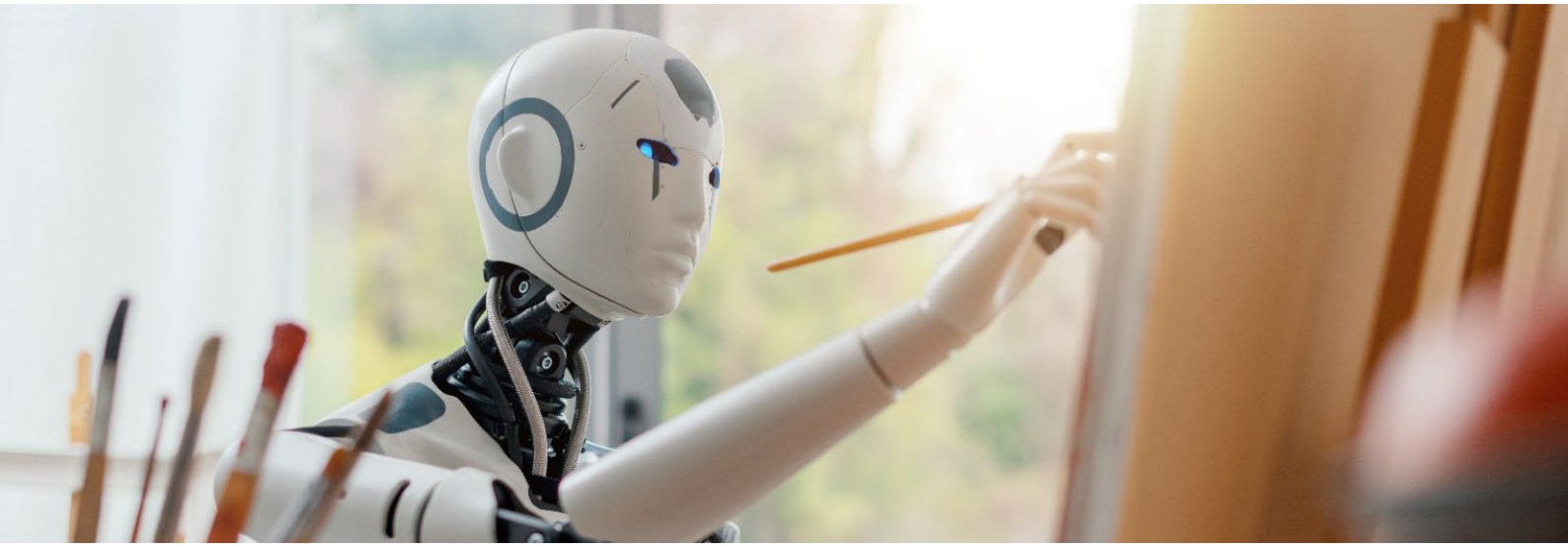
Rathenau Instituut

Generatieve AI



Rathenau Scan

Inleiding
Weinig technologieën roepen zoveel discussie op als generatieve artificiële intelligentie (GAI). Met systemen als ChatGPT en Bard is een stap gezet naar computers die tal van taken kunnen uitvoeren, maar hoe ver de kunstmatige intelligentie (artificiële intelligentie, of AI) reikt, is onduidelijk. Sommige experts denken dat computers zo krachtig worden dat ze het voortbestaan van de mens bedreigen. Andere experts vinden dit overtrokken, of wijzen op de risico's op korte termijn, zoals vooroordelen en incorrecte output.

Deze scan maakt de balans op: wat is GAI, wat kan het nu, en wat misschien later? Welke kansen, risico's voor publieke waarden en beleidsopties zijn ermee verbonden? De scan is ontwikkeld op verzoek van het ministerie van Binnenlandse Zaken en Koninkrijksrelaties, gebaseerd op kortlopend onderzoek met literatuurstudie, werksessies en interviews, en is bedoeld voor beleidsmakers en politici.

Inhoud

Samenvatting

Waarom een Rathenau Scan over generatieve AI?
De naam generatieve AI (GAI) verwijst naar AI-systemen die geautomatiseerd content kunnen maken, op verzoek van een gebruiker. Zo kan je een systeem vragen een samenvatting te maken, of een foto te creëren in de stijl van de schilder Van Gogh. Sinds de lancering van ChatGPT in november 2022 experimenteren miljoenen gebruikers wereldwijd met deze technologie. De technologie heeft al impact op de samenleving, en de verwachtingen van wat het de maatschappij gaat brengen zijn hooggespannen. Deze scan biedt een overzicht van de kansen, risico's en handelingsopties verbonden met GAI.

Is generatieve AI nieuw?
Generatieve AI bouwt voort op bestaande AI-technieken en vormt een subgroep van lerende AI-systemen. Tegelijkertijd hebben generatieve AI-systemen een aantal onderscheidende eigenschappen:

- ten eerste zijn generatieve AI-systemen aanzienlijk beter in taal dan andere AI-systemen;
- ten tweede kunnen de systemen goed met verschillende 'modaliteiten' werken, zoals beeld, geluid, video, spraak, en zelfs zaken als eiwitstructuren en chemische verbindingen;
- ten derde krijgen generatieve AI-systemen een algemene training, die de basis biedt voor allerhande specifieke toepassingen.

Om deze redenen kunnen GAI-systemen veel verschillende taken uitvoeren, in tegenstelling tot veel andere AI-systemen die vallen onder de noemer '*narrow AI*', en getraind zijn voor één specifieke taak.

Wat kun je met generatieve AI?
In deze scan onderscheiden we vier rollen die GAI-systemen kunnen vervullen. Een GAI-systeem is in te zetten als:

1. leerinstrument: bijvoorbeeld om informatie op te zoeken, of te fungeren als vraagbaak bij het maken van huiswerk;
2. productietool: het systeem maakt iets in opdracht van een gebruiker. Op de werkvloer experimenteren velen hier al mee.
3. complexe probleemoplosser: bijvoorbeeld in de wetenschap, waar GAI-systemen helpen bij het vouwen van eiwitstructuren, o.a. met het oog op de ontwikkeling van nieuwe medicijnen;
4. het creëren van een ervaring: sommige gebruikers vinden het fijn of fascinerend om te communiceren met GAI-systemen. GAI-systemen fungeren dan bijvoorbeeld als metgezel. Zo maakte iemand al een chatbot die zijn overleden geliefde imiteerde.

Ondanks deze mogelijkheden, kent de technologie de nodige beperkingen. Generatieve AI-systemen berekenen het meest waarschijnlijke antwoord. Dat kan leiden tot onjuiste antwoorden of discriminerende content. Ook zijn de onderliggende algoritmes dusdanig

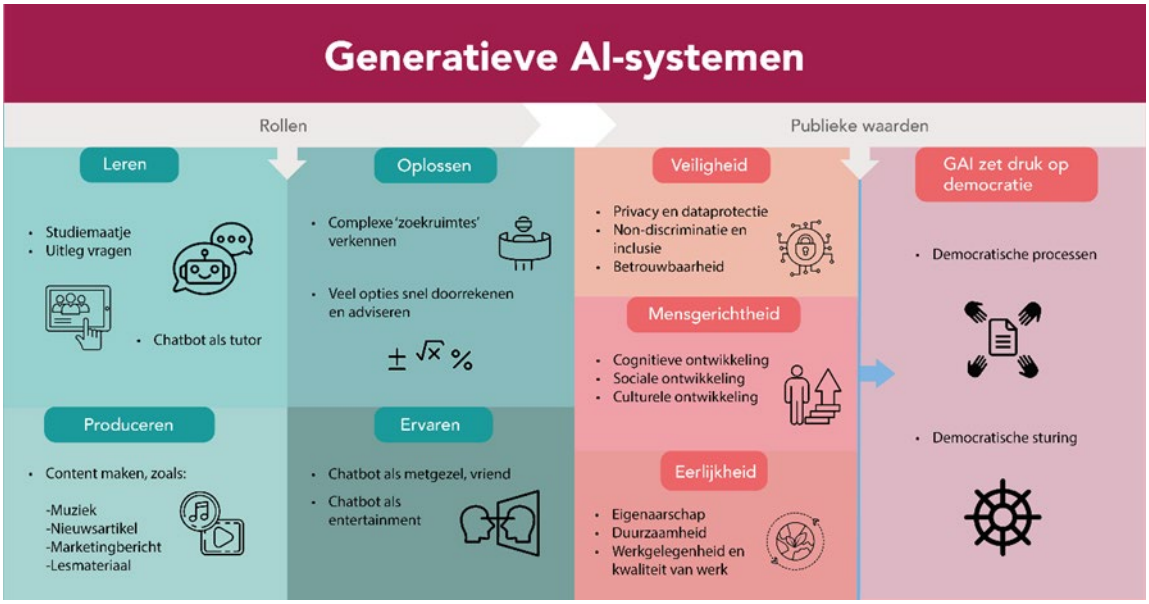
complex dat de werking van de systemen voor mensen slechts beperkt begrijpelijk is – ook voor de ontwikkelaars. Daardoor is de technologie voorlopig nog niet goed genoeg om in te zetten in belangrijke processen, zoals bij het stellen van medische diagnoses.

