoor terecht immense macht

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

suggereert voor diegenen die het brein van de octopus controleren, bestaan er mechanismen voor feedback van gebruikers in de vorm van marktkrachten, die de opties van diegenen die beslissingen nemen ernstig beperken.

Als verbindende netwerken zich steeds dieper in de samenleving nestelen, wat voor soort samenleving kunnen we dan verwachten? Welk handelingsvermogen zal AI hebben en wat zal de impact ervan zijn op het menselijke handelingsvermogen? Als we handelingsvermogen definiëren als het vermogen om doelen te stellen, je omgeving aan te voelen en een redelijke handelswijze te plannen en uit te voeren om die doelen te bereiken, dan kun je stellen dat veel technische systemen – van de eenvoudige thermostaat tot de AI die een aandelenportefeuille beheert – een zekere mate van handelingsvermogen vertonen (als we even accepteren dat handelingsvermogen geen bewustzijn vereist). Op dit punt is het gepast om een andere metafoor naar voren te halen om de relatie tussen mensen en hun machines en hulpmiddelen te beschrijven: de cyborg, een "organisme dat zijn functie heeft hersteld of zijn vermogens heeft vergroot door de integratie van een of andere kunstmatige component of technologie die afhankelijk is van feedback"[3]. Het begrip feedback is hier belangrijk. Iemand met een houten been is geen cyborg, terwijl iemand met een geavanceerde myo-elektrische prothese dat technisch gezien wel is. Typisch voor het moderne archetype is een directe verbinding en feedback tussen de prothese en de hersenen, iets wat al wordt gevonden in cochleaire implantaten. Verder worden cognitieve functies bij moderne cyborgs vaak gedeeltelijk uitgevoerd of verbeterd door computersystemen. Het kernidee is dat in een cyborg het handelingsvermogen gesitueerd kan worden in het biologische of het digitale deel van het brein, of in beide. We denken hier niet aan een mogelijke cyborgtoekomst voor mensen, maar eerder aan wat er gebeurt als de besluitvorming in de hele maatschappij verdeeld raakt tussen het menselijke en het digitale.

Om te bekijken hoe verbindende netwerken de samenleving beïnvloeden, roepen we de meedogenloze antagonist uit Star Trek in herinnering: de Borg, een collectief van cyborgs verenigd via een "collectieve geest" en geleid door de Borgkoningin. Hoewel de koningin het collectief leidt en het enige lid is met persoonlijke autonomie, is ze ook het product van alle ervaringen en herinneringen van alle leden van het collectief.

Waar worden de beslissingen genomen in een maatschappij waar alle individuen voortdurend verbonden zijn met cognitieve machines en met elkaar via verbindende netwerken? Als AI een geweldig cadeau voor de verjaardag van je partner voorstelt door hun digitale spoor te beoordelen, heb je die beslissing dan echt alleen maar genomen omdat je een suggestie hebt goedgekeurd die veel beter is dan elk idee

[3] Joseph Carvalko (2012) *The Techno-human Shell-A Jump in the Evolutionary Gap*. Sunbury Press.

dat je zelf had kunnen bedenken? Wanneer een AI-systeem cv's screent voor een sollicitatiegesprek, wordt de besluitvorming dan niet al gedeeld tussen mens en machine? Als een autonoom AI-wapen een vijandelijke soldaat selecteert en doodt, lag de beslissing dan echt bij de operator die een breed doel aangaf ("neem dit militaire complex in") of bij de ontwikkelaar die het AI-besluitvormingskader programmeerde zonder enige kennis van de specifieke kenmerken van deze specifieke beslissing (en vaak zonder begrip van hoe specifieke beslissingen worden genomen)? We kunnen stellen dat dit proces van "verplaatsing van het handelingsvermogen" naar digitale systemen tot op zekere hoogte al aan de gang is in onze samenlevingen. Hoe vaak worden we in ons dagelijks leven niet geconfronteerd met het botte "De computer zegt nee", waar we ons een halve eeuw geleden waarschijnlijk doorheen hadden kunnen worstelen, maar waarbij we nu geconfronteerd worden met een ondoorgrondelijk computersysteem of een mens die erdoor gebonden is? Digitale systemen geven nu al vorm aan belangrijke aspecten van de architectuur van de moderne samenleving. Ze zijn simpelweg op zo'n saaie manier ingebouwd in het weefsel van de samenleving dat we er nauwelijks aandacht aan besteden, behalve wanneer we af en toe tegen een scherm vloeken. Hoewel sociale relaties, normen, markten, rechtssystemen en bureaucratieën samenlevingen hebben gevormd van de oudheid tot de laatmoderne geschiedenis, waren deze structuren uitsluitend ingebed in menselijk handelingsvermogen. Met de komst van het mainframe, de personal computer en daarna de cloud, zien we een verschuiving van het handelingsvermogen van het menselijke naar het digitale domein. Hoeveel mensen zijn hun job kwijtgeraakt door een geautomatiseerde formule die door een managementconsultant werd toegepast – of zelfs door een fout in een Excel-sheet? Hoeveel koppels zijn er gevormd of uit elkaar gegaan door algoritmische suggesties? Aandelenmarkten zijn ingestort door algoritmische handel zonder dat iemand echt begrijpt waarom[4]. Deze trends zullen exponentieel en op talloze manieren worden bevorderd door de opkomst van steeds krachtigere AI-machines. Ik poneer de stelling dat we op weg zijn naar een Borgmaatschappij, waar de verstrengeling van menselijk en digitaal handelingsvermogen steeds nauwer wordt en onmogelijk te ontwarren is. Net zoals het in het verleden bijna onmogelijk was om individueel handelingsvermogen te scheiden van de resulterende socio-historische krachten, zal het onmogelijk worden om te beslissen waar handelingsvermogen menselijk of digitaal is.

Hoewel doemscenario's niet de meest nuttige zijn en de huidige voorspellingen van een existentieel risico van Skynet/Terminator AI misschien afleiden van meer geloofwaardige zorgen, kunnen we ons redelijkerwijs zorgen maken over hoe ver deze verplaatsing van het handelingsvermogen zal gaan. De menselijke autonomie wordt al aanzienlijk beperkt door sociale structuren. Zelfs als we vrij zijn om beslissingen te nemen, zijn de beslissingen die we kunnen nemen en de beschikbare opties sterk beperkt. Hoeveel verder zou deze autonomie verschrompelen in de

[4] https://nl.wikipedia.org/wiki/Zwarte_maandag

Borgmaatschappij? Wie zal door de verbazingwekkende mogelijkheden van AI zijn individuele autonomie zien vergroten en wie zal een marionet worden van digitale systemen en alle overblijvende niet-geautomatiseerde tweederangstaken uitvoeren? Wie wordt aan de kant van de weg achtergelaten en volkomen overbodig gemaakt? Zijn de miljardairs van Silicon Valley onze Borgkoningin? Zullen we binnen een paar generaties volledig geassimileerd zijn in een maatschappij die totaal onherkenbaar is voor ons huidige zelf? Zijn mensen op lange termijn wel nog nodig voor het voortbestaan van de Borgmaatschappij? Kunnen we deze historische dynamiek verzachten of moeten we accepteren dat verzet zinloos is?

De vloek van Turing

Luc Steels, VUB

De Turingtest

In een beroemd artikel dat in 1950 werd gepubliceerd, stelde Alan Turing een test voor om te bepalen of een computerprogramma in staat was om te denken (Turing, 1950). Zijn voorstel was gebaseerd op een spel waarbij mensen proberen te raden of ze met een man of een vrouw te maken hebben, puur op basis van vragen en antwoorden die via stukjes papier worden uitgewisseld. Turing gaf hier een kleine draai aan: om te testen of een bepaalde machine, om precies te zijn een computerprogramma, intelligent was, stelde hij een spel voor waarbij je moest raden of je met een programma of een mens te maken had. Het programma zou als intelligent worden beschouwd als de beoordelaar niet in staat was om een onderscheid te maken tussen de twee, met andere woorden als het computerprogramma de beoordelaar misleidde zodat deze geloofde dat hij of zij te maken had met een mens terwijl de interactie in feite plaatsvond met een computerprogramma.

Turing stierf helaas te vroeg om grote technische bijdragen te kunnen leveren aan Artificiële Intelligentie, wat erg jammer is, want hij zou daar zeker toe in staat zijn geweest. Alleen de Turingtest blijft dus over als zijn belangrijkste nalatenschap met betrekking tot AI. In die tijd was er echter zeker geen eensgezindheid over de vraag of deze test een goede was. De meeste wetenschappers en ontwikkelaars die aan Artificiële Intelligentie werken, beschouwen de Turingtest als een zijspoor. Marvin Minsky, een van de grondleggers van AI en winnaar van de Turingprijs, noemde het bijvoorbeeld "een grap"[5]. Je zult tevergeefs zoeken naar papers over de Turingtest in het enorme aantal papers dat de afgelopen decennia al is gepubliceerd in AI-tijdschriften of -conferenties, behalve misschien om het nut ervan in twijfel te trekken. Toch gebruiken veel filosofen en psychologen de

[5] https://www.youtube.com/watch?v=3PdxQbOvAlI vanaf 23:40.

Turingtest als hun belangrijkste middel om te bespreken of AI vooruitgang heeft geboekt of niet. De komst van ChatGPT heeft dit idee alleen maar versterkt. Veel gebruikers raakten ervan overtuigd dat ChatGPT inderdaad de Turingtest heeft doorstaan omdat de teksten die het produceert, gebaseerd op menselijke prompts, vaak coherent en grammaticaal correct blijken te zijn. Het is moeilijk geworden om ze te onderscheiden van door mensen gemaakte teksten.

Problemen met de Turingtest

Bij nader onderzoek zien we twee problemen met de Turingtest:

1. De kern van de Turingtest is gebaseerd op misleiding: het is voldoende om te doen alsof je intelligent gedrag vertoont om voor de test te slagen. De interne mechanismen doen er niet toe. Voor het soort machine learning dat ten grondslag ligt aan de huidige generatieve AI, waaronder ChatGPT, zijn de processen waarmee het getrainde systeem tot conclusies komt trouwens volledig ondoorzichtig. Machine learning levert een zwarte doos af die niet kan worden geopend, zelfs als we dat zouden willen, omdat het gedrag van het systeem wordt bepaald door miljarden numerieke parameters die geen duidelijke, door mensen te begrijpen interpretatie hebben. We kunnen de intelligentie ervan dus alleen evalueren door gedrag van het systeem teweeg te brengen en van buitenaf te observeren.

Om te begrijpen dat misleiding een vreemde manier is om de vooruitgang in een wetenschappelijk vakgebied te bepalen, kun je dit vergelijken met de manier waarop andere wetenschappelijke disciplines zichzelf uitdagingen stellen en evalueren of ze vooruitgang boeken. Laten we de biologie als voorbeeld nemen. Vooruitgang in de biologie wordt zeker NIET beoordeeld aan de hand van de vraag of biologen erin slagen kunstmatige planten te maken of kunstmatige wezens die er op een bedrieglijke manier uitzien als echte. Biologie gaat over het begrijpen van de aard en de oorsprong van het leven. Het is waar dat een tak van de biologie, synthetische biologie genoemd, of soms "kunstmatig leven", dezelfde methodologie gebruikt die AI gebruikt om de geest te begrijpen, namelijk het bouwen van kunstmatige systemen die biologische functies vertonen, zoals het laten groeien van een celmembraan of nierdialyse (Langton, 1997). Het doel is echter niet om te misleiden, maar om te begrijpen hoe biologische functies materieel gerealiseerd kunnen worden, waarom ze belangrijk zijn voor het overleven van organismen en hoe deze functies in de evolutie kunnen zijn ontstaan.

Het doel van AI-onderzoek was vanaf het begin om bij te dragen aan grote wetenschappelijke vragen over de geest, zoals over de relatie tussen geest en materie, mechanistische verklaringen van leren, de aard van de vrije wil, d.w.z. de realisatie van autonomie en handelingsvermogen, enz. (Minsky, 1954). Vandaag de dag wordt dit doel grotendeels overschaduwd door de enorme commerciële druk op AI-ontwikkelaars om met informatietechnologieën te komen die de

enorme winsten kunnen genereren waar financiers die miljarden dollars in AI hebben geïnvesteerd om vragen. Het is echter dringend noodzakelijk dat de oorspronkelijke wetenschappelijke doelen weer op de agenda komen, al was het maar om te voorkomen dat er halfbakken toepassingen worden gebouwd en uitgebracht die mogelijk een zeer negatief effect op de samenleving hebben.