Wat staat er bij de opkomst van GAI op het spel?
Generatieve AI brengt veel risico's met zich mee die publieke waarden onder druk kunnen zetten. We clusteren de risico's in deze scan in drie thema's. Ten eerste zijn er zorgen over de **veiligheid** van GAI-systemen: ze kunnen de privacy van gebruikers schenden, vooroordelen uiten en foute informatie geven. Bovendien zijn de systemen dermate complex, dat ontwikkelaars en externe partijen de werking niet geheel kunnen doorgronden, wat het lastig maakt om nu en in de toekomst risico's te voorkomen.

Ten tweede is het de vraag hoe **mensgericht** de systemen zijn: wat gaan de systemen betekenen voor onze cognitieve, sociale en culturele ontwikkeling. Stimuleren chatbots onze creativiteit? Ontleren we sociale omgangsvormen, als we vaker interacteren met een GAI-systeem? Verwerken we ons verdriet via chatbots die onze overleden geliefde imiteren? Kortom: wat betekent het om mens te zijn in een wereld van robots?

Ten derde leven er vragen over hoe **eerlijk** de systemen zijn: wie profiteert van deze systemen, en wie draagt de lasten? Hoe beschermen we het werk van creatieve beroepen? Welke banen gaan veranderen en hoe zorgen we voor goed werk? En hoe gaan we om met de impact op het milieu?

Tot slot signaleren we een rode draad: de invloed van GAI op onze **democratie**. GAI kan democratische processen bemoeilijken, zoals het publieke debat en politieke besluitvorming, en vanwege de toenemende machtspositie van enkele techbedrijven het vermogen aantasten om in tal van maatschappelijke domeinen digitale technologie democratisch te sturen.



Wat moet er nu gebeuren?
In deze scan concludeert het Rathenau Instituut dat generatieve AI risico's in de digitale samenleving versterkt en nieuwe risico's introduceert. Beleidsmakers op nationaal, Europees en internationaal niveau hebben zich de afgelopen jaren ingespannen om AI in goede banen te leiden, en veel wordt verwacht van de aankomende Europese AI-verordening. Het is echter onduidelijk hoe de abstracte normen in deze wet voor het naleven van mensenrechten vorm gaan krijgen in de praktijk. Wanneer is bijvoorbeeld het risico op discriminatie tot een acceptabel niveau teruggebracht? En voor wie is dat acceptabel? Ook voor andere juridische kaders en beleid speelt de vraag of zij risico's van generatieve AI voldoende ondervangen.

De hamvraag is dus: zijn de gedane inspanningen voldoende? Het is een reële mogelijkheid dat het huidige en voorgenomen beleid niet opgewassen zijn tegen de impact van generatieve AI-systemen, bijvoorbeeld op het gebied van non-discriminatie, veiligheid, desinformatie, mededinging en de uitbuiting van werknemers. Het is daarom zaak dat het kabinet een strategie uitwerkt om de grip van de samenleving op deze technologie te verbeteren. Dat begint met het evalueren van het Nederlands en Europees beleid en toetsen waar dat aanscherping behoeft. Ook is het zaak toezicht maximaal te ondersteunen, afspraken te maken met ontwikkelaars en de samenleving te waarschuwen voor de risico's van GAI. Wereldwijd worden de risico's van GAI-systemen serieus genomen; dat zou iedere Nederlandse burger en instantie ook moeten doen.

Het Rathenau Instituut geeft het kabinet vijf handelingsopties mee om haar strategie vorm te geven:
1. Creëer het vermogen om schadelijke GAI-toepassingen van de markt te halen;
2. Zorg voor toekomstbestendige juridische kaders;
3. Investeer in internationaal AI-beleid, om mondiale innovatieprocessen van technologiebedrijven bij te sturen;
4. Stel een ambitieuze agenda op voor maatschappelijk verantwoorde GAI;
5. Stimuleer maatschappelijk debat over de wenselijkheid van GAI.

1. Wat is generatieve AI?

1.1. Introductie

De naam generatieve AI verwijst naar artificiële-intelligentiesystemen die geautomatiseerd content kunnen maken, op verzoek van een gebruiker. Denk hierbij aan teksten zoals een sollicitatiebrief, programmeercode of een schoolopstel, beelden zoals schilderijen of foto's, video's en geluiden zoals de stemmen van verschillende personen. De gebruiker hoeft geen programmeercode te beheersen, en kan via menselijke taal (*prompts*) interacteren met het generatieve systeem. De ontwikkeling van generatieve AI-systemen is in een stroomversnelling geraakt, omdat een nieuwe algoritmische methode het mogelijk maakte veel efficiënter data te verwerken en daarbij met veel meer variabelen te rekenen. Dit heeft geleid tot complexe algoritmische modellen die indrukwekkende taaltaken kunnen uitvoeren, ook wel grote taalmodellen of *large language models* genoemd.[1] Hieronder leggen we deze technologische doorbraak uit (zie paragraaf 1.2). Deze scan richt zich op generatieve AI-systemen (hierna GAI-systemen), omdat ze een veelvoud van taken kunnen uitvoeren en daarmee een significante impact op de samenleving zullen hebben.

Het meest bekende GAI-systeem is de chatbot Chat-GPT, ontworpen door het Amerikaanse bedrijf OpenAI. Snel na de lancering van Chat-GPT3.5, in november 2022, werd het al ingezet door miljoenen gebruikers wereldwijd. Maar er zijn meer generatieve systemen, zoals de beeldengenerator DALL-E (ook van OpenAI), de Co-Pilot van Microsoft en de chatbot Bard van Google.

1.2. Hoe werkt het?

Zoals gezegd is GAI gebaseerd op grote taalmodellen, *large language models* (LLM). Deze softwaremodellen berekenen welk woord het meest waarschijnlijke volgende woord is in een zin, zoals: 'De paus gelooft in …' [God]. Ook het meest waarschijnlijke ontbrekende woord kan uitgerekend worden: 'Ik eet een ….. met kaas' [boterham]. Nieuwe modellen kunnen niet alleen woorden voorspellen, maar ook geluidsfragmenten in een muziekreeks, of beelden in een samenstelling van pixels. Deze modellen kunnen dus omgaan met meerdere 'modaliteiten', en kunnen tekst, beeld en geluid ook combineren. Ze worden daarom ook wel *large multimodal models* (LMM's) genoemd. ChatGPT4 kan zowel tekst als beelden genereren.

Om in beeld, taal en geluid patronen te ontdekken en te voorspellen moeten de modellen uitvoerig getraind worden. Dat gebeurt met bestaande *machine learning* technieken, zoals neurale netwerken.[2] Daarbinnen is een nieuwe ontwikkeling opgekomen: bepaalde algoritmische modellen, genaamd *transformers*, zorgen dat veel

---

1 Met taaltaken worden taken op het gebied van taal bedoeld, zoals samenvatten, vraag en antwoord geven, vertalen e.d.

2 Generatieve AI kan worden beschouwd als een subgroep binnen AI, en specifiek als een subgroep binnen *machine learning* en *deep learning* AI-technieken. AI kan worden omschreven als een systeem dat intelligent gedrag vertoont, door de omgeving te analyseren en acties te ondernemen – met enige mate van autonomie – om specifieke doelen te bereiken, zie European Commission High Level Expert Group on AI, 2019.

meer tekst geanalyseerd kan worden. Dit helpt om de context van een woord of zin beter te verwerken: een losse zin krijgt immers pas betekenis als je ook de gehele pagina hebt gelezen.[3] Deze nieuwe modellen blijken een cruciale stap ten opzichte van eerdere AI-systemen.[4]

De ontwikkeling van een GAI-systeem kent grofweg vijf fases (zie figuur 1).

1. Dataverzameling

Allereerst worden er grote hoeveelheden trainingsdata verzameld. Denk hierbij aan de onvoorstelbare hoeveelheid teksten, foto's en video's die online staan, gesprekken op sociale media en alle boeken en wetenschappelijke literatuur die je kan digitaliseren. Veel van die data is publiekelijk toegankelijk, maar het kan ook besloten datasets betreffen die eigendom zijn van een partij en worden opgekocht.[5]

2. Datacuratie

Vervolgens kan de data gefilterd worden, bijvoorbeeld door data te anonimiseren, dubbelingen uit de verzameling te halen of specifieke woorden te verwijderen. Hierbij spelen algoritmes en mensenwerk beide een rol.

3. Training

In de derde fase vindt de training plaats, ook wel pre-training genoemd.[6] Het taalmodel probeert in de omvangrijke database veel verschillende soorten patronen te ontdekken. Hierbij gebruikt het miljarden parameters, ofwel variabelen die verschillende waardes kunnen hebben.[7] Zowel de datasets als het aantal parameters zijn enorm groot: GPT3 maakte gebruik van 570 GB aan data en 175 miljard parameters.[8] Het kan nog groter, geschat wordt dat GPT4 getraind werd met 1,75 biljoen tot een triljoen parameters.

4. Aanpassing

Na deze training wordt het model in de vierde fase verder aangepast en verfijnd, om geschikt te zijn voor bepaalde taken. Zo kan een GAI-systeem gevoed worden met medische termen en literatuur om ingezet te worden in het medisch domein. Het systeem kan ook specifieke training ondergaan om de kans te verkleinen op racistische of kwetsende uitlatingen, zoals bij GPT4 is gebeurd (ook wel *alignment* technieken genoemd). Hierbij worden mensen ingezet om de output van de modellen te beoordelen: is dit haatzaaiende, beledigende, discriminerende of anderszins illegale of

---

3 De transformer doet nog meer: er komt een 'nieuw' aandachtsmechanisme dat aangeeft welke contextinformatie belangrijk is. Input kan bovendien parallel worden verwerkt in plaats van woord voor woord – een belangrijke beperking van de neurale netwerken die tot dan toe voor taal gebruikt werden. Zie Vaswani et al., 2017.

4 Brown et al., 2020; Kaplan et al., 2020

5 De huidige trend is dat vrij beschikbare internetgegevens afnemen. Dit komt onder meer doordat diverse platforms hun data meer afschermen. Zo hebben Reddit en X (voorheen Twitter) maatregelen genomen om te voorkomen dat gegevens van hun platforms zijn te ‘schrapen’. Daarnaast kan de relatieve waarde van internetgegevens afnemen naarmate het internet meer AI-gegenereerde inhoud begint te bevatten. Zie o.a. Vipra & Myers West, 2023

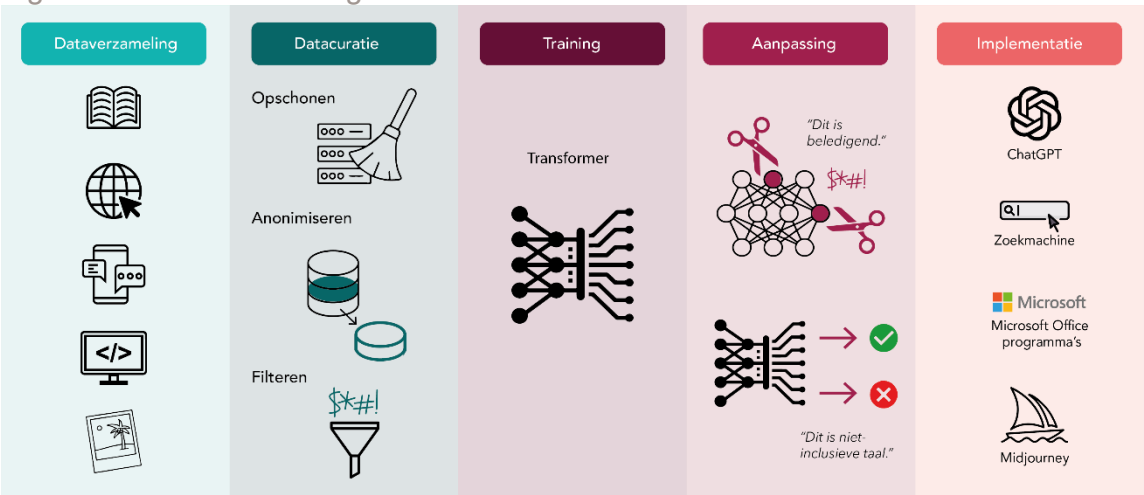
6 Nu deze termen zijn uitgelegd, wordt de afkorting van ChatGPT inzichtelijk: Generative Pre-Trained Transformer

7 Parameters verwijzen naar het aantal variabelen waarmee het model getraind is, bijvoorbeeld de gewichten in neurale netwerken of co-efficiënten in lineaire regressie.

8 OpenAI heeft verschillende varianten van GPT3 uitgebracht, ook wel de GPT3 ‘familie’ genoemd. We verwijzen hier naar het grootste model in de GPT3 familie.

ongewenste content? Ook kan een systeem met instructies verder verfijnd worden (*prompt engineering*). Een programmeur kan bijvoorbeeld met redeneerstappen laten zien hoe een model op het gewenste antwoord kan komen. Eigenlijk wordt dus een specifiekere toepassing gebouwd op een algemene fundering – daarom worden grote taalmodellen ook wel *foundation models* genoemd.[9] Deze werkwijze verschilt significant van '*narrow AI*'-systemen, die op basis van een specifieke dataset getraind worden voor een specifieke taak.

Figuur 1: Ontwikkelfases generatieve AI[10]



5. Implementatie

Ten slotte wordt het model toegepast in de praktijk. Die toepassing kan verschillende vormen aannemen: als een chatbot, een beeldengenerator (Midjourney) of verwerkt in een zoekmachine. In hoofdstuk 2 staan we stil bij mogelijke toepassingen. In het geval van de chatbot kan de gebruiker wederom met *prompts* verschillende resultaten genereren, en een productieve interactie aangaan met de generatieve AI. Zo kan je met een chatbot brainstormen over de titel van een boek, waarbij de AI op basis van bepaalde instructies mogelijkheden genereert. Het antwoord hangt mede af van hoe de instructie gegeven wordt. De gebruiker kan ook met redeneerstappen proberen het AI-systeem iets te leren om het juiste antwoord te krijgen.[11]

1.3. Wie maakt het?

Voortbouwend op hun AI-kennis en producten, hebben sinds 2018 meerdere technologiereuzen grote taalmodellen ontwikkeld, waaronder OpenAI, Google, Meta, Microsoft in Amerika en Baidu in China. Ze hebben niet alleen modellen gemaakt gericht op taal, maar ook voor andere modaliteiten, zoals CodeX (OpenAI), eiwitstructuren (zoals AlphaFold van Deepmind) en robotica (zoals PaLM-E van

9 Bommasani et al., 2022
10 Figuur aangepast op basis van Bandi et al., 2023; Bommasani et al., 2022; Zhao et al., 2023.
11 Zo kan je een rekenvraag stellen die een chatbot niet goed weet te beantwoorden. Door bij de rekenvraag ook het stappenplan mee te geven voor hoe de som kan worden opgelost, leert de chatbot hoe het in het vervolg tot het juiste antwoord kan komen.