Een deel van het opnieuw tot leven brengen van een wetenschappelijke modus operandi in AI is het vinden van manieren om vooruitgang te evalueren in termen van vooruitgang op het gebied van de fundamentele vragen die AI probeert te helpen oplossen. Daarvoor is de Turingtest niet de juiste manier. We moeten dieper nadenken over wat voor soort fundamentele vragen er op het spel staan en ze vervolgens formuleren in termen van haalbare uitdagingen met verifieerbare uitkomsten, vergelijkbaar met de manier waarop David Hilbert in 1900 zijn beroemde uitdagingen aan wiskundigen formuleerde (Gray, 2000), van Hemmen en Sejnowski 23 nog onopgeloste problemen in de neurowetenschappen formuleerden (van Hemmen & Sejnowski, 2006), of Johan Hansson in 2015 de 10 grootste onopgeloste problemen in de natuurkunde formuleerde (Hansson, 2015).

Er heerst een wijdverspreide opvatting, helaas ook onder sommige verantwoordelijken voor het wetenschapsbeleid en de uitvoering ervan, dat AI geen wetenschap is en dat het een kwestie is van knutselen tot er op magische wijze iets opmerkelijks opduikt, zoals ChatGPT. Dit is een vergissing. ChatGPT is gebaseerd op een lange reeks wetenschappelijke inzichten en experimenten die teruggaan tot de jaren 1950. De beperkingen van generatieve AI op basis van statistisch neuraal leren waren al bekend en werden al besproken in de jaren 1960, net als mogelijke oplossingen die werden ontwikkeld in de jaren 1970 en 1980. Net zoals chemische technologie gebaseerd is op scheikunde of geneeskunde op biologie, heeft ook de technologie van AI een solide wetenschappelijke basis nodig.

2. Het tweede probleem met de Turingtest is dat wij mensen gemakkelijk kunnen worden misleid, omdat we onvermijdelijk een antropomorfische houding aannemen bij het beschrijven en verklaren van het gedrag van complexe systemen. Deze houding schrijft spontaan overtuigingen, verlangens en intenties toe aan entiteiten en gaat ervan uit dat ze kennis hebben en rationele beslissingen nemen (Dennett, 1987). De antropomorfische houding komt ons goed van pas in de omgang met andere mensen, maar we passen ze vaak ook metaforisch toe op fysieke entiteiten (vooral in sjamanistische en rituele contexten, Turner, 1974), levende entiteiten zoals huisdieren (McFarland, 2008) of machines, zoals wanneer we zeggen "mijn auto wil niet starten" of "de computer begrijpt mijn verzoek niet".

De overdreven toepassing van het intentionele perspectief betekent dat het geen betrouwbare basis is om te beoordelen of een bepaald AI-systeem intelligent is of het alleen maar lijkt te zijn. Pas als we meer te weten komen over de interne mechanismen van een systeem, kunnen we beoordelen of een systeem het echt

verdient om intelligent genoemd te worden. Wanneer we achteraf te weten komen hoe een opmerkelijk resultaat wordt bereikt, zijn we vaak teleurgesteld, vooral als er een of andere truc wordt gebruikt die niet overeenkomt met wat wij denken dat intelligentie vereist. Veel mensen zijn bijvoorbeeld verbaasd dat ChatGPT geen toegang heeft tot de betekenis van een tekst, hoewel dat wel zo lijkt. De output is volledig gebaseerd op het voorspellen van het volgende woord in een zin, waarbij rekening wordt gehouden met de context en enorme hoeveelheden bestaande teksten, inclusief prompts. Wanneer mensen zich hiervan bewust worden, beginnen ze te denken dat ChatGPT toch niet zo intelligent is.

Teleurstelling wanneer men zich realiseert hoe een AI-systeem werkt is ook de reden waarom de definitie van wat AI is of heeft bereikt is verschoven in de loop van de geschiedenis van het vakgebied. Schaken, differentiaalvergelijkingen oplossen, de juistheid van een ingewikkeld wiskundig bewijs controleren of het treinverkeer voor een heel land plannen werden allemaal beschouwd als vaardigheden die buiten het bereik van machines liggen en een mensachtige intelligentie vereisen, totdat er AI-programma's opdoken die deze vaardigheden konden bereiken – toen werden ze niet langer geacht intelligentie te vereisen. Dit is natuurlijk een overdreven reactie in de andere richting. Het is niet omdat we begrijpen dat leven gebaseerd is op biochemie dat levende organismen niet langer als levend worden beschouwd.

Vooruitgang in AI meten

AI-wetenschappers zijn zich terdege bewust van deze twee problemen met de Turingtest. Deels als reactie hierop heeft de machine learning-gemeenschap de Turingtest omgevormd tot een meer objectieve methodologie, die min of meer standaard is in de ingenieurswetenschappen. De recente literatuur staat vol met artikelen die eerst een concretere test voorstellen (bijvoorbeeld afbeeldingen correct labelen of een tekst in een andere taal vertalen), vervolgens kwantitatieve maatstaven definiëren om het prestatieniveau voor deze test objectief vast te stellen, en ten slotte experimenteel en systematisch verifiëren hoe een bepaalde AI-techniek presteert tijdens het uitvoeren van de test. Dezelfde test wordt ook aan menselijke proefpersonen gegeven, zodat kan worden vergeleken in hoeverre het AI-systeem presteert ten opzichte van menselijke proefpersonen.

Deze kwantitatieve vergelijkende methodologie heeft een enorme impuls gegeven aan het onderzoek naar machine learning, waarbij tests en datasets worden gedeeld en scoreborden dagelijks aankondigen welk team tot nu toe het beste resultaat heeft[6]. Ze heeft geleid tot een race naar "beter dan menselijke" resultaten voor een breed scala aan intellectuele taken. Aanvankelijk werden specifieke algoritmes en trainingsdatasets gebruikt, elk afgestemd op een specifieke taak.

[6] https://www.kaggle.com/

Maar na een voldoende aantal successen is het doel op dit moment om systemen voor te stellen die veelzijdig zijn voor veel taken – zonder de algoritmes telkens aan te passen of nieuwe trainingsgegevens te introduceren. Nog ambitieuzer is het uiteindelijke doel om de menselijke intelligentie te overtreffen en te komen tot AGI (Artificial General Intelligence, kunstmatige algemene intelligentie), die elke taak kan uitvoeren die de menselijke intelligentie kan uitvoeren, maar dan beter (Kurzweil, 2005).

Maar ondanks de meer objectieve aard van deze methodologie, zijn er nog steeds veel fundamentele gebreken, die eigenlijk ook al overvloedig zijn besproken in de technische literatuur. De voorgestelde tests blijken om een aantal redenen problematisch: (i) We hebben te maken met de klassieke problemen die alle vormen van empirisch testen hebben: representativiteit van steekproeven, verborgen vooroordelen, uitschieters, onvoorziene contextuele effecten; (ii) Intelligentie gaat over het omgaan met een open wereld waarin voortdurend snelle veranderingen plaatsvinden. Testsets raken onvermijdelijk snel verouderd en tests met vaste trainings- en testsets gaan niet na in hoeverre een systeem met verandering om kan gaan; (iii) Veel aspecten, vooral deze die met betekenis te maken hebben, kunnen niet geoperationaliseerd worden voor automatische toepassing op grote aantallen uitkomsten; daarom worden benaderende maatstaven gebruikt. Een goed voorbeeld is machinevertaling. Bestaande maatstaven zoals BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), LEPOR (Length-Penalty, Precision, n-gram Position Difference Penalty and Recall) enz. vergelijken allemaal oppervlaktekenmerken, namelijk de gelijkenis van woorden en woordreeksen (n-grammen), maar niet of de betekenis van een bronzin wordt weergegeven door de vertaling. De reden is simpel: het is veel moeilijker om betekenissen te operationaliseren en een automatisch criterium toe te passen op een grote testset.

Het gevolg van deze moeilijkheden is dat AI-ontwikkelaars victorie kraaien over beter-dan-menselijke prestaties voor de testsets en de prestatiemetingen die ze hebben, maar dat hun AI-systemen desondanks falen onder echte omstandigheden, vooral in open omgevingen waar regelmatig ongewone gebeurtenissen plaatsvinden. De fouten die beeldherkenningssystemen maken bij het herkennen van verkeersborden vormen een goed voorbeeld (Pavlitska et al., 2023). Het is voldoende dat er wat vuil op een verkeersbord zit, een sticker, of dat de lichtomstandigheden anders zijn dan bij de trainingsset, en een bord kan dramatisch verkeerd worden gecategoriseerd, met mogelijk gevaarlijke gevolgen. Dit probleem is een van de redenen waarom zelfrijdende auto's niet veilig worden geacht. Het mislukken van IBM's Watson Health-systeem is een ander illustratief voorbeeld. Diagnostici worden voortdurend geconfronteerd met ongewone gevallen die geen standaardpatroon volgen en die diepgaandere modellen en denkwijzen vereisen. De ongewone gevallen zullen nauwelijks zichtbaar zijn in de testgegevens. Dus ook in dit domein blijkt overdreven optimisme op basis

van hoge succespercentages bij testgegevens niet op te gaan in de echte wereld (Strickland, 2019).

De vloek van Turing

Wat is dan de vloek van Turing? Het is het risico om geobsedeerd te raken door de gepubliceerde maatstaven en deel te nemen aan een race om daarvoor te optimaliseren, waarbij de werkelijke omstandigheden en vooral de veranderingen die onvermijdelijk zijn in open omgevingen uit het oog worden verloren. Deze obsessie houdt het risico in dat een gemeenschap van onderzoekers die aan een specifiek onderwerp werken naar doodlopende wegen worden geleid. Ze belet dat er middelen worden besteed aan andere benaderingen die het slecht doen volgens de maatstaven, zeker in de beginfase, maar die op lange termijn juist kunnen leiden tot diepgaandere onderzoeksresultaten.

De vloek van Turing is ook het risico om de vergelijkende methodologie te serieus te nemen. Nadat een AI-systeem een positief en mogelijk hoger resultaat heeft behaald voor bepaalde maatstaven, wordt ons gevraagd te accepteren dat het systeem klaar is voor algemeen gebruik en een waardige, adequate vervanging voor een competente mens. We worden gevraagd om dat "gecertificeerde" systeem verantwoordelijkheid en handelingsvermogen te geven, zoals we bij een mens zouden doen, en om het te vertrouwen. We krijgen te horen dat de menselijke expert – de radioloog, architect, leraar, programmeur, vrachtwagenchauffeur, onderzoeker, journalist, kunstenaar – niet langer nodig is en dat het zinloos is om nieuwe menselijke experts op te leiden of in dienst te nemen.

Deze mening is natuurlijk schandaleus en wordt niet gedeeld door iedereen die in AI werkt (zeker niet door mij). Maar ze is wel gangbaar onder leidinggevende leden van de machine leer-gemeenschap. Een beroemde illustratie is de bewering in 2016 van Geoff Hinton (een winnaar van de Turingprijs) dat "mensen nu moeten stoppen met het opleiden van radiologen. Het is gewoon overduidelijk dat deep learning het binnen vijf jaar beter gaat doen dan radiologen"[7]. We zijn nu vele jaren verder dan de datum van Hinton's voorspelling en – gelukkig – zijn er geen aanwijzingen dat radiologen ontslagen worden. Sterker nog: er zijn er te weinig.

Een andere illustratie zijn soortgelijke beweringen dat programmeurs in de zeer nabije toekomst overbodig zullen worden omdat grote taalmodellen het lijken te kunnen, zoals wordt geïllustreerd door ChatGPT (getraind op enorme datasets van codevoorbeelden) of GitHub Copilot (Microsoft). Op het eerste gezicht zijn de resultaten erg indrukwekkend. Deze tools bieden stukjes code en gezaghebbend klinkende uitleg, gebaseerd op een uitgebreide trainingsset van door mensen gemaakte code en tutorials en tekstboeken. Maar helaas zijn er ook flagrante

[7] https://www.youtube.com/watch?v=2HMPRXstSvQ

fouten met hoge veiligheidsrisico's, zodat programmeurs wordt verteld deze tools nooit te gebruiken tenzij je zelf weet hoe je de betrokken code moet schrijven (Vaidya & Asif, 2023). De fouten zijn deels te wijten aan het feit dat foutieve code deel uitmaakte van de trainingsset en aan het feit dat generatieve AI zal afwijken van het meest waarschijnlijke patroon om een duidelijke schending van auteursrechten te voorkomen. De op AI gebaseerde coderingstools kunnen leiden tot een productiviteitsboost voor alledaagse taken, maar je moet een zeer vaardig programmeur zijn om ze goed te kunnen gebruiken. Dit is waarschijnlijk een patroon dat we in de meeste expertisedomeinen zullen zien.

Conclusie

Het doel van dit korte essay was om te beargumenteren dat de Turingtest en de vergelijkende methodologie die ermee gepaard gaat met een grote korrel zout genomen moet worden. Bijgevolg is het gebrek aan respect voor menselijke expertise dat wordt getoond door sommige AI-predikers niet gerechtvaardigd, evenmin als de druk om AI agressief te verspreiden naar alle hoeken van de samenleving op dit punt in de ontwikkeling van de technologie. AI is zeker een aanjager van verandering, maar men is zich onvoldoende bewust van de beperkingen ervan.

Referenties

Gray, Jeremy (2000) *The Hilbert Challenge*. Oxford University Press, Oxford.

Hannson, Johan (2015) The 10 biggest unsolved problems in physics. http://www.diva-portal.org/smash/get/diva2:996740/FULLTEXT01.pdf

Turing, Alan (1950) Computing Machinery and Intelligence. *Mind*, LIX (236): 433-460, doi:10.1093/mind/LIX.236.433

Kurzweil, Ray (2005) *The singularity is near*. Viking, New York.

Langton, Ray (1997) *Artificial Life. An Overview*. The MIT Press, Cambridge Ma.

McFarland, David (2008) *Guilty Robots, Happy Dogs: The Question of Alien Minds*. Oxford University Press, Oxford.

Minsky, Marvin (1954) Matter, mind and models. Updated regularly and reprinted in Minsky, M (ed.) (1968) *Semantic Information Processing*. The MIT Press, Cambridge Ma. https://web.media.mit.edu/~minsky/papers/MatterMindModels.html

Strickland, Eliza (2019) How IBM Watson Overpromised and Underdelivered on AI Health Care. *IEEE Spectrum*. https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

Pavlitska, Svetlana, Nico Lambing and Marius Zoelner (2023) Adversarial Attacks on Traffic Sign Recognition: A Survey. arXiv:2307.08278 [cs.CV]

Turner, Victor (1974) *Dramas, Fields, and Metaphors: Symbolic Action in Human Society*. Cornell University Press, Cornell, NY.

Vaidya, Jaideep and Hafiz Asif (2023) A Critical Look at AI-Generated Software. *IEEE Spectrum*. https://spectrum.ieee.org/ai-software

van Hemmen, Leo and Terry Sejnowski (2006) *23 Problems in Systems Neuroscience*. Oxford University Press, Oxford.

AI als motor voor verandering op onze kijk op handelingsvermogen

Johan Wagemans, KU Leuven

In haar provocerende essay "AI as an Agent of Change", geschreven voor de KVAB-Denkerscyclus van 2023, heeft Helga Nowotny gereflecteerd over een aantal interessante overeenkomsten en verschillen tussen de drukpers en AI als aanjagers van verandering. Met een bewonderenswaardige geleerdheid en welsprekendheid analyseert ze een breed scala aan historische, sociale en culturele aspecten van deze twee revolutionaire technologieën. Ze presenteert een indrukwekkende synthese van haar bevindingen, trekt er interessante conclusies uit en plaatst ze in een adembenemend breed perspectief, met belangrijke lessen voor overheden, beleidsmakers, bedrijven, organisaties en individuele stakeholders, waaronder wetenschappers en, in zekere zin, alle burgers die geconfronteerd worden met de nieuwe ontwikkelingen die AI hen brengt.

In dit korte commentaar zal ik me concentreren op het begrip "handelingsvermogen", dat centraal staat in de analyse. In hoeverre kunnen we echt handelingsvermogen toekennen aan AI? Helga Nowotny koppelt het handelingsvermogen van machines aan hun functies en bedoelingen: "(...)machines zijn gebouwd om bepaalde functies te vervullen. Er zijn menselijke bedoelingen in vastgelegd." Vervolgens wijst ze erop dat mensen de neiging hebben om handelingsvermogen aan machines toe te kennen, gebaseerd op "een diepgewortelde antropomorfische neiging om het gedrag van een andere entiteit of object te zien in termen van mentale eigenschappen". Dit blijft "… relatief onschuldig als het gaat om vertrouwde technologieën. (…) Als het echter om AI gaat, kan het ... veranderen in een gevaarlijk dwingende illusie dat we in de aanwezigheid zijn van een denkend wezen zoals wijzelf." In haar boek *In AI We Trust* heeft Helga hieromtrent een interessante paradox geïdentificeerd: "We gebruiken AI om onze controle over de toekomst en onzekerheid te vergroten, maar tegelijkertijd vermindert de performativiteit van AI, de macht die het heeft om ons te laten handelen op manieren die het voorspelt, ons handelingsvermogen met betrekking tot de toekomst. Dit gebeurt wanneer we vergeten dat wij mensen de digitale technologieën hebben gecreëerd waaraan we handelingsvermogen toekennen. Als hier niets aan wordt gedaan, zou het zelfs kunnen leiden tot de

terugkeer van een deterministisch wereldbeeld waarin de meeste mensen geloven dat AI hen beter kent dan zijzelf, inclusief hun toekomst."

In de rest van dit commentaar zal ik proberen duidelijk te maken dat kennis van de factoren die menselijk gedrag sturen of die ten grondslag liggen aan onze voorkeuren altijd al is uitgebuit en dat er in feite een continuüm is van het verliezen van controle over ons eigen gedrag. AI is in dit opzicht niet zo nieuw, het doet het alleen beter en op minder transparante manieren. Als we langs de schappen van een supermarkt lopen, op zoek naar een artikel dat we nodig hebben, zijn we ons er misschien niet van bewust dat de duurste producten op ooghoogte staan, terwijl de goedkopere merken hoger of lager staan en dus moeilijker te bereiken zijn. Dit economische gedrag van het supermarktpersoneel, dat meer winst wil maken, maakt perfect gebruik van wetenschappelijke inzichten in aspecten die het gedrag van de gemiddelde consument bepalen (bv. beperkte aandachtsspanne, normale perceptuele processen, automatische besluitvorming met weinig cognitieve controle). Zoekmachines maken gebruik van ons zoekgedrag in het verleden om hits voor ons te sorteren, en bijkomende economische krachten (voornamelijk bedrijven die betalen om hun website hoger te krijgen) zullen ook een belangrijke rol spelen, misschien buiten het bewustzijn en de controle van de gebruiker zelf. Aanbevelingssystemen op Netflix, Spotify, Instagram en dergelijke gaan nog een stap verder, hoewel veel consumenten dit misschien wel leuk vinden.

Wat betreft het specifieke geval van de esthetiek van afbeeldingen, maken alle camera's in onze smartphones tegenwoordig gebruik van ingebouwde software om betere foto's te maken, en kunnen generatieve antagonistennetwerken worden gebruikt om de esthetische kwaliteiten van foto's te verbeteren in een soort post-fotografische "zwarte doos"-bewerkingsfase, gebaseerd op machinaal lerende netwerken die zijn getraind met tienduizenden afbeeldingen. Wetenschappers die de visuele perceptie bestuderen, kunnen de parameters die worden aangepast onderzoeken en proberen de factoren te begrijpen die de esthetische waardering bepalen (zoals verbeterd contrast, kleur, scherpte, diepte enz.), maar inzicht in deze factoren betekent niet dat we controle krijgen over wat we met de beelden doen.

In een meer recente ontwikkeling, computationele esthetica genoemd, worden modellen voor machine learning getraind om menselijke esthetische voorkeuren voor afbeeldingen te voorspellen. Tot nu toe is dit nog niet erg succesvol (hoewel de grote techbedrijven hier veel geld en onderzoekstijd in investeren), deels omdat de kwaliteit van de trainingsgegevens beperkt is, en omdat esthetische voorkeur verre van universeel is. In plaats daarvan spelen cultureel of sociaal bepaalde invloeden en zeer persoonlijke factoren ook een grote rol. Moeten we bang zijn voor zulke ontwikkelingen? Zolang we ons bewust zijn van de doelen van de systemen die we gebruiken, en zolang we op zijn minst iets begrijpen van de essentie van machine learning op basis van eerder verkregen gegevens van

andere menselijke gebruikers, zijn we nog steeds in staat om controle te houden over ons eigen gedrag. We kunnen deze systemen gebruiken zonder er "controle aan over te dragen". Ik ben zelf betrokken bij een groot project waarin we de menselijke esthetische voorkeur voor beelden willen voorspellen, verklaren en begrijpen. We vertrekken van bestaande machine learning-modellen, maar we zullen ze verrijken met onze kennis over de menselijke perceptie, inclusief de centrale rol die perceptuele organisatie speelt in de interactie met het geheugen, emoties, expertise, persoonlijkheid enzovoort. Ons doel is niet om te concurreren met de grote techbedrijven, maar om belangrijke wetenschappelijke vooruitgang te boeken. Met deze kennis kunnen we vervolgens het publiek informeren over alle factoren die een rol spelen en hoe ze hun esthetische ervaringen kunnen verrijken door gebruik te maken van deze kennis.

Ik denk dat dit op één lijn ligt met Helga Nowotny's pleidooi "om af te stappen van het simplistische binaire utopisch-dystopische denkschema" en haar nadruk op "de belangrijke rol die de wetenschap moet spelen om aan het publiek uit te leggen hoe AI werkt". Voor mij is AI nog steeds een machine, geen echte entiteit, omdat de ogenschijnlijke intentionaliteit en handelingsvermogen altijd worden geïnitieerd door mensen: zij die de neurale netwerken opzettelijk ontwikkelen en trainen om een specifieke functie uit te voeren, meestal nauw gedefinieerd en bijna altijd gebaseerd op trainingsgegevens die onbedoeld worden aangeleverd door duizenden, zo niet miljoenen mensen.

4. Reacties van beleidsmakers

Hoe doet Vlaanderen het op het gebied van AI?

Bart De Moor, KU Leuven

Kunstmatige Intelligentie (AI) – wellicht beter te omschrijven als *assisterende intelligentie* – werd onlangs in een indrukwekkend en bijzonder uitvoerig rapport[8] van de Nederlandse WRR (*Wetenschappelijke Raad voor het Regeringsbeleid*), beschreven als de "Nieuwe Systeem Technologie".

Inderdaad, de opeenvolgende industriële revoluties van de afgelopen driehonderd jaar zijn allemaal gekenmerkt door de introductie van nieuwe "systeemtechnologieën", waarvan de impact de wereldwijde samenleving drastisch heeft gewijzigd. Denk aan energieproductie en -verbruik (steenkool en stoom, fossiele brandstoffen zoals aardolie, elektriciteit, kerncentrales enz.), de mechanisering, robotisering en automatisering van de maakindustrie en industriële processen, de revoluties in mobiliteit (treinen, auto's, vliegtuigen enz.), de globale informatisering (computers) en de revoluties in communicatietechnologieën (audiovisuele media, world wide web, internet enz.).

Allemaal zijn het voorbeelden van *systeemtechnologieën* die, eenmaal ze de wereld hadden veroverd, gebleven zijn. Systeemtechnologieën bouwen op doorbraken in vorige generaties: het world wide web en het internet bouwen op onze wereldwijde communicatiesystemen, op computers (wet van Moore), op (software)automatisering en zonder energievoorziening zouden ze gewoonweg onmogelijk zijn.

AI, als nieuwe systeemtechnologie en extra laag van deze evolutie, steunt op de informatiewetenschappen (inclusief wiskunde, software engineering enz.) en -technologieën (computers en servers), communicatie (wereldwijde connectiviteit) en het delen van gegevens (bv. draadloze interacties, websites en databases enz.). AI stelt ook enorme eisen aan energie en vermogen, eisen die we maar al te vaak als vanzelfsprekend beschouwen. De tsunami van sensoren en data, ook wel *het nieuwe goud* genoemd, creëert een toenemend aantal datagestuurde diensten en toepassingen in wetenschappelijk onderzoek, gezondheid en geneeskunde, industrie, mobiliteit, overheidsdiensten én ons dagelijks leven.