Google).[12] De techbedrijven beschikken vaak of zelf, of door samenwerkingen, over zowel de benodigde infrastructuur (*computing resources*), als de data om de modellen te trainen, de modellen, en de softwareprogramma's van eindgebruikers. Zo heeft Microsoft miljarden geïnvesteerd in OpenAI, levert Microsoft samen met NVIDIA de supercomputers voor OpenAI, en heeft het aangekondigd GPT in haar kantoorsoftware te integreren.

Omdat de training van die modellen veel rekenkracht, hardware en data vergt, zijn met name grote techbedrijven tot nu toe in staat geweest om een taalmodel te trainen. Afhankelijk van de grootte van het model duurt dat enkele dagen tot maanden. Die computerkracht kost geld. De schattingen lopen uiteen, en zijn erg afhankelijk van het precieze model en de gebruikte hardware. Als iemand van de grond af een taalmodel zou willen trainen, met vergelijkbare computerkracht als de modellen van de grote techbedrijven, lopen de kosten al snel richting 100 miljoen U.S. dollar.[13] Als de hardware nog verder wordt opgeschaald, kunnen deze kosten aanzienlijk hoger zijn. Bovenop de training komen de operationele kosten. Zo kost het in de lucht houden van ChatGPT naar schatting 700.000 U.S. dollar per dag.[14] Dit soort rekenkracht vergt ook veel energie- en waterverbruik. Wetenschappers en bedrijven werken aan efficiëntere manieren om generatieve AI-systemen te ontwikkelen (zie hoofdstuk 3).

Verschillende systemen kennen varianten van openheid, waardoor bedrijven, wetenschappers een particuliere eindgebruikers zelf aan de slag kunnen met de taalmodellen van anderen. Dit kan bijvoorbeeld via een licentie op de broncode.[15] In de zomer van 2023 bracht Meta Llama 2 uit, waarvan de broncode onder voorwaarden beperkt toegankelijk is. Zodoende zijn er startups ontstaan die op basis met bestaande generatieve AI-systemen specifieke diensten leveren, zoals Jasper en Grammarly. Er is echter pas echt sprake van open ontwikkeling als de broncode, datasets en andere trainingsinformatie door iedereen ingezien kan worden. Voorlopig bestaat er één succesvol open systeem, BLOOM, gemaakt door een collectief van wetenschappers en ontwikkelaars.

---

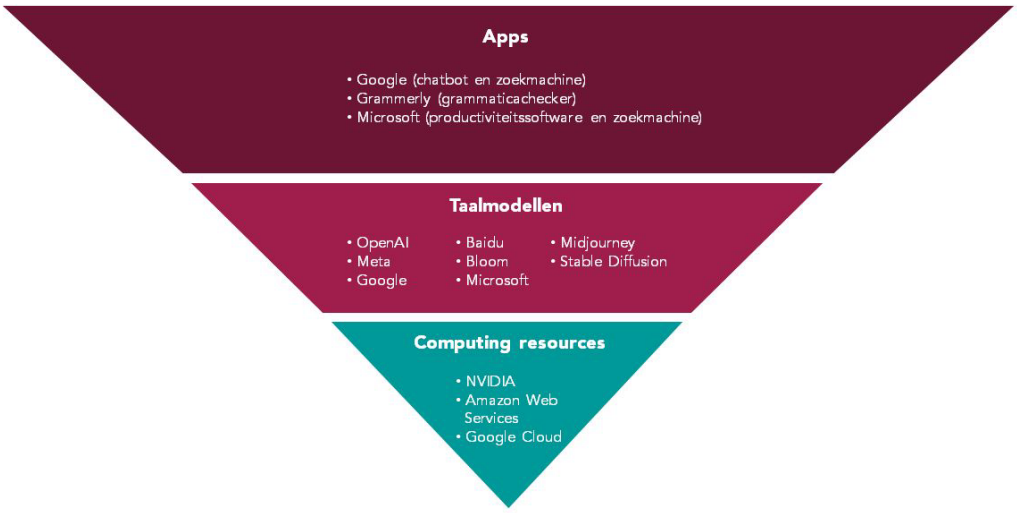
12 De grote techbedrijven hebben elk meerdere modellen en modaliteiten ontwikkeld – we noemen hier slechts enkele voorbeelden. De modellen werken op een vergelijkbare manier, in alles wat kan worden uitgedrukt in een rij van tekens kunnen patronen worden ontdekt. Aan de hand van die patronen kan je voorspellingen doen over mogelijke combinaties van tekens.

13 Simon, Julien, n.d.

14 Mok, 2023

15 Voor een overzicht van gradaties in de mate van openheid, zie Solaiman, 2023.

Figuur 2: Vereenvoudigd overzicht van technologie-onderdelen en actoren



Bron: Rathenau Instituut

1.4. Hoe goed is generatieve AI?

De resultaten van generatieve AI zijn indrukwekkend, en een enorme stap voorwaarts in het creëren van kwalitatief goede teksten. Ten opzichte van oudere AI-methodes kunnen uiteenlopende taken veel beter uitgevoerd worden.[16] Zo kunnen de nieuwste modellen, zoals Bard van Google, veel inhoudelijker en correcter vragen beantwoorden dan bijvoorbeeld de AI-spraakassistenten Alexa en Google Assistant. Het herkent taal beter, en het kan taken uitvoeren die voorheen niet geautomatiseerd konden worden, zoals het herschrijven van een tekst in de stijl van een bekende auteur, of het uitleggen van een natuurkundig principe. Maar het meest vernieuwende is de hoeveelheid verschillende taken die generatieve AI kan uitvoeren. Als het funderende model eenmaal is ontwikkeld, kan het met relatief weinig moeite toegespitst worden op een specifieke taak. Bovendien zijn de systemen te koppelen aan andere programma's, zoals zoekmachines, beurswaardes en websites, waarmee de systemen over actuele informatie beschikken en nog meer taken kunnen uitvoeren, zoals het boeken van een vliegticket of (ver)kopen van aandelen.[17] Ontwikkelaars integreren steeds meer functionaliteiten in hun diensten, zoals Meta AI in Whatsapp of stem en spraak in ChatGPT.[18]

Niettegenstaande de indrukwekkende prestaties, kennen de huidige generatieve AI-systemen ook beperkingen, die te maken hebben met de onderliggende eigenschappen van deze systemen.

16 Zie voor een overzicht van de voortgang op verschillende taken, en een vergelijking met hoe mensen (zowel experts als leken) scoren: Rudolph et al., 2023

17 Lorenz et al., 2023; Zhao et al., 2023. Eerdere modellen, waaronder GPT3, waren dit niet, en konden geen antwoorden geven over gebeurtenissen die na de trainingsfase plaats hadden gevonden. GPT3 is getraind op data uit de periode tot eind 2021.

18 Meta, 2023; OpenAI, 2023

Generatieve AI is gebaseerd op data en statistiek

Generatieve AI is in de kern gebaseerd op data en statistiek, oftewel kansberekening en waarschijnlijkheid. Bij het creëren van output doen de systemen in feite een 'geïnformeerde gok'. Het blijkt inmiddels dat deze statistiek kan leiden tot 'verzonnen' antwoorden, waarbij het systeem een gebeurtenis vermeldt die niet heeft plaatsgevonden (ook wel 'hallucineren' genoemd). Zo stelde Google's Bard bij de introductie in Nederland dat CDA-minister Hugo de Jonge in de race zou zijn voor het lijsttrekkerschap van de VVD.[19] Systemen gebaseerd op data en statistiek kunnen ook vooroordelen bevatten, die kunnen leiden tot discriminerende content. Gebruikers en ontwikkelaars rapporteren dagelijks fouten in de systemen. De ontwikkelaars proberen dit soort fouten te voorkomen door datasets te cureren, van menselijke feedback gebruik te maken en andere zogeheten *alignment* technieken toe te passen.[20] Maar deze datacuratie biedt tot dusver niet de garantie dat er geen fout meer gemaakt wordt.