Nieuwe systeemtechnologieën gaan vaak hand in hand met of zijn gebaseerd op nieuwe wetenschappelijke inzichten en technologische doorbraken. Ze resulteren op hun beurt in nieuwe toepassingen, nieuwe bedrijfsmodellen, nieuwe voordelen

[8] https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

en onvermijdelijk (soms onvoorziene) overlast en risico's voor burgers en samenleving. Omwille van de immense impact op alle dimensies van ons dagelijks leven, hebben overheden op mondiaal, nationaal en regionaal niveau geen andere keuze dan zich te engageren voor de ontwikkeling van nieuwe regelgeving, de goedkeuring van wereldwijde protocollen, de inzet van infrastructuur, het beperken van risico's en de bescherming van alle stakeholders. Met als opzet dat instellingen, bedrijven en burgers optimaal kunnen beschikken over de nieuwe technologieën. En natuurlijk zijn deze nieuwe technologieën vaak een kans om nieuwe belastingen te heffen. AI als nieuwe systeemtechnologie is zeker geen uitzondering op al deze eigenschappen.

Systeemtechnologieën veroorzaken gewoonlijk niet-technische tekortkomingen in de manier waarop onze samenleving functioneert. Er zijn *democratische* tekortkomingen, omdat vaak niet alleen burgers, maar ook besluitvormers (bv. in parlementen) geen vat krijgen op nieuwe technologieën en niet begrijpen hoe de voordelen en bedreigingen de samenleving kunnen beïnvloeden. Regelmatig leidt dit tot overregulering, wat innovatie kan afremmen. Er zijn *juridische* tekortkomingen, omdat parlementen en regeringen vaak te traag zijn in het formuleren en goedkeuren van adequate wetten. Soms doen ze dit pas wanneer er ongelukkige voorvallen zijn of wanneer ze verrast en overweldigd zijn door wereldwijde ontwikkelingen. Tenslotte zijn er *ethische tekortkomingen*. Die ontstaan niet in het "*hoe*" van nieuwe wetenschap en technologie, maar in het "*wat*". Niet in *hoe* nieuwe AI-toepassingen geïmplementeerd worden, maar eerder in *wat* de voorziene of onvoorziene impact en gevolgen zouden kunnen zijn. Ethiek behandelt noodzakelijke keuzes uit een heel spectrum van opties die door de vooruitgang in wetenschap en technologie mogelijk zijn.

Vlaanderen, als een van de drie gewesten in België, is geen uitzondering op het vlak van de omgang met de wereldwijde impact van de nieuwe AI-systeemtechnologie. Vijf jaar geleden besliste de Vlaamse Regering om een ambitieus AI-Programma Vlaanderen[9] te lanceren, met een waarde van 32 mio €/jaar, dat drie pijlers omvat: een *onderzoeksprogramma*[10] van 12 mio €/jaar, *een extra budget* (15 mio €/jaar) *voor R&D-beurzen*[11] *aan bedrijven* die innovatieve AI-toepassingen ontwikkelen, en 5 mio €/jaar voor "ondersteunende maatregelen": een *Kenniscentrum Data & Maatschappij*[12], de *Vlaamse AI Academie*[13] en verschillende communicatie-initiatieven, waaronder het *AI citizen science initiatief AMAI*[14].

---

9 https://www.ewi-vlaanderen.be/en/flemish-ai-plan/broad-context
10 https://www.flandersairesearch.be/nl
11 https://www.vlaio.be/nl/begeleiding-advies/digitalisering/artificiele-intelligentie
12 https://data-en-maatschappij.ai/
13 https://www.vaia.be/nl/?lang=nl
14 https://amai.vlaanderen

Het Vlaams AI-Onderzoeksprogramma (Flanders AI Research Program, afgekort als FAIR) omvat de vijf universiteiten in Vlaanderen (Leuven, Gent, Antwerpen, Brussel en Hasselt) en de vier strategische onderzoekscentra (imec (nanotechnologie), VIB (biotechnologie), VITO (milieu, energie en duurzaamheid) en Flanders Make (productie). Er zijn tientallen professoren en hoofdonderzoekers bij betrokken, honderden promovendi en postdocs, in meer dan 40 onderzoeksgroepen.

In de eerste editie (FAIR 1.0, 2019-2023) werden de onderzoeksactiviteiten georganiseerd in onderzoeksuitdagingen ("*Grand Challenges*") enerzijds en toepassingen ("*Use Cases*") anderzijds. In de tweede periode van 5 jaar (FAIR 2.0, 2024-2028) zijn twee onderzoeksuitdagingen geformuleerd. De eerste onderzoeksuitdaging (de grootste van de twee) focust op *AI Driven Data Science* en probeert *complexe besluitvorming en de creatie van bruikbare inzichten uit de exploitatie van data te ondersteunen.* De tweede onderzoeksuitdaging, *Situated AI*, is gericht op *het ondersteunen van complexe taken in een dynamische omgeving met (semi-)autonome AI-systemen die in realtime met elkaar en met mensen samenwerken.* Beide onderzoeksuitdagingen delen gemeenschappelijke translationele waarden en doelstellingen met duidelijk gedefinieerde criteria. Deze criteria zijn mensgericht, verantwoordelijk, veerkrachtig, performant, data-efficiënt en duurzaam op het vlak van energie.

Tijdens het ontwerpen van FAIR 1.0 ontdekten we een schijnbare paradox bij de interactie met potentiële gebruikers van AI-technologie. Allemaal, zonder uitzondering, erkennen ze een dringende vraag naar meer AI in hun biotoop. Maar AI is een "containerbegrip" dat een grote verscheidenheid aan problemen en oplossingen dekt. Het is zeer belangrijk om de "verwachting van AI" te managen, omdat potentiële gebruikers vaak wonderoplossingen verwachten die met de huidige AI simpelweg niet mogelijk zijn en in veel gevallen zelfs in de toekomst nooit realiseerbaar zullen zijn. Wanneer potentiële gebruikers hun behoeften in specifieke termen willen omschrijven, ontbreekt het hen regelmatig aan typisch AI-jargon. Zo slagen ze er niet in om precies aan te geven wat het probleem is waarvoor ze een AI-oplossing zoeken. Met andere woorden, de paradox die we ontdekten impliceert dat er enerzijds *een brede vraag is naar meer AI*, maar dat er anderzijds een absolute, semantische drempel is bij het formuleren of "articuleren" van een exacte AI-vraag. Dit is de paradox van de *vraagarticulatie*.

Er zijn verschillende maatregelen genomen om die leemte op te vullen.

VAIA, de Vlaamse AI-Academie, organiseert elk jaar honderden activiteiten rond AI in het kader van levenslang leren. Het oplossen van de paradox van de vraagarticulatie staat er centraal.

VLAIO, het agentschap dat AI-R&D in bedrijven subsidieert, is gestart met de organisatie van heel wat informatietrajecten op maat, waarbij goede praktijken op de voorgrond staan.

In het onderzoeksprogramma zijn verder meer dan 30 toepassingen gelanceerd die een rolmodel zijn om toekomstige potentiële gebruikers te inspireren. De toepassingen zijn in 4 clusters gegroepeerd: Gezondheid, Planeet en Energie, Industrie en Maatschappij. In de cluster Gezondheid zijn de toepassingen *thuis monitoring, realtime, "real-world" biomedische monitoring, medische beeldvorming, "single-cell" moleculaire biologie, "digital twin"-cardiologie, AI in intensieve zorgen, gepersonaliseerde dermatologie, sportmonitoring.* In de cluster Planeet en Energie zijn er rolmodellen op het vlak van *monitoring van de natuurlijke omgeving, geo-stedelijke platformen en AI voor het slimme elektriciteitsnet in Vlaanderen.* In de cluster Industrie staan *slimme machines, monitoring en controle van productiemachines, productoptimalisatie, "straight-through" digitalisering van productie, prognostiek en gezondheidsbeheer voor activa en renovatie* centraal. In de cluster Maatschappij tenslotte lanceren we voorbeelden over *AI om publieke tewerkstellingsinitiatieven te optimaliseren, digitale menswetenschappen en AI voor onderwijs en opleiding.*

Al deze initiatieven zijn terug te vinden in de ambitieuze plannen voor FAIR 2.0, de tweede fase van het Vlaamse onderzoeksprogramma voor 2024-2028. Een minder punt is dat, tegen alle verwachtingen in, de huidige Vlaamse Regering beslist heeft om de budgetten voor het gehele AI-programma niet te verhogen. In 2024 blijft het budget status quo, ondanks een extreem positieve beoordeling van FAIR 1.0, het onderzoeksprogramma 2019-2024, door onze internationale wetenschappelijke adviesraad en door een onafhankelijk comité van internationale experten. Dit is natuurlijk betreurenswaardig, want in alle regio's en landen die aan Vlaanderen grenzen, zijn de R&D-budgetten voor AI aanzienlijk toegenomen en die toename zal op korte termijn standhouden, in navolging van de internationale trends in de VS en China. Daarom hopen alle stakeholders in Vlaanderen oprecht dat de toekomstige Vlaamse Regering vanaf 2024 de internationale groeitrends zal bijbenen en haar beleid ermee zal synchroniseren.

Samenvattende toespraak voor het evenement "AI as an Agent of Change"

Lucilla Sioli, Kunstmatige Intelligentie en Digitale Industrie, Europese Commissie

Als directeur van het Directoraat "Kunstmatige Intelligentie en Digitale Industrie" van DG CONNECT, in de Europese Commissie, omvat mijn portefeuille een breed spectrum aan verantwoordelijkheden. Deze omvatten het formuleren van AI-beleid ter ondersteuning van de ontwikkeling van AI in de EU. Dit heeft betrekking op het bevorderen van onderzoek, innovatie en de inzet van AI, het zorgen voor voldoende talent in de EU, maar ook op het regelgevingskader voor het gebruik van AI, de AI-verordening, en het stimuleren van de ontwikkeling en ingebruikname van betrouwbare AI. Daarnaast houd ik toezicht op het bestuur van AI, waaronder de bevordering van betrouwbare AI op internationaal niveau.

Bij het bespreken van het cruciale onderwerp van de rol van AI als aanjager van maatschappelijke verandering, wil ik uw aandacht vestigen op de inzichtelijke aanbevelingen die Helga Nowotny in haar verslag doet en die zijn ontworpen om een meer onderbouwde, humanistische en evenwichtige benadering van de ontwikkeling van AI en de integratie ervan in de samenleving te bevorderen. Hieronder wil ik benadrukken hoe deze aanbevelingen aansluiten bij de belangrijkste beleids- en regelgevingskaders van de Europese Commissie (EC), namelijk het gecoördineerde plan voor artificiële intelligentie en de AI-verordening.

In haar eerste aanbeveling benadrukt Nowotny de noodzaak van een grote publiekscampagne om burgers voor te lichten over AI, de diverse toepassingen en de inherente beperkingen ervan, vanuit het perspectief van digitaal humanisme. Dit komt redelijk goed overeen met het gecoördineerde plan van de EC, dat de nadruk legt op een mensgerichte aanpak en de noodzaak om digitale vaardigheden en AI-geletterdheid onder burgers te stimuleren. Het sluit ook aan bij de focus van de AI-verordening op transparantie en verantwoording, zodat burgers de AI-systemen begrijpen en vertrouwen. Als onderdeel van de geplande acties zal de Commissie via het actieplan voor digitaal onderwijs 2021-2027 stages in digitale domeinen ondersteunen, met de nadruk op AI-vaardigheden, en de ontwikkeling van ethische richtlijnen voor leerkrachten over AI en datagebruik bij het onderwijzen en leren. Daarnaast worden de lidstaten in het kader van het plan aangemoedigd om de vaardigheidsdimensie in hun nationale AI-strategieën te verfijnen en te implementeren, in samenwerking met de sociale partners. Dit omvat het bevorderen van computationeel denken, het creëren van AI-onderwijsprogramma's, het vergroten van de beschikbaarheid van AI-opleidingen en het ondersteunen van effectieve AI-onderwijstechnologieën.

Nowotny's tweede aanbeveling pleit ervoor om prioriteit te geven aan fundamenteel onderzoek naar AI in de stijl van de Europese Onderzoeksraad (ERC), waarbij een focus op humanistische aspecten wordt voorgesteld en de onbalans tussen publieke en private financiering van AI-onderzoek wordt aangepakt. Wat betreft de behoefte aan onderzoek en innovatie op het gebied van AI worden in het gecoördineerde plan van de EC verschillende belangrijke initiatieven beschreven. Deze omvatten de oprichting van Europese partnerschappen in Horizon Europe voor het stimuleren van innovatie op het gebied van AI, data en robotica, waarbij de nadruk ligt op een mensgerichte en betrouwbare Europese visie op AI. Verder wil het plan, in een top-downbenadering, de uitmuntendheid van de EU op het gebied van AI-onderzoek en -innovatie versterken door financiering te bieden voor de ontwikkeling van de volgende generatie AI-systemen, die groener, autonomer, transparanter en niet-bevooroordeeld zullen zijn. Het plan introduceert ook het AI Networks of Excellence-initiatief om een sterke Europese onderzoeksalliantie te bevorderen. Daarnaast bevordert het beleidskader het vertrouwen in AI-systemen, ethische AI-ontwikkeling en multidisciplinair onderzoek. Tot slot stimuleert het de versnelling van private en publieke investeringen door gebruik te maken van EU-

financiering die beschikbaar is via programma's als Digitaal Europa (DEP), Horizon Europe (HE) en de faciliteit voor herstel en veerkracht (RRF), bijvoorbeeld door faciliteiten op te zetten die testen en experimenteren vergemakkelijken.

Nowotny's laatste aanbeveling roept op tot robuuste steun voor onderzoek naar de impact van AI op de samenleving, vooral op gebieden die waarschijnlijk niet door de grote bedrijven zullen worden bestudeerd, om de nuttige toepassingen van AI te begrijpen en sociale schade te voorkomen. Ze benadrukt ook de noodzaak van interdisciplinaire benaderingen bij het begrijpen en toepassen van AI. Dit sluit aan bij het engagement in de AI-verordening om ervoor te zorgen dat AI de fundamentele rechten en waarden respecteert. De AI-verordening volgt namelijk een risicogebaseerde benadering waarbij de regels gericht zijn op AI-systemen die worden gebruikt in contexten waarin de fundamentele rechten en veiligheid van mensen in gevaar zijn. Bovendien bevordert het gecoördineerde plan, zoals reeds besproken, de ontwikkeling en ingebruikname van mensgerichte, betrouwbare, veilige en duurzame AI-technologieën.

Samengevat komen de aanbevelingen van Nowotny goed overeen met de specificaties van het gecoördineerde plan inzake AI van de EC en de AI-verordening, met enkele genuanceerde verschillen. Ik ben daarom blij met de aanbevelingen van Nowotny, omdat ze een waardevol perspectief bieden dat de bestaande plannen en regelgeving aanvult en verrijkt.

Tot slot, met betrekking tot de toekomstperspectieven, bereiden we verschillende maatregelen voor, van fundamenteel onderzoek tot implementatie en financieringsinstrumenten, om Europa op de kaart te zetten wat betreft generatieve AI.

5. Verslagen van de Stakeholder Workshops

Stakeholder Workshop I – 12 september 2023[15]

Thema: AI als aanjager van verandering: Hoe worden AI en ChatGPT gebruikt, ervaren en ondersteund in het formele onderwijs en in de informatie aan en opleiding van het bredere publiek in Vlaanderen?

De Denker en de stuurgroep hebben vooraf een aantal vragen uitgedeeld aan de deelnemers.

- Wat zijn de huidige toepassingen van AI/ChatGPT in en buiten het klaslokaal en met welk doel?
- Wat is jouw ervaring en die van je studenten tot nu toe?
- Is er nieuwsgierigheid en enthousiasme?
- Welke beperkingen en problemen heb je ondervonden?
- Wat is de rol van de leerkracht? Hoe kan hij of zij helpen?
- Voelen leerkrachten zich alleen gelaten? Hebben ze ondersteuning nodig van andere disciplines?
- Zijn de institutionele richtlijnen voldoende?
- Wat is er nog meer nodig om een productief gebruik voor iedereen mogelijk te maken?
- Waar zou je over 3 jaar willen zijn?
- Denk aan nieuwe toepassingen, vragen, bereik: heb je een voorbeeld?
- Hoe voorzie je te navigeren tussen positieve en negatieve toekomstbeelden?

Een belangrijke vaststelling die de deelnemers deelden, is dat er al een wijdverspreide en diverse reeks acties en initiatieven bestaat rond het thema generatieve AI voor onderwijs en onderzoek en communicatie naar de maatschappij en het grote publiek (zie bijlage[16]).

Generatieve AI wordt in Vlaanderen al op grote schaal gebruikt in het lager, middelbaar en universitair onderwijs, in de vrije tijd en in professionele omgevingen.

---

[15] Deelnemers: Helga Nowotny (denker), Charlotte Vandooren (imec/RVO-Society), Johan Suykens (Ingenieurswetenschappen, hoofd van het programma Master of Artificial Intelligence, KU Leuven), Cynthia Van Hee (Taalkunde, UGent), Els Lefever (Taalkunde, UGent), Hugo De Man (Ingenieurswetenschappen, KU Leuven, imec, KVAB), Joos Vandewalle (Ingenieurswetenschappen, KU Leuven, coördinator, KVAB).

[16] Bijlage Initiatieven en acties rond het gebruik van AI/ChatGPT in Vlaanderen voor onderwijs, opleiding en informatie buiten de onderzoeksgemeenschap: Aanbod voor onderwijs (amai.vlaanderen), AI voor Vlaanderen, AI4Belgium, Scivil, Kenniscentrum Data & Maatschappij, Digisoc, amai! imec-SMIT-VUB, imec-UGent-IDLab, imec-UA-IDlab, CiTiP-KUL, RVO-Society Brightlab, Nerdland, burgerprojecten, pptx ChatGPT voor scholen, FARI-conferentie VUB-ULB 11-12 sept. 2023, enz.

We kunnen hier een aantal voorbeelden noemen, zoals het corrigeren van taal, het genereren van rapporten, het beoordelen en samenvatten van documenten en onderzoeksartikelen, als zoekmachine enz. Over het algemeen is het publiek zich ervan bewust dat deze diensten nog maar aan het begin staan en in de toekomst waarschijnlijk bij ons zullen blijven.

De stakeholders zijn het erover eens dat generatieve AI en algemene AI-diensten geen magie zijn, maar vooral een wiskundige optimalisatie van voorspellende modellen die worden getraind met een enorme rekenkracht en met behulp van enorme datasets. Bovendien verbruikt ze enorme hoeveelheden energie, kan de dataset onbetrouwbaar en bevooroordeeld zijn en is er weinig respect voor de privacy en het auteurschap van teksten. Het publiek en de jongere generaties zijn zich hier echter onvoldoende van bewust.

De grote bedrijven hebben de controle over de gegevens en de media, en de individuele gebruiker heeft geen inzicht in en begrip van het proces, de waarde, de beperkingen en de geldigheid en daardoor vaak een vals geloof in de betrouwbaarheid ervan. De grote bedrijven hebben deze technologie te vroeg op het publiek losgelaten, met als doel winst te maken. Universiteiten en openbare onderzoekscentra beschikken niet over dezelfde hoeveelheid middelen en moeten zich daarom beperken tot kleinere en meer gerichte thema's, die met meer objectiviteit en eerlijkheid worden behandeld. Toch zijn er nog heel wat onderzoeksvragen open, zoals de uitlegbaarheid van generatieve AI, de correctheidsbewijzen, hallucinaties enz. Als we de controle loslaten, besteedt AI de geproduceerde kennis uit op een oncontroleerbare en onverklaarbare manier. Het publiek is zich er niet volledig van bewust dat het niet communiceert met een mens, maar met een machine die gebruik maakt van oncontroleerbare en vaak onbetrouwbare enorme datasets.

Daarom is er een nieuwe vorm van wetenschapscommunicatie nodig. Wetenschappers moeten een dialoog aangaan met de burger en uitleggen dat wetenschap 'georganiseerd scepticisme' is. Ze dienen het publiek niet voor te houden dat het weigeren van wetenschappelijk bewijs hetzelfde is als een gebrek aan begrip. Ze moeten uitleggen hoe wetenschap werkt, via interactie en experimenten, en dat ze op deze manier kan bijdragen aan hun leven, werk en onderwijs. Publieke betrokkenheid bij de wetenschap wordt steeds belangrijker. De wetenschap heeft de morele plicht om te voorkomen dat mensen gedesoriënteerd raken.

De overheid, het bedrijfsleven, het onderwijs, de media en de kunsten zijn zich bewust van en geïnteresseerd in deze nieuwe diensten en zijn bereid om actie te ondernemen. Het onderwerp staat dus open voor een publiek debat waaraan gewone burgers, jong en oud, graag deelnemen.

De studenten van het programma Master of Artificial Intelligence zijn enthousiast over de gebruiksmogelijkheden ervan, bijvoorbeeld om een eerste idee over een

onderwerp te krijgen, om een tekst te corrigeren of om een programmeercode te genereren. Ze zijn zich ook bewust van de beperkingen. Aangezien ChatGPT vaak niet betrouwbaar is wat betreft feitelijke informatie, moet men de programma's verder debuggen, referenties verder controleren op hun juistheid en het bestaan ervan, enz. ChatGPT is nog geen betrouwbare tool, hoewel het wel al een aantal indrukwekkende functies heeft. Bovendien kan het niet worden gebruikt met gevoelige of persoonlijke gegevens vanwege privacykwesties, of met gegevens of onderwerpen die onder een geheimhoudingsovereenkomst vallen. In het programma Master of Artificial Intelligence aan de KU Leuven is een sjabloonbestand gecreëerd dat door de studenten moet worden ingevuld voor alle rapportering in het kader van cursusopdrachten en voor de masterproef. Het bevat bovendien een gedragscode met betrekking tot alle mogelijke vormen van gebruik van ChatGPT en andere AI-hulpmiddelen bij het schrijven. Er wordt ook uitgelegd hoe je aan het examenreglement kunt voldoen. Aan de KU Leuven zijn richtlijnen ontwikkeld voor een verantwoord gebruik van generatieve AI-tools in onderzoek en onderwijs, waarbij ook rekening is gehouden met de gedragscode die is voorgesteld door het programma Master of Artificial Intelligence.[17]

De kwaliteit van tools zoals ChatGPT zou moeten verbeteren. Vooral feitelijke informatie en referenties moeten correct en betrouwbaar zijn.

Als feitelijke informatie op de juiste manier wordt geïmplementeerd, zal AI een zeer belangrijk en ontwrichtend hulpmiddel worden voor de toekomst waarin veel opwindende nieuwe dingen mogelijk zullen worden; anders zijn de vooruitzichten somberder. De inzet van AI-technologie zal vermoedelijk ook nieuwe, nog onvoorziene problemen scheppen. Het is op dit moment moeilijk te voorspellen hoe deze technologie zich zal ontwikkelen.

De opleiding van taalleerkrachten en vertalers wordt nu al vrij verregaand beïnvloed. Hun opleidingsinstellingen moeten natuurlijk een verantwoord en inclusief gebruik van deze digitale diensten omvatten en voorbereidingen treffen voor een breder gebruik. Ook moeten de beperkingen van de diensten en de essentiële toegevoegde waarde van menselijke kennis, standpunten, emoties en wijsheid worden uitgelegd. De vooruitzichten voor professionals in deze vakgebieden zijn dus aan het veranderen, maar zijn niet somber.

De rol van ingenieurs, datawetenschappers en ontwerpers van deze diensten verandert, en hun opleiding zou moeten volgen. Hun werkzaamheden zullen niet langer los staan van het publiek en van de specialisten in de menswetenschappen

---

[17] https://nieuws.kuleuven.be/en/content/2023/ku-leuven-opts-for-responsible-use-of-generative-artificial-intelligence-in-researchand-education
https://www.kuleuven.be/onderwijs/student/onderwijstools/artificiele-intelligentie
https://research.kuleuven.be/en/integrity-ethics/integrity/practices/genai_nl/genAI_NL

en de sociale wetenschappen. Er is momenteel op dit gebied sprake van zowel een groeiende bewustwording als actie op internationaal vlak en in Vlaanderen.[18]

Hoewel de deelnemers een diverse achtergrond, activiteitendomein en carrièrepad hebben, is er een gemeenschappelijk begrip en gedeelde overtuiging dat de besproken thema's een aantal belangrijke inzichten, problemen en aandachtspunten bevatten die een allesomvattende aanpak vereisen in Vlaanderen, zowel voor het formele onderwijs van lagere scholen over middelbare scholen tot universiteiten als voor het grote publiek in hun dagelijkse activiteiten binnen gezin en vrije tijd. Zo'n plan moet worden geïnstitutionaliseerd en er moet worden gewerkt met deskundigen ter plaatse die zich richten op verschillende doelgroepen. Dit moet leiden tot een dieper begrip, demystificatie, een evenwichtige kijk op de voordelen en een kritische houding. Op deze manier worden de mensen de aanjagers van verandering.