Bovendien is er nog een ander risico: als een systeem in nieuwe trainingsrondes getraind wordt met de foutieve data die eerder is gegenereerd, kan de werking van het model verslechteren, waardoor het systeem in een neerwaartse spiraal terecht kan komen van foutieve data en gemankeerde training.[21] Dit wordt *model collapse* genoemd. Synthetische data kunnen hierin een rol spelen. Normaal gesproken verwijzen data naar iets in de werkelijkheid, zoals iemands naam of geboorteplaats. Maar GAI-systemen kunnen ook data creëren, ofwel synthetiseren, die niet naar specifieke personen of gebeurtenissen verwijzen. Als GAI vervolgens weer traint aan de hand van deze data, kan het beeld van de werkelijkheid verder verslechteren.

Er is veel discussie over de vermogens van generatieve AI-systemen om in hun berekeningen de wereld correct weer te geven, en de verschillen tussen menselijk begrip en de werking van algoritmes. Mensen denken in concepten, zoals 'een vliegtuig' of 'een boterham', en kunnen oorzakelijke verbanden doorgronden en in nieuwe situaties toepassen. Het is de vraag in hoeverre een generatief AI-systeem concepten en oorzaken überhaupt statistisch kan representeren.[22] Sommige experts stellen dat dit probleem onoplosbaar is, maar niet iedereen is het daarmee eens. De systemen kunnen dingen doen die vroeger voor onmogelijk werden gehouden, en regelmatig vinden er doorbraken plaats die nog niet goed door onderzoekers begrepen worden. Voorlopig is er te weinig bekend over wat de modellen precies wel en niet kunnen. Bestaande methoden zijn ongeschikt om de vaardigheden van generatieve AI-systemen, zoals abstract redeneren, te toetsen. Hier wordt ook onderzoek naar gedaan.[23]

---

19 Quekel & Hoijtink, 2023

20 Het bedrijf Anthropic maakt AI-systemen die op basis een normenkader, 'een grondwet', zelf feedback geven op de resultaten van GAI-systemen, dit noemt het bedrijf constitutional AI. Zie Anthropic, 2022; Bai et al., 2022

21 Wong, 2023

22 Bender et al., 2021; Bender & Koller, 2020; Floridi, 2023; Gurnee & Tegmark, 2023

23 Bandi et al., 2023; Kaddour et al., 2023; Zhao et al., 2023

De taalmodellen zijn complex en daardoor moeilijk te interpreteren

Ontwikkelaars en onderzoekers hebben beperkt begrip van precieze algoritmische werking van taalmodellen.[24] Dat betekent dat de prestaties en resultaten van de modellen beperkt voorspelbaar zijn, wat kan leiden tot risico's voor betrouwbaarheid en veiligheid. Het feit dat ontwikkelaars en onderzoekers zelf maar beperkt begrip hebben van de taalmodellen, betekent ook dat zij eerdere aannames over de eigenschappen van de modellen soms moeten herzien. Stanfordonderzoekers betwijfelden dit voorjaar dat LLM's ineens iets nieuws kunnen leren, zoals voorheen werd gedacht.[25] Door andere meetmethoden werd duidelijk dat de modellen niet plotseling, maar juist stap voor stap leerden. Een saillante ontdekking, omdat de hooggespannen verwachtingen over de kansen en risico's van generatieve AI samenhangen met het idee dat de modellen zonder (weinig) extra training nieuwe taken kunnen leren. Veel wetenschappers vinden daarom dit soort onderzoek naar de onderliggende werking van de modellen essentieel.

---

24 Bandi et al., 2023; Bommasani et al., 2022; Bowman, 2023

25 Deze eigenschap wordt ook wel 'emergentie' genoemd. Een algoritme maakt zich ineens een vaardigheid eigen die het eerder nog niet beheerste. Zie Schaeffer et al., 2023

Kader 1 **Artificial general intelligence?**

De ontwikkeling van generatieve AI wordt door sommige experts gezien als stap in de richting van *artificial general intelligence* (AGI): systemen die zo goed zo veel taken kunnen uitvoeren, en zo gestructureerd redeneren, dat ze een zeer hoge mate van autonomie bezitten. Het gevaar hiervan kan zijn dat deze systemen op grote schaal handelingen gaan uitvoeren op basis van doelen die ontwikkelaars en gebruikers niet voor ogen hebben: het *alignment* probleem. Daarom waarschuwen diverse techbedrijven en experts voor toekomstige, existentiële risico's van generatieve AI voor de mensheid. Andere experts uiten kritiek op deze zienswijze. Zij zien de ontwikkeling richting AGI als speculatief en vinden onder meer dat deze zienswijze afleidt van al aanwezige risico's van generatieve AI: het geven van onbetrouwbare antwoorden, verspreiding van desinformatie, vooroordelen en gebrek aan transparantie.

Ongeacht de vraag wat de modellen kunnen of nog gaan kunnen, is duidelijk dat er geen sprake meer is van zogenaamde '*narrow AI*': een AI-systeem dat getraind wordt voor – en goed is in – slechts één specifieke taak. De brede variëteit aan taken die GAI-systemen kunnen uitvoeren zal naar verwachting alleen maar toenemen. Dat sorteert nu al impact in economie en maatschappij – positief én negatief – en de waaier aan kwesties waar de samenleving mee te maken krijgt, groeit daarmee ook.

1.5. Conclusie

Ondanks de beperkingen leveren generatieve AI-systemen indrukwekkende prestaties, en de volledige potentie van de technologie is nog niet benut. De systemen zullen waarschijnlijk op basis van nog complexere statistische patroonherkenning en groei in rekenkracht beter worden in diverse taken, meer modaliteiten als taal, tekst en video kunnen combineren en vaker gekoppeld worden aan andere systemen en programma's. Dit maakt het dus ook lastiger af te bakenen waar generatieve AI ophoudt of begint. Het is een open vraag wat AI-systemen in de toekomst precies wel en niet kunnen, zeker als je bedenkt dat tijdens de ontwikkeling van de systemen menselijke beoordeling van de data en de output onmisbaar is. Het is niet zonder reden dat de verwachtingen hooggespannen zijn, maar de technologie kent ook serieuze problemen die de vooruitgang kunnen frustreren. In het volgende hoofdstuk gaan we in op de toepassingen en kansen die in verschillende maatschappelijke domeinen verwacht kunnen worden.

2. Wat wordt ervan verwacht?

2.1. Introductie

Dit hoofdstuk kijkt naar de manieren waarop de technologie in maatschappelijke domeinen benut kan worden. In veel domeinen experimenteren wetenschappers, bedrijven, werknemers en particulieren met GAI-systemen. We lichten voor deze scan enkele domeinen uit: onderwijs en wetenschap, defensie en cybersecurity, arbeidsmarkt en gezondheidszorg. Onze inventarisatie is deels gebaseerd op voorbeelden uit de praktijk, maar bestaat met name uit onderbouwde inschattingen van wetenschappers, bedrijven en journalisten. Vanwege de recente ontwikkelingen bestaan er nog relatief weinig empirische wetenschappelijke studies die toepassingen in de praktijk bestuderen. De meeste artikelen proberen kansen te voorspellen of ethische kwesties van mogelijke toepassingen te beoordelen.[26] De nadruk ligt in dit hoofdstuk op de kansen van GAI, de maatschappelijke en ethische risico's bespreken we in hoofdstuk 3. Wel noemen we hier mogelijke twijfels over de effectiviteit van GAI. We sluiten dit hoofdstuk af met vier rollen die GAI kan vervullen: als leerinstrument, productietool, probleemoplosser en als bron van ervaringen, zoals verwondering en sociale interactie.