Stakeholder Workshop II – 15 september 2023[19]

Thema: AI als aanjager van verandering: Wat is de multidisciplinaire ervaring van de Vlaamse onderzoekers die actief zijn in fundamenteel AI-onderzoek en AI-toepassingen met betrekking tot de maatschappelijke aspecten van AI, en meer specifiek ChatGPT?

De Denker en de stuurgroep hebben vooraf een aantal vragen uitgedeeld aan de deelnemers. De deelnemers gaven eerst enkele opmerkingen over het algemene thema.

"Verantwoorde AI" is al een belangrijk onderdeel van het Vlaams AI-onderzoeksprogramma[20], waarbij verschillende deelnemers betrokken zijn. Dit omvat een uitgebreid programma voor strategisch basisonderzoek in artificiële intelligentie, gebaseerd op de noden en wensen van bedrijven, organisaties, de overheid en haar burgers. Al is duidelijk dat de meeste deelnemers zich bewust zijn van de belangrijkste maatschappelijke problemen, dit geldt minder voor het publiek. Met

---

[18] Het algemeen belang centraal stellen in AI, data en robotica onderzoek, https://www.fari.brussels/nl

[19] Deelnemers: Helga Nowotny (denker), Sabine Demey (imec/FlandersAI), Tony Belpaeme (Ingenieurswetenschappen, UGent), Tijl DeBie (Ingenieurswetenschappen, UGent), Stein Aerts (Geneeskunde, VIB, KU Leuven), Sien Moens (Computerwetenschappen, KU Leuven), Peter Spyns (Departement EWI, Vlaanderen), Tias Guns (KU Leuven), Thomas Demeester (UGent), Pedro Goncalves (nerf), Guillermo Perez (UAntwerpen), Hugo De Man (Ingenieurswetenschappen, KU Leuven, imec, KVAB), Joos Vandewalle (Ingenieurswetenschappen, KULeuven, coördinator, KVAB), Ine Van Hoyweghen (KU Leuven, coördinator, KVAB), geschreven antwoorden door Johan Suykens (Ingenieurswetenschappen, KU Leuven).

[20] https://www.flandersairesearch.be/nl

sommige problemen kan rekening worden gehouden in de fase van het fundamenteel onderzoek en het ontwerp van de methoden voor machine leren, terwijl de meeste moeten worden aangepakt binnen de specifieke toepassingscontext wanneer de methoden voor machine leren worden toegepast in diensten en producten.

1 - Hoe ga je om met de problematische aspecten van AI (vooroordelen, uitlegbaarheid, verantwoording, transparantie, fairness)?

Over het algemeen beschouwden de deelnemers AI als gunstig voor de wetenschap, met veel mogelijkheden voor het gebruik ervan in hun onderzoekspraktijk. Vooral het gebruik van generatieve AI werd beschouwd als een game-changer voor de wetenschap. Ze erkenden echter allemaal dat er rekening moet worden gehouden met belangrijke problemen en risico's van AI. Problematische aspecten van AI zijn vooroordelen, uitlegbaarheid, verantwoording, transparantie en eerlijkheid.
Vooroordelen en eerlijkheid zijn sterk met elkaar verbonden en op dezelfde manier is uitlegbaarheid gekoppeld aan transparantie in AI. Uitlegbaarheid/transparantie is een voorwaarde voor verantwoording en eerlijkheid, maar ook voor het vertrouwen en de machtiging van eindgebruikers en consumenten. Wanneer men ze rangschikt naar belangrijkheid, wordt uitlegbaarheid/transparantie op het hoogste niveau geplaatst, omdat ze in zekere zin fundamenteler is.

Aspecten van vooroordelen in besluitvorming kunnen bijvoorbeeld verband houden met het vooroordeel in een classificator, de bemonstering van diversiteiten, en andere. Er is veel onderzoek gedaan naar uitlegbare AI. Men richt zich ook op reproduceerbaar onderzoek en open source software, wat leidt tot meer transparantie. Wat de verantwoording van AI betreft, is er een onderscheid nodig tussen algemeen methodologisch onderzoek naar machine learning en AI in een specifieke toepassingscontext. Men kan een intuïtieve mening hebben over eerlijkheid, over wat goed of fout is, maar dit kan afhangen van de toepassingscontext en het kan moeilijk zijn om dit op een objectieve manier te kwantificeren. Omdat bovendien de definitie van eerlijkheid tussen verschillende disciplines (bv. informatica, recht, economie, filosofie, sociologie) uiteenloopt, is er behoefte aan transdisciplinair onderzoek naar eerlijkheid in AI.

Er wordt in Vlaanderen actief onderzoek gedaan naar verschillende van deze kwesties. Eén onderzoeker houdt zich expliciet bezig met uitlegbaarheid en transparantie. Binnen het veld van uitlegbare AI is er een groep experts in formele methoden die dit algoritmisch benaderen vanuit twee invalshoeken. Ten eerste ontwikkelen ze alternatieve/aangepaste AI-technieken die verklaringen geven of een efficiënte controle van mogelijke verklaringen mogelijk maken. De tweede voorgestelde benadering is om een "Vanilla" AI-model te nemen en te controleren of het voldoet aan een specificatie in een taal die (door experts) als verklaarbaar wordt beschouwd (gerelateerd aan de verificatie van computersystemen). Een ander lid werkt vooral aan de transparantie van generatieve modellen: het in-

bouwen van een extra controle in voorstellingen en neurale architecturen (bv. in tekst-naar-beeld-synthese/diffusie). Andere leden in Vlaanderen doen onderzoek naar causale modellen, causale relaties en logisch redeneren. Over het algemeen gaven de deelnemers aan dat er behoefte is aan fundamenteel onderzoek naar generatieve AI.

- Welke zaken zijn voor jou het meest problematisch en waarom?

Vooral uitlegbaarheid en eerlijkheid zijn moeilijke kwesties om aan te pakken: vaak ziet men een afweging tussen nauwkeurigheid en uitlegbaarheid. Eerlijkheid is meestal een subjectief begrip. Een andere observatie is dat de gegevens en hun verdelingen waarop onze huidige grote basismodellen worden getraind, niet bekend zijn. Bovendien zijn alle details van de modellen die gebruikt worden voor het trainen van de basismodellen onbekend. Regelgevers en wetenschappers moeten toegang hebben tot de interne werking van deze modellen – hoe ze zijn getraind en met welke datasets. Als deze modellen worden omgezet in gesloten producten, zullen ze niet beschikbaar zijn voor grondige inspectie, replicatie en testen. Het kwantificeren van de verantwoording is misschien wel het moeilijkst, omdat deze nauw verbonden is met juridische aspecten. Meer in het algemeen: is de programmeur, het bedrijf, de software of zijn de gegevens verantwoordelijk als er iets fout gaat door een bug of een aanval? Door AI toe te voegen aan de mogelijke boosdoeners wordt dit alleen maar complexer, want je hebt de training van de gegevens, de ontwerper van het AI-algoritme, de persoon die het AI-kader heeft gekozen, enzovoort.

De deelnemers vermelden ook de bredere maatschappelijke impact van AI. Er werd bijvoorbeeld gediscussieerd over de veiligheid van My AI op Snapchat, en vooral chatbots met ChatGPT werden als zorgwekkend beschouwd, vooral voor kinderen, omdat ze vatbaar zijn voor propaganda en cyberaanvallen.

Hoewel sommige leden optimistisch zijn over technologie in het algemeen en AI in het bijzonder, en daarbij verwijzen naar overregulering zoals dat is gebeurd met genetisch gemodificeerd voedsel, beschouwen de meeste leden regulering als noodzakelijk om maatschappelijke schade te voorkomen. De Europese Commissie heeft in april 2021 een voorstel gedaan voor een EU-regelgevingskader voor artificiële intelligentie (AI)[21]. Het voorgestelde rechtskader richt zich op de specifieke toepassing van AI-systemen en de eraan verbonden risico's. De Commissie stelt voor om een technologie neutrale definitie van AI-systemen op te nemen in de EU-wetgeving en om een classificatie voor AI-systemen vast te stellen met verschillende eisen en verplichtingen die zijn afgestemd op een

[21] AI-verordening
https://www.europarl.europa.eu/news/nl/headlines/society/20230601STO93804/ai-verordening-eerste-regels-voor-artificiele-intelligentie

"risico gebaseerde aanpak". Sommige AI-systemen met "onaanvaardbare" risico's zouden verboden worden.

- Als we het eens zijn over de noodzaak van regulering, welke technische hindernissen moeten dan overwonnen worden?

Het is relatief eenvoudig om een "verlanglijstje" op te stellen van wenselijke eigenschappen van AI-systemen (bijvoorbeeld met betrekking tot vooroordelen, uitlegbaarheid, verantwoording, transparantie, eerlijkheid). Er kunnen echter fundamentele theoretische beperkingen bestaan van wat haalbaar is (bv. afwegingen tussen robuustheid en nauwkeurigheid, prestatie en uitlegbaarheid, transparantie en concurrentievermogen van het bedrijf, enz.). Het is vaak moeilijk om in zeer algemene termen over AI-systemen te praten, omdat je uiteindelijk rekening moet houden met de specifieke toepassingscontext. Aangezien men rekening moet houden met het AI-systeem, de gegevens, de toepassingscontext, de ontwerper, de gebruiker en het doel van het AI-systeem, zal het vaak een interdisciplinaire en transdisciplinaire taak zijn om het AI-systeem te realiseren en te implementeren. Er is meestal expertise nodig uit verschillende vakgebieden.

In de loop van de voorbije jaren zijn verschillende begrippen van uitlegbaarheid/transparantie en vooroordelen/eerlijkheid wiskundig geformaliseerd, wat een vereiste is om ze te kunnen opleggen aan AI-algoritmes en -systemen. Er is echter een grote kloof tussen deze wiskundige begrippen en de juridische perspectieven op dezelfde concepten. Zelfs in de eenvoudigste AI-opzet van binaire classificatie kan eerlijkheid op meerdere, onverenigbare manieren worden geformaliseerd en de verschillen daartussen zijn, zelfs als ze echt van belang zijn voor individuen en de samenleving, moeilijk uit te leggen aan niet-technische mensen, zoals zakelijke besluitvormers en juridische of ethische adviseurs. De modellen veranderen zeer snel, we kunnen het ons niet veroorloven om gewoon te zeggen dat het zwarte dozen zijn, want dat zijn ze niet. We hebben experts nodig die de berekeningen en optimalisatieprocessen begrijpen en op de hoogte blijven van de nieuwste evaluaties in de technologie, en deze wetenschappers moeten toegang krijgen tot de training van de modellen. Verschillende deelnemers zijn van mening dat regelgeving waarschijnlijk niet zal werken en vergelijken dit met de regelgeving over cookies, die in de praktijk het wijdverspreide gebruik van cookies niet heeft beperkt.

Eén deelnemer stelde dat overheden zich zouden moeten onthouden van het reguleren van computersystemen of AI in het bijzonder, omdat de nationale en internationale wetgeving te traag gaat om de ontwikkelingen op het gebied van AI bij te houden. In plaats daarvan moet prioriteit worden gegeven aan het aanmoedigen van bedrijven om standaardprocessen op te zetten voor het valideren en verifiëren van AI-systemen. Zie bijvoorbeeld het verificatie- en validatieplan voor NASA-technologie (uit hun handboek voor systeemengineering dat online beschikbaar is). Dit zou de discussie concreter moeten maken.

In antwoord op dit probleem bevat de AI-verordening van de EU meerdere bepalingen over (onder andere) uitlegbaarheid/transparantie en vooroordelen/eerlijkheid, maar deze blijven vaag vanuit het perspectief van computerwetenschappers en AI-systeemontwikkelaars. De vraag hoe deze vage eisen vertaald kunnen worden naar hun concrete implementatie blijft onbeantwoord.

- Wat moeten we doen om deze hindernissen weg te nemen? Wat is de rol van bedrijven en de rol van individuele onderzoekers?

De eisen die worden gesteld op het vlak van transparantie verschillen momenteel tussen universiteiten en bedrijven, wat een onevenwicht veroorzaakt, vooral omdat fundamenteel AI-onderzoek steeds vaker ook in (grote) bedrijven wordt gedaan.