2.2. Kansen in maatschappelijke domeinen

Onderwijs en wetenschap

In de wetenschappelijke literatuur worden tientallen taken genoemd die GAI in het onderwijs kan vervullen, zowel voor leerlingen en studenten, als voor docenten. Voorbeelden zijn samenvattingen maken, lesmateriaal schrijven, lessen inplannen en het op maat maken van lesstof, bijvoorbeeld voor een specifieke leerstijl of beperking. Verder kunnen de systemen helpen bij het beoordelen van het werk van leerlingen en studenten, of toetsen opstellen. Een chatbot kan ook bij het leren ondersteuning bieden als real-time vraagbaak, studiemaatje of bron van ideeën en suggesties.[27] De prestaties van generatieve AI-systemen zijn dusdanig dat ze voor diverse schoolvakken of universitaire vakken kunnen slagen. Zo haalde ChatGPT een voldoende voor een examen rechten en een examen geneeskunde.[28] Multimodale GAI-modellen kunnen deze ervaring en hulp in de toekomst verder verrijken.

Veelgenoemde kansen voor het gebruik van generatieve AI in het onderwijs zijn tijdwinst en efficiënte, hogere kwaliteit van de lesstof, betere leeropbrengst en het bevorderen van de motivatie van leerlingen. Doordat chatbots materiaal snel kunnen personaliseren, bijvoorbeeld in toegankelijke tekst of een andere taal, bieden de bots ook kansen om inclusie te bevorderen. Tegelijkertijd zijn er twijfels. Zo wijzen studies op de beperkte betrouwbaarheid van taalmodellen, waardoor docenten en leerlingen niet helemaal kunnen vertrouwen op de output, en tijd kwijt zijn met controles. Ook zijn GAI-

---

26 Sohail et al., 2023

27 Farrokhnia et al., 2023; Jeon & Lee, 2023; Lo, 2023; Sabzalieva & Valentini, 2023

28 Choi et al., 2023; Kung et al., 2023; Lo, 2023. De taalmodellen zijn niet in ieder vakgebied even goed, zie bijvoorbeeld Lo, 2023.

systemen niet didactisch getraind, en zal niet iedereen met de systemen overweg kunnen.[29]

Naast onderwijs biedt GAI ook kansen in de wetenschap.[30] Studenten en wetenschappers kunnen met GAI literatuuronderzoek doen en ideeën aangereikt krijgen.[31] In diverse wetenschappelijke tijdschriften zijn al artikelen verschenen waarin ChatGPT als medeauteur is opgevoerd. Meta ontwikkelde in 2022 een speciaal model om wetenschappers te helpen: Galactica. Het systeem werd onder meer getraind op wetenschappelijke artikelen, encyclopedieën, wetenschappelijke boeken en online lesmateriaal. Het moest artikelen kunnen samenvatten, wiskundige problemen helpen oplossen en wetenschappelijke teksten kunnen schrijven. Meta moest Galactica echter binnen enkele dagen offline halen, omdat het foute resultaten en vooroordelen vertoonde.

Er bestaan ook hoge verwachtingen over het vermogen van GAI-systemen om complexe zoekproblemen op te lossen.[32] Deze modellen worden op specifieke data getraind, onder meer op medische beelden en teksten, eiwitstructuren en wiskundige problemen.[33] Zo bestaat de zogeheten ‘zoekruimte’ voor het zoeken naar medicijnen uit ongeveer 10^{23} tot 10^{63} molecuulsamenstellingen die verkend kunnen worden.[34] Medische AI-modellen, zoals Alphafold van Deepmind, kunnen gebruikt worden om de zoekruimte veel sneller door te lopen. In de chemie is de hoop dat modellen kunnen helpen bij het zoeken naar synthetische moleculen en materialen. Hier is echter wel voldoende hoogwaardige data voor nodig.[35]

Defensie en cybersecurity

Toepassingen van generatieve AI-systemen worden in tenminste twee veiligheidsdomeinen onderzocht: defensie en cybersecurity. Op het terrein van defensie kondigde het Amerikaanse ministerie van Defensie een *Generative AI Taskforce* aan, die onder meer zou kijken naar toepassingen voor inlichtingenverzameling en om administratieve processen te verbeteren.[36] Toepassing op het slagveld ligt een stuk verder weg, maar er wordt al wel over nagedacht. Zo zouden GAI-systemen ontwikkeld kunnen worden die de strategische beraadslaging versterken.[37] Denk aan een GAI die uitrekent welk aanvalsplan succesvol zou kunnen zijn, of welk aanvalsplan gepaard gaat met te hoge risico's. Ook beraadslaging omtrent de interpretatie van oorlogsrecht in specifieke situaties zou ondersteund kunnen worden door GAI. Ten slotte kan je ook denken aan GAI-systemen die advies geven op het gebied van logistiek.

---

29 Blodgett & Madaio, 2021; Jeon & Lee, 2023; Lodge et al., 2023; Malinka et al., 2023; Rahman & Watanobe, 2023
30 Harrer, 2023; Janssen et al., 2023; Marr, 2023; Van Buchem et al., 2021
31 Grünebaum et al., 2023; Harrer, 2023; Koncz, 2023; Kung et al., 2023; Qi et al., 2023
32 Acosta et al., 2022; Lang et al., 2023; Nógrádi et al., 2023
33 Callaway, 2022; Vogt, 2023
34 Wang et al., 2023. Insilico Medicine gebruikt GAI bijvoorbeeld voor ‘target identificatie’, wat een (potentieel) medicijn voor idiopathische longfibrose heeft opgeleverd. Zie hiervoor Field, 2023; Philippidis, 2023.
35 Nature editorial, 2023
36 U.S. Department of Defense, 2023
37 Baughman, 2023

De militaire toepassing van GAI hangt ook weer samen met de discussie over autonome wapensystemen, zoals drones. Wetenschappers, staten en de secretaris-generaal van de Verenigde Naties waarschuwen dat er altijd sprake moet zijn van betekenisvolle menselijke controle bij automatische besluitvorming door systemen.[38]

Ook op het terrein van cybersecurity biedt GAI kansen. Zo kan een GAI-systeem gevraagd worden om een cybersecuritysysteem te evalueren: waar zitten de zwakke plekken waarvan kwaadwillenden kunnen profiteren?[39] Ook kunnen systemen zoals Microsoft's Security Co-Pilot cybersecurityprofessionals in hun werk ondersteunen, door vragen te beantwoorden en te helpen snel op incidenten te reageren.[40] De GAI wordt getraind op basis van de data en IT-omgeving van de gebruiker, en past het advies daar ook op toe.

Voor zowel defensie als cybersecurity kunnen GAI-toepassingen tegenvallen. Zo is accuratesse essentieel bij het beveiligen van een IT-omgeving en het verzamelen van inlichtingen, en is het letterlijk van levensbelang bij het nemen van besluiten op het slagveld. Je moet erop kunnen vertrouwen dat een GAI geen domme fouten maakt, of door de tegenpartij gesaboteerd kan worden. Veel zal dus afhangen van de kwaliteit van GAI.

Arbeidsmarkt

Sinds de introductie van ChatGPT is er veel belangstelling voor de potentiële impact van generatieve AI op de arbeidsmarkt.[41] Op het oog is de belofte van generatieve AI indrukwekkend. Het kan nieuwe content creëren zoals tekst, beeldmateriaal, geluid of een combinatie hiervan, en ingezet worden voor het produceren van nieuwsberichten, artikelen, reclameteksten, samenvattingen, recepten, computercode of zelfs hele muzieknummers.[42] Een recente studie liet zien dat GAI al wordt ingezet om realtime gesprekssuggesties en mogelijke antwoorden te geven aan klantenservice medewerkers.[43] Uit deze studie bleek ook dat vooral minder ervaren medewerkers hiervan profiteerden, aangezien het systeem de meest productieve en succesvolle werknemers als voorbeeld neemt.

De verwachting is dat GAI-toepassingen kunnen leiden tot een substantiële verhoging van de productiviteit en efficiëntie.[44] Zo benoemt de OESO[45] 'transformerende' effecten van generatieve AI, onder meer omdat de systemen in zoveel sectoren zijn in te zetten:

---

38 United Nations, 2023
39 Al-Hawawreh et al., 2023
40 Jakkal, 2023
41 Chohan, 2023; Gmyrek et al., 2023; Knight, 2023; Villasenor & West, 2023
42 Zo verscheen er in het voorjaar plotseling een met AI gegenereerd nummer van Drake en The Weeknd. Het nummer ging in korte tijd viraal maar werd al snel offline gehaald door streamingsdiensten en online platforms, zie Coscarelli, 2023. Zie verder Bronzwaer, 2023.
43 Brynjolfsson et al., 2023
44 Alshurafat, 2023; Cardon et al., 2023; Noy & Zhang, 2023
45 Lorenz et al., 2023

van de zorg tot de rechtspraak, van de industrie tot het bankwezen. Ook kunnen de systemen eenvoudig bestaande taalbarrières doorbreken en internationale handel stimuleren. Door deze brede toepasbaarheid kan in principe elk beroep geraakt worden. Toch zijn de eerste verwachtingen dat de impact zich vooral laat gelden bij *white collar* banen, zoals kenniswerkers en managers, in tegenstelling tot eerdere automatiseringsgolven.[46] Vervolgstudies zullen moeten uitwijzen of deze verwachtingen uitkomen.

Belangrijke vragen zijn in hoeverre GAI-systemen banen zullen overnemen, in hoeverre mensen met de systemen zullen gaan samenwerken en tot op welke hoogte de technologie nieuwe banen mogelijk maakt. Denk aan een GAI-systeem dat taaladvies geeft tijdens het schrijven van een tekst, of een mens die het werk van het GAI-systeem naloopt. Banen bestaan uit bundels van taken, en technologie neemt vaak slechts een deel van het takenpakket over.[47] Onderzoekers van de University of Pennsylvania en OpenAI, het bedrijf achter ChatGPT, berekenden dat bij 80% van de banen minstens 10% van de taken geautomatiseerd zou kunnen worden met generatieve AI.[48] Werkenden houden zo tijd over om andere dingen te doen, en leren weer nieuwe vaardigheden aan. De International Labour Organisation (ILO) verwacht mede daarom dat GAI-systemen mensenwerk vooral zullen ondersteunen.[49]

Discussies over de toepassing van GAI in de cultuursector geven voorbeelden van hoe takenbundels zouden kunnen veranderen. Hoewel GAI in potentie kan worden ingezet om volledig zelfstandig kunst te genereren, is het goed mogelijk dat de technologie vooral zal worden gebruikt in de fase waarin het ruwe werk tot stand komt. De kunstenaar speelt dan een belangrijke rol aan het begin van projecten, als al dan niet met behulp van GAI ideeën worden ontwikkeld.[50] Ook aan het einde van projecten, als werk gepolijst moet worden, kan de kunstenaar nog aan zet zijn.[51]

Er kan een nieuw begrip ontstaan van wat kunst is, en van wie zich artiest mag noemen, waarbij mogelijk een nieuw soort artiest opkomt die profiteert van de mogelijkheden van generatieve AI. Ook leidt het gemak waarmee kunst met chatbots gemaakt kan worden tot een enorme hoeveelheid nieuw werk.[52] Ook nieuwe kunstgenres kunnen ontstaan; digitalisering maakte al eerder muziekgenres als Trance en Drum ’n bass mede mogelijk. Kunstenaars die AI voor hun werk willen gebruiken, zullen zich waarschijnlijk meer gaan bekwamen in het schrijven van *prompts* en nieuwe manieren bedenken waardoor ze hun individuele stempel op een werk kunnen

---

46 Chui et al., 2023; Gmyrek et al., 2023; Gownder & O’Grady, 2023. In vorige automatiseringsgolven waren met name routinematige taken gevoelig voor automatisering. Generatieve AI is ook in staat om niet-routinematige taken uit te voeren.
47 Rathenau Instituut, 2015; Went et al., 2015
48 Eloundou et al., 2023
49 Gmyrek et al., 2023
50 Epstein et al., 2023
51 Hugenholtz & Quintais, 2021
52 Epstein et al., 2023

drukken.[53] Ook is het denkbaar dat de toename van kunst gemaakt met behulp van AI zorgt voor een heropleving van ambachten en handgemaakt werk.[54]

Gezondheidszorg

Volgens de AI Monitor Ziekenhuizen 2023 en het Amsterdam UMC bieden LLM's kansen voor de medische praktijk.[55] Allereerst kan GAI ingezet worden voor informatiemanagement en administratie- en schrijftaken, en zo zorgprofessionals ontlasten en meer tijd geven voor patiënten.[56] LLM's kunnen gesprekken en informatie documenteren, transcriberen, samenvatten, classificeren en controleren.[57] Zo kunnen ze gebruikt worden voor efficiënte analyse van elektronische gezondheidsdossiers en andere data-archieven.[58] Een arts zou een systeem bijvoorbeeld kunnen vragen om metingen uit een operatieverslag te halen, geschikte deelnemers voor klinische trials te selecteren of gevaarlijke medicijninteracties in behandeldossiers te signaleren.[59] Ook kan GAI gebruikt worden voor assistentie, zoals afspraken inplannen, medicatie-innames managen, voor het produceren van informatieve content, zoals het maken van websites, bijsluiters en instructiefilmpjes, en voor het simplificeren van medisch jargon.[60]

De hoop is dat GAI systemen ook zijn in te zetten om misdiagnoses te voorkomen en behandelingen te verbeteren. Dit zou bijvoorbeeld basisartsen kunnen helpen om diagnoses te stellen zonder specialistische doorverwijzing.[61] Hierbij is het natuurlijk de vraag of dit verantwoord is, gezien de fouten die GAI-systemen kunnen maken.

GAI kan ook gecombineerd worden met andere technologieën, zoals draagbare sensoren en sensoren waarmee patiënten op afstand gemonitord kunnen worden. Of met de medische digitale representaties (*digital twins*) van Unlearn.AI, die met GAI-modellen voorspellen hoe de gezondheid van individuele patiënten kan veranderen in verschillende scenario's.[62] Ook zijn er combinaties mogelijk met Brain Computer Interfaces's (BCI's). Modellen kunnen hersenactiviteit sneller en accurater dan eerdere technieken decoderen tot tekst. Daardoor kunnen bijvoorbeeld mensen die niet kunnen praten door verlamming van de ziekte ALS, beter communiceren.[63] Ook is recent een *AI thought decoder* gekoppeld aan een Brain-Spine Interface, waardoor een verlamde

---

53 Epstein et al., 2023
54 Hugenholtz & Quintais, 2021
55 Janssen et al., 2023; Sparnaaij et al., 2023
56 Harrer, 2023; Van Buchem et al., 2021
57 Marr, 2023; Van Buchem et al., 2021
58 Harrer, 2023; Sweeney, 2021
59 Janssen et al., 2023; Marr, 2023
60 Harrer, 2023; Koncz, 2023; Loh, 2023; Marr, 2023
61 Het onderzoek van Raso et al., 2018 gaat over AI in het algemeen. Voor studies specifiek over GAI en LLM's, zie Acosta et al., 2022; Bell et al., 2023; Janssen et al., 2023; Lang et al., 2023; Nógrádi et al., 2023. Nógrádi et al. suggereren op basis van een onderzoek met prompts van neurologische symptomen, dat ChatGPT met hogere waarschijnlijkheid correcte diagnoses stelt dan een basisarts. Acosta et al. gaat in op de kansen van multimodale medische LLM's. Lang et al. gaan in op de de kansen die GAI ten opzichte van eerdere AI kan bieden voor betrouwbare *medical imaging.*
62 Marr, 2023; Unlearn.AI, n.d.
63 Ravindran, 2023; Tang et al., 2023; Whang, 2023a; Willett et al., 2023

man enigszins kon lopen.[64] Door andere wetenschappers is gesuggereerd dat toevoeging van GAI-software aan BCI-decoders een systeem op zou kunnen leveren dat dromen (hersenactiviteit) kan omzetten naar kunst, geluid of video, in plaats van alleen tekst.[65]

Naast professionals kunnen ook burgers GAI-chatbots gebruiken voor medische vragen of mentale steun. Bijvoorbeeld met een *prompt* als 'Wat is het beste voedingsplan voor een diabetespatiënt met hoge bloeddruk?'[66] Onderzoek suggereert zelfs dat chatbots sympathieker en uitgebreider antwoorden dan dokters.[67] Dit bevestigt de medeoprichter van mentale gezondheidsapp Koko, die GPT-3 toevoegde aan de applicatie, als toevoeging naast het advies van medische professionals. Berichten werden wel slechter beoordeeld zodra mensen ontdekten dat een machine betrokken was.[68] Uit een vergelijkende studie naar verschillende chatbots volgt dat er nog weinig bewijs is dat ze duidelijk voordeel kunnen bieden.[69]

2.3. Reflectie: de vier rollen van generatieve AI

Uit het hierboven beschreven overzicht komt naar voren dat GAI verschillende rollen speelt, vaak op hetzelfde moment. De eerste rol is die van leerinstrument. Deze rol is goed te zien in het onderwijs: leerlingen kunnen via GAI-chatsystemen informatie opzoeken en met GAI in gesprek gaan. Dit stelt leerlingen in staat snel veel informatie in een context te plaatsen. In het dagelijks leven gebruikt vrijwel iedereen digitale zoekmachines, en GAI kan de gebruiksvriendelijkheid hiervan verbeteren. Daarom hebben Google en Microsoft GAI ook in hun zoekprogramma's geïntegreerd. Het is echter de vraag in hoeverre docenten en leerlingen kunnen vertrouwen op de kwaliteit van de systemen, en welke vaardigheden leerlingen door het gebruik precies verwerven.

GAI wordt ook gebruikt als productietool: de gebruiker wil dat GAI iets maakt. In het onderwijs willen leerlingen een paper of een samenvatting hebben, terwijl leraren lesmateriaal laten maken. In de cultuursector wordt geëxperimenteerd met GAI die muziek componeert, literatuur schrijft of schilderijen maakt. Hierbij kan de mens samenwerken met GAI: zo kan een schrijver eerst zelf een tekst maken en het systeem vragen deze om te zetten naar een andere stijl of taal. Iemand die muziek wil produceren kan uitvoerige feedback geven op de deuntjes die GAI voortbrengt. Het productievermogen van GAI is snel en schaalbaar. Als een GAI-systeem eenmaal iets kan maken van een bepaalde kwaliteit, kan het daar eindeloos op variëren en nieuwe opdrachten uitvoeren. De belofte van deze rol is dat GAI goedkoop en snel producten kan leveren van hoge kwaliteit. De vraag is of die hoge kwaliteit inderdaad geleverd kan worden, én of we wel willen dat GAI bepaalde taken overneemt.

---

64 Lorach et al., 2023; Whang, 2023b
65 Kelsey, 2023a, 2023b
66 Harrer, 2023
67 Ayers et al., 2023; Korteweg, 2023
68 Ingram, 2023
69 Pandey & Sharma, 2023

De derde rol is die van probleemoplosser. De hoop is dat met behulp van GAI bepaalde complexe vraagstukken sneller zijn op te lossen. Voorbeelden hiervan zijn de generaal op het slagveld die moet besluiten waar bommen afgeworpen worden, en daarvoor advies kan krijgen via een evaluatie door een GAI-systeem, of de inzet van GAI bij de ontwikkeling van nieuwe medicijnen en biochemische structuren. In dit soort gevallen zou GAI antwoord proberen te geven op vragen die mensen niet, of slechts moeizaam, kunnen beantwoorden. Bij deze rol geldt natuurlijk dat de antwoorden voldoende moeten kloppen. Een ‘hallucinerend’ systeem kan door niemand worden vertrouwd – laat staan in de operatiekamer of op het slagveld. Bovendien is het zaak dat mensen wel kunnen achterhalen hoe de GAI tot bepaalde resultaten is gekomen, wat door de complexiteit van berekeningen moeilijk kan zijn.

Ten slotte wordt GAI ook gebruikt als nieuwe ervaring. Sommige gebruikers vinden het fascinerend om te communiceren met een systeem dat getraind is op immense datasets, en in staat is om op een begrijpelijke, en zelfs vriendelijke manier terug te praten. Zo is de chatdienst Replika ontwikkeld om je digitale vriend te zijn. Een chatsysteem zou een metgezel kunnen worden die je regelmatig raadpleegt, zoals verbeeld in de film Her. Op dit vlak is de vraag of de ervaringen die we hebben met GAI-systemen ons als mensen verrijken, of dat ze weinig aan onze levens toevoegen, of zelfs ongezond voor ons zijn.

2.4. Conclusie

GAI is een technologie die in veel maatschappelijke domeinen impact zal hebben, en voor allerlei taken ingezet zal worden. Het is een nieuwe technologie en het is vaak nog de vraag hoe effectief en werkbaar de technologie in de praktijk zal zijn, ook gezien de diverse technische zwaktes van GAI. GAI kan complexe taaltaken met indrukwekkende snelheid en precisie volbrengen, maar het is onzeker of de technologie betrouwbaar genoeg zal zijn voor de ondersteuning van besluitvorming in ziekenhuizen of op het slagveld. GAI zal in ieder geval de gehele samenleving raken en daarmee ook tal van publieke waarden. Hier gaan we in het volgende hoofdstuk verder op in.

3. Welke publieke waarden staan op het spel?

Dit hoofdstuk biedt een analyse van de risico's van generatieve AI voor de bescherming van publieke waarden. De analyse is gebaseerd op een studie van wetenschappelijke en grijze literatuur.[70] De risico's clusteren we in drie thema's: veiligheid, mensgerichtheid en eerlijkheid. Per thema benoemen we ook de publieke waarden die op het spel staan.

Met publieke waarden doelen we op datgene wat in een samenleving belangrijk wordt gevonden en waarvoor systematische bescherming nodig wordt geacht.[71] Het vertrekpunt voor onze clustering vormen verschillende inventarisaties van publieke waarden.[72] We beperken ons echter niet tot deze lijsten, omdat uit een analyse van de maatschappelijke impact van technologie zich ook nieuwe kwesties kunnen aandienen. Tot slot reflecteren we aan het einde van dit hoofdstuk op de betekenis van GAI voor de democratie.

3.1. Veiligheid

GAI-systemen kunnen op een aantal manieren mensen schade toebrengen en publieke waarden onder druk zetten: privacy- en dataprotectierechten kunnen worden geschonden, de systemen kunnen discrimineren en op verschillende manieren onbetrouwbaar zijn. Samen zetten ze de veiligheid van GAI-systemen onder druk.

Privacy en dataprotectie

De training en het gebruik van generatieve AI-systemen kunnen leiden tot schendingen van iemands privacy- en dataprotectierechten. Trainingsdata van de GAI-modellen kunnen persoonlijke informatie over iemand bevatten of als output persoonlijke informatie van iemand delen. In de Algemene Verordening Gegevensbescherming (AVG) is vastgelegd onder welke voorwaarden persoonsgegevens verwerkt mogen worden. Diverse Europese dataprotectietoezichthouders hebben daarom OpenAI om opheldering gevraagd, en onderzoeken of de verwerking in lijn is met de AVG. De Italiaanse toezichthouder verbood ChatGPT al tijdelijk in afwachting van zo'n onderzoek. De onderzoeken kunnen betekenen dat ontwikkelaars hun modellen moeten aanpassen. Die aanpassingen kunnen lastig zijn: zo gaf OpenAI al aan dat het

---

70 Hierbij is gezocht naar artikelen en overzichtstudies in wetenschappelijke databases over maatschappelijke en ethische risico's van generatieve AI en/of grote taalmodellen met de volgende zoektermen: ethical, moral, societal, social risks, implications, challenges, impacts of generative AI and LLM.

71 We volgen hierbij de interpretatie van studies uit de bestuurskunde, zie bijvoorbeeld Bozeman, 2007; Bruijn & Dicke, 2006; Nabatchi, 2018; Riemens et al., 2021. Er bestaan meerdere publieke waarden en deze zijn eerder dynamisch dan statisch