De AI-ontwikkelingen bevinden zich in een versnellingsfase. Bovendien heeft AI een stabiele omgeving nodig om te kunnen floreren. De geopolitieke status van de wereld is echter instabiel, wat fundamentele problemen veroorzaakt. Daarom zal het essentieel zijn om een wereldwijde internationale overeenkomst over gemeenschappelijke principes te bereiken.

Ingenieurs moeten worden opgeleid en getraind. Als we geen AI-experts trainen in de bovenstaande vereiste vaardigheden, hebben we een gebrek aan bedrijfsmedewerkers en onderzoekers met deze vaardigheden. Meer in het algemeen moeten we studenten ICT, ingenieurswetenschappen, wiskunde en statistiek opleiden om de interne werking te begrijpen van de (voor natuurlijke taalverwerking, beeldverwerking, spraak enz.) gebruikte neurale architecturen, om de onderliggende wiskunde en optimalisatieprocessen te begrijpen. Dit wordt beschouwd als een morele plicht. Bovendien moeten deze experts worden opgeleid over expertise uit andere disciplines (recht, sociologie, ethiek). Daarnaast vormen deze kritieke vaardigheden een basis voor innovatie en zijn ze absoluut nodig in het snel veranderende AI-landschap. Er zijn volledig nieuwe methoden en instrumenten nodig om deze kloof tussen technische benaderingen en sociale/juridische/ethische perspectieven te overbruggen, gericht op het betrekken van alle stakeholders met zeer uiteenlopende achtergronden en belangen. Dergelijke instrumenten moeten helpen bij het ontwerpen van een aanpak voor deze problemen die breed wordt gedragen, gegarandeerd wettelijk is en technisch haalbaar is.

De deelnemers geloven dat onderzoek een belangrijke bijdrage kan leveren aan deze uitdaging. Hiervoor is samenwerking tussen verschillende disciplines van cruciaal belang (toepassingen, AI-basismodellen, formalisering van eerlijkheid/vooroordelen, gebruikersinterfaces, recht en ethiek, sociologie enz.).

2 - Er bestaan veel ethische richtlijnen die elkaar deels overlappen, maar vaak zijn ze onvoldoende gespecificeerd.

- Wat ontbreekt er het meest om ze te controleren en te implementeren? Zijn er sancties nodig?

Wat typisch is voor richtlijnen en voorschriften, is dat ze vaak vaag of onvoldoende gespecificeerd zijn om veel en vooral toekomstige situaties te dekken. Daarom moeten deze worden geëvalueerd en geïnterpreteerd in het licht van de meest recente technologische ontwikkelingen. Als sancties inhouden dat bepaalde technologieën verboden worden, kan dit de technologische vooruitgang belemmeren. Op dit moment moet men zich vooral richten op AI-toepassingen met een hoog risico, zoals gedefinieerd in bijvoorbeeld het voorstel van de EU voor een AI-verordening https://www.europarl.europa.eu/news/nl/headlines/society/20230601STO93804/ai-verordening-eerste-regels-voor-artificiele-intelligentie

Het is ook het beste om dit voor elke toepassingssector te controleren, rekening houdend met de specifieke toepassingscontext.

Ondanks de vele resterende uitdagingen begint de onderzoeksgemeenschap grip te krijgen op kwesties als transparantie/uitlegbaarheid, vooroordelen/eerlijkheid, privacy enz. Anderzijds is er onvoldoende aandacht voor de bredere maatschappelijke, pedagogische en psychologische effecten van AI.

Wat volgens een ander lid het meest ontbreekt, zijn richtlijnen en regelgeving rond sociale chatbots en het gebruik van "conversational AI" in de vorm van socialemediabots. (Aangezien dit niet verboden of zelfs maar gereguleerd is, zijn sancties natuurlijk niet aan de orde.)

Wat betreft het reguleren van dergelijke toepassingen van AI is één deelnemer van mening dat de risicogebaseerde benadering van niet-generatieve AI in het ontwerp van de AI-verordening van de EU zeer verstandig is. Helaas bestaat de vrees dat de amendementen van de Raad van de EU en in het bijzonder van het Europees Parlement, die gericht zijn op de generatieve AI-technieken en de basismodellen die aan deze toepassingen ten grondslag liggen en die dus afwijken van de risicogebaseerde benadering, onnodige obstakels zullen opwerpen voor innovatie in generatieve AI en basismodellen, terwijl tegelijkertijd de risicovolle toepassingen daarvan niet adequaat worden onderkend en gereguleerd. Met name de risico's op het gebied van grootschalige desinformatie en manipulatie worden dan onderschat, vooral bij toepassingen waarbij mensen waarschijnlijk een emotionele band aangaan met AI-systemen, zoals sociale chatbots.

- Wat staat de controle en implementatie in de weg?

Nu AI steeds sneller evolueert, zijn nieuwe ontwikkelingen en toekomstige AI-systemen en -diensten moeilijk te voorspellen. Het publiek en de politiek zijn zich onvoldoende bewust van de potentiële risico's, omdat deze subtieler lijken te zijn dan bijvoorbeeld directe discriminatie of inbreuken op de privacy op gebieden die in het ontwerp van de AI-verordening als risicovol worden aangemerkt.

- Verder kijkend dan de ethiek: Als we rekening willen houden met de waarschijnlijke maatschappelijke gevolgen, hoe kunnen we dat dan het beste doen?

De deelnemers benadrukten de relevantie van onderzoek naar de maatschappelijke impact van AI met betrekking tot het huidige onderzoek op dit gebied in Vlaanderen. Er is op zijn minst een publiek en politiek debat nodig over de rol die we aan mensachtige AI-systemen willen geven in onze samenleving. Voorzichtigheid is vooral geboden voor minderjarigen, vooral bij AI-systemen die waarschijnlijk leiden tot een emotionele binding, laat staan bij systemen die voor dergelijke doeleinden op de markt worden gebracht. Deze moeten worden gereguleerd op het niveau van de toepassing (zoals het geval was in het oorspronkelijke ontwerp), in plaats van op het niveau van de technologie (zoals in de laatste amendementen, in ieder geval voor generatieve AI en basismodellen).

Er kunnen ernstige gevolgen voor de arbeidsmarkt worden verwacht. Sommige jobs kunnen in omvang afnemen of verdwijnen, en andere kunnen veranderen, maar AI zal nieuwe mogelijkheden blijven bieden voor jobs in veel disciplines.

3 - Als onderzoekers hebben we ook de verantwoordelijkheid om kritisch denken, digitale geletterdheid en vertrouwen bij te brengen aan de jongere generatie en het publiek.

- Hoe kunnen we een democratisch debat over het bestuur van AI voeren?

Voor het grote publiek heeft het geen zin om deze algemene vragen te stellen, omdat dit leidt tot zinloze discussies die de bevolking alleen maar polariseren. In plaats daarvan moeten de debatten zich richten op concrete vragen en AI-praktijken. Het is dus het beste om een democratisch debat te voeren over specifieke AI-onderwerpen met een grote impact, zoals ChatGPT, zelfrijdende auto's, AI-gezondheidstoepassingen enzovoort, waarbij de voor- en nadelen van nieuwe ontwikkelingen worden besproken. Zorg ervoor dat een dwarsdoorsnede van de samenleving en van de wetenschappelijke gemeenschap betrokken is bij dit debat – d.w.z. het debat moet worden ondersteund door AI-experts, maar mag niet door hen worden gedomineerd. Dit zal ervoor zorgen dat er rekening wordt gehouden met de bredere maatschappelijke context en de behoeften en zorgen van alle leden van de samenleving. In dit opzicht moeten we de fouten vermijden die we tijdens de pandemie gezien hebben.

Het is ook belangrijk om technologische experts en experts in andere disciplines op te leiden over dit snel veranderende veld.

We moeten het publiek bewust maken van de waarde van de gegevens die ze creëren. De gegevens die wij als publiek aan de grote techbedrijven geven, zijn misschien wel meer waard dan de diensten die de grote techbedrijven in ruil daarvoor aan het publiek geven. Misschien kunnen we andere bedrijfsmodellen bedenken.

Onlangs toonden de doctoraatsstudenten AI in Vlaanderen veel interesse in de maatschappelijke aspecten van AI tijdens een studiedag. In de ingenieursfaculteiten van de Vlaamse universiteiten zouden studenten graag meer opleiding krijgen over maatschappelijke aspecten van AI. Dit soort opleiding zou moeten doorsijpelen naar de maatschappij. Burgers moeten leren hoe chatbots worden aangedreven door neurale netwerken en wat hun training inhoudt – mechanismen die kritisch denken vereisen dat verder gaat dan louter digitale geletterdheid.

- Kan AI een publiek goed worden? Onder welke omstandigheden?

De meeste deelnemers hopen dat dit zal gebeuren. De academische wereld zou hier samen met de open source-gemeenschap een cruciale rol in moeten spelen. Als basismodellen een publiek goed kunnen worden, is dat ook goed voor het klimaat omdat het trainen van deze modellen veel energie kost. Europese academische instellingen zouden stimulansen kunnen geven om dergelijke publieke modellen te bouwen.

AI is al een publiek goed, omdat de meeste ontwikkelingen worden beschreven in voorpublicaties die online worden geplaatst en gratis toegankelijk zijn, als een vorm van open wetenschap. Wat geen publiek goed is, is de gedetailleerde kennis over hoe je AI-modellen moet trainen en gebruiken. Nog belangrijker is dat de gegevens die gebruikt worden om ze te trainen ook geen publiek goed zijn. Naast het verplicht stellen van programmeren in het middelbaar onderwijs, zou de basis van het gebruik van AI-technieken waarschijnlijk ook verplicht moeten worden, omdat deze een standaard hulpmiddel zullen worden bij programmeren en computertechniek in het algemeen.

- Wat is de rol van interdisciplinariteit in dit alles?

Alle deelnemers zijn ervan overtuigd dat diepgaande interdisciplinariteit niet alleen belangrijk, maar ook van vitaal belang is tijdens de verschillende stadia van AI-onderzoek en -ontwikkeling. Er kunnen verschillende niveaus van inzicht in de technologieën worden overwogen, en niet alle onderzoekers hebben dezelfde diepgang nodig. Aan de ene kant zijn er de computer-, informatie- en datawetenschappen en de wiskundige grondslagen, en aan de andere kant

verschillende disciplines in de geesteswetenschappen, zoals sociologie, rechten, psychologie, onderwijs, taalkunde en antropologie, en daarbovenop de interactie met de samenleving en de burgers.

6. Conclusies en aanbevelingen van de Denkerscyclus

Over een periode van bijna een jaar is deze Denkerscyclus erin geslaagd om interessante reflecties, nieuwe discussies en een meer duurzaam bewustzijn te genereren – zowel in Vlaanderen als binnen de KVAB – over de belangrijke rol van AI in onze samenleving en in tal van wetenschappelijke activiteiten. De bekende interdisciplinaire en visionaire denker prof. Helga Nowotny stimuleerde deze verschillende activiteiten enorm met haar bijdragen. In haar studie *In AI We Trust* uit 2021 behandelt ze onthullende historische parallellen en presenteert ze belangrijke interdisciplinaire inzichten. Samen met de recente vooruitgang en het wijdverspreide gebruik van generatieve AI heeft dit geleid tot een beter begrip van de voortdurende transformatie in de wetenschap en de maatschappij van vandaag. De deelnemers waren het duidelijk eens over de dringende behoefte aan gezamenlijke acties en intensieve samenwerking tussen experts in AI-wetenschap en -technologie en wetenschappers in de sociale en geesteswetenschappen. Het is duidelijk dat algemene AI en generatieve AI veel nieuwe mogelijkheden bieden voor zowel gebruikers als wetenschappers, maar tegelijkertijd gaan ze gepaard met een reeks potentiële risico's en sociale schade, zoals vooroordelen, ongewenste profilering/discriminatie, misbruik van gegevens en toenemende sociale ongelijkheid.

Bovendien is er zowel in de wetenschap als bij de gebruikers behoefte aan betere informatie, meer inzicht en meer reflectie over ethisch handelen. Het is bijvoorbeeld belangrijk om het bewustzijn te stimuleren dat AI-tools geen magie zijn, maar dat ze worden geproduceerd door wiskundige optimalisatie met behulp van een enorme computerkracht en op basis van enorme datasets. Op al deze fronten zijn er grote zorgen. De wiskundige optimalisatie en de algoritmes zijn solide, maar ze bieden geen rechtvaardiging of verklaring voor de AI-producten als zodanig. De computerkracht en de grote datasets zijn wereldwijd in handen van bijna-monopolies. Bovendien ligt de energie die computers vandaag de dag nodig hebben om deze modellen bij te werken al op het niveau van het energieverbruik van een redelijk groot land. Met andere woorden: aan de verschillende datasets hangt een prijskaartje en, erger nog, ze kunnen verkeerde informatie bevatten die kan leiden tot valse resultaten.

In deze context bespraken en becommentarieerden experts, stakeholders, AI-praktijkmensen en de stuurgroep een inspirerende tekst van Helga Nowotny die vooraf was uitgedeeld. Men kan zeggen dat er een goede overeenstemming was tussen alle deelnemers, zoals beschreven in de verslagen van de stakeholderbijeenkomsten, terwijl de verschillende experts inspirerende reflecties bijdroegen (zie hoofdstuk 5). Op basis van de verschillende discussies werd een reeks van drie specifieke aanbevelingen geformuleerd.

Aanbeveling 1: **We bevelen aan een brede publiekscampagne te lanceren** onder het motto "AI voor burgers – burgers voor AI" om burgers te ondersteunen bij het gebruik van AI in hun dagelijks leven en voor een betere samenleving.

Het doel is om het begrip van de werking van AI en digitale systemen te verdiepen en te verspreiden, het potentieel van huidige en toekomstige toepassingen en het gebruik ervan te onderzoeken, en te leren over hun beperkingen.

De vele reeds bestaande en nieuwe initiatieven moeten een officieel mandaat krijgen om

1. de educatieve inspanningen gericht op deze doelen onderling te coördineren;
2. hun respectievelijke doelgroepen (leeftijdsgroepen, formele en informele settings); de middelen en materialen die ze gebruiken, testen en ontwikkelen (bv. voor leraren in het basis- en secundair onderwijs); en vormen van samenwerking met universiteiten, media, de kunsten en het bedrijfsleven te specificeren en in kaart te brengen;
3. voldoende ruimte te creëren voor een voortdurende uitwisseling van ervaringen en wederzijds leren, over academische disciplines en generaties heen;
4. ervoor te zorgen dat alle onderwijsinspanningen een digitaal humanistisch perspectief bevatten (en dus veel verder gaan dan digitale geletterdheid) https://informatics.tuwien.ac.at/digital-humanism/

Daartoe moet een solide institutioneel kader worden opgezet en voorzien van de nodige financiële en personele middelen, in eerste instantie voor een periode van drie jaar, en hernieuwbaar na evaluatie.

Aanbeveling 2: **We bevelen aan om fundamenteel AI-onderzoek een hoge prioriteit te geven** en uit te voeren volgens de lijnen van de Europese Onderzoeksraad (ERC) (bottom-up, uitgaand van een hoofdonderzoeker). Dit dient als tegenwicht voor de dominantie van een eendimensionaal "technologisch oplossingsdenken", dat alternatieven negeert en/of terzijde schuift bij de keuze van onderzoeksproblemen, methoden en technieken. Hierdoor ontstaat bovendien een meer humanistisch begrip van de reikwijdte en de diepte van de menselijke ervaring en wat het betekent om mens te zijn.

De huidige overconcentratie van de financiering van AI-gerelateerd O&O in de private sector leidt tot een zorgwekkend onevenwicht voor (voornamelijk) universitair onafhankelijk onderzoek met betrekking tot de toegang tot rekenkracht, trainingsgegevens, het aantrekken van talent en het pionieren in nieuwe onderzoeksrichtingen. In het belang van AI als een publiek goed moeten deze nadelen worden aangepakt.

Als onderzoeksgebied is AI, inclusief machine leren en generatieve AI, relatief jong, terwijl een historisch perspectief grotendeels ontbreekt, vooral in Europa. Hierdoor bestaat er een grote kans op het verlies van waardevolle technische kennis, wiskundige concepten, technieken en wetenschappelijke inzichten. Veelbelovende onderzoekslijnen werden vaak voortijdig afgesloten. Alleen een sterke focus op fundamenteel onderzoek kan de aanzet geven tot hun herontdekking en de verdere verkenning van historische paden die niet werden bewandeld.

Aanbeveling 3: **We bevelen een krachtige ondersteuning aan van onderzoek naar de maatschappelijke impact van AI wat betreft aspecten en gebieden die waarschijnlijk niet worden opgepakt door de grote internationale bedrijven.**

Omdat we nog maar aan het begin staan van het systematisch volgen en analyseren van de mogelijke nuttige toepassingen van AI voor verschillende groepen in de samenleving en het leren over het vermijden van sociale schade, is het cruciaal om de snel evoluerende ervaringen, stemmen en behoeften van burgers mee te nemen.

Studenten AI en aanverwante technische gebieden (en hun docenten) moeten worden aangemoedigd om een perspectief van digitaal humanisme op te nemen in hun technische opleiding en praktijk. Ook studenten in de geesteswetenschappen en sociale wetenschappen (en hun docenten) moeten meer vertrouwd raken met de technische aspecten.

Dit zijn de voorwaarden voor een meer en beter gefundeerde inter- en zelfs transdisciplinariteit, die dringend nodig is.

Bijlage 1 – CV van de Denker

Helga Nowotny is emeritus hoogleraar wetenschaps- en technologiestudies aan de EHT in Zürich en stichtend lid en voormalig voorzitter van de Europese Onderzoeksraad (ERC).

Ze heeft onderwijs- en onderzoeksfuncties bekleed aan universiteiten en onderzoeksinstellingen in verschillende landen in Europa en is nog steeds actief betrokken bij onderzoeks- en innovatiebeleid op Europees en internationaal niveau. Momenteel is ze onder andere lid van de raad van bestuur van de Falling Walls Foundation (Berlijn); vicevoorzitter van de Lindau Ontmoetingsdagen met Nobelprijslaureaten; Senior Fellow aan de School of Transnational Governance van het EUI (Florence); lid van de raad van het IEA de Paris; lid van de Oostenrijkse Raad voor onderzoek en technologische ontwikkeling; voorzitter van de wetenschappelijke adviesraad van de Complexity Science Hub Vienna; en ze was gastprofessor aan de Nanyang Technological University (Singapore). Ze ontving meerdere eredoctoraten, onder andere van de Universiteit van Oxford en het Weizmann Institute of Science in Israël.

Ze heeft veel gepubliceerd in het domein van wetenschaps- en technologiestudies en over sociale tijd. Haar meest recente publicatie *In AI we Trust. Power, Illusion and Control of Predictive Algorithms* werd in 2021 gepubliceerd door Polity Press.

Bijlage 2 – Leden van de stuurgroep

Ine Van Hoyweghen – Coördinator / KVAB KMW / KU Leuven
Joos Vandewalle – KVAB KTW / KU Leuven
Marc De Mey – KVAB KMW / UGent
Lieven Verschaffel – KVAB KMW / KU Leuven
Johan Wagemans – KVAB KMW / KU Leuven
Luc Bonte – KVAB KNW
Luc Steels – KVAB KNW / VUB
Paul Verstraeten – KVAB KTW
Hugo De Man – KVAB KTW
Bart De Moor – KVAB KTW / KU Leuven
Anne-Mie Van Kerckhoven – KVAB KK / AMVK
Ann Dooms – Alumna JA / VUB
Inez Dua – KVAB-medewerker

RECENTE STANDPUNTEN

61. Luc Bonte, Aimé Heene, Paul Verstraeten e.a. – *Verantwoordelijk omgaan met digitalisering. Een oproep naar overheden en bedrijfsleven, waar ook de burger toe kan/moet bijdragen*, KVAB/Klasse Technische Wetenschappen, 2018.
62. Jaak Billiet, Michaël Opgenhaffen, Bart Pattyn, Peter Van Aelst – *De strijd om de waarheid. Over nepnieuws en desinformatie in de digitale mediawereld*, KVAB/Klasse Menswetenschappen, 2018.
63. Christoffels Waelkens. – *De Vlaamse Wetenschapsagenda en interdisciplinariteit. Leren leven met interdisciplinaire problemen en oplossingen*, KVAB/Klasse Natuurwetenschappen, 2019.
64. Patrick Onghena – *Repliceerbaarheid in de empirische menswetenschappen*, KVAB/Klasse Menswetenschappen, 2020.
65. Mark Eyskens – *Als een virus de mensheid gijzelt. Oorzaken en gevolgen van de Coronacrisis*, KVAB/Klasse Menswetenschappen, 2020.
66. Jan Rabaey, Rinie van Est, Peter-Paul Verbeek, Joos Vandewalle - *Maatschappelijke waarden bij digitale innovatie: wie, wat en hoe?, KVAB - Denkersprogramma 2019*, KVAB/Klasse Technische Wetenschappen, 2020.
67. Oana Dima (auteur), Dirk Inzé, Hubert Bocken, Pere Puigdomènech, René Custers (eds)., *Genoombewerking voor veredeling van landbouwgewassen. Toepassingen van CRISPR-Cas9 en aanverwante technieken*, ALLEA-KVAB/Klasse Natuurwetenschappen, 2020.
68. Marie-Claire Foblets, *De multiculturele samenleving en de democratische rechtsstaat – Hoe vrijwaren we de sociale cohesie?*, KVAB/Klasse Mensweten-schappen 2020
69. Joost Van Roost, Luc Van Nuffel, Pieter Vingerhoets e.a., *De rol van gas in de Belgische energietransitie – Aardgas en Waterstof,* KVAB/Klasse Technische Wetenschappen, 2020.
70. Richard Bardgett, Joke Van Wensem, *Bodem als natuurlijk kapitaal* – KVAB Denkersrapport 2020, KVAB/Klasse Technische Wetenschappen, 2021
71. Jos Smits e.a., *Multifunctionele eilanden in de Noordzee*, KVAB/Klasse Technische Wetenschappen, 2021.
72. Elisabeth Monard, red., *Kunst, Wetenschap en Technologie in Symbiose*, KVAB/Klasse Technische Wetenschappen, 2021.
73. Jan Wouters, Maaike De Ridder, *De problematiek van de rechtsstaat en democratische legitimiteit binnen de Europese Unie*, KVAB/Klasse Menswetenschappen, 2021.
74. Hilde Heynen, Bart Verschaffel, e.a., *Architectuurkwaliteit vandaag, Reflecties over architectuur in Vlaanderen*, KVAB/Klasse Technische wetenschappen en Klasse Kunsten, 2021.
75. Godelieve Laureys & Kristiaan Versluys e.a., *Language Matters, Taalgebruik en taalbeleid aan de Vlaamse universiteiten,* KVAB/Klasse Menswetenschappen, 2022.
76. Bea Cantillon, *Het armoedevraagstuk en de tragiek van de welvaartsstaat, Zeven termen voor een nieuw sociaal contract*, KVAB/Klasse Menswetenschappen, 2022.
77. Joos Vandewalle, Marc Acheroy e.a *Een oproep tot een versnelde digitale transformatie voor België,* ARB/KVAB, 2022.
78. Jo Tollebeek, Marc Boone en Karel van Nieuwenhuyse, *Een Canon van Vlaanderen, Motieven en bezwaren*, KVAB Klasse Menswetenschappen, 2022.
79. Luc Taerwe e.a., *Duurzaam Beheer van Infrastructuur, Niet alleen een kwestie van budgetten*, KVAB/Klasse Technische Wetenschappen, 2022.
80. Willem Salet, Marleen Spiekman, Staf Roels, Tom Coppens, Ivo Van Vaerenbergh, *Naar klimaatneutrale woongebouwen in 2050, KVAB Denkersprogramma 2022*, KVAB/Klasse Technische Wetenschappen, 2022.
81. Sabina Leonelli, Stephan Lewandowsky, *De reproduceerbaarheid van het onderzoek in Vlaanderen: Feitenonderzoek en aanbevelingen – KVAB Denkersrapport 2022*, KVAB/Klasse Technische Wetenschappen en Menswetenchappen, 2022.
82. Elisabeth Monard e.a., *Vrij onderzoek noodzakelijk voor maatschappelijke uitdagingen, Ruimte voor wetenschap op initiatief van de onderzoeker*, KVAB/Klasse Technische Wetenschappen, Klasse Menswetenschappen, Jonge Academie, 2023.
83. Herman De Dijn, Gita Deneckere, Danny Praet, Jo Tollebeek, Sabine Verhulst, *Een noodzakelijk goed., Over het blijvend belang van de geesteswetenschappen*, KVAB/Klasse Menswetenschappen, 2023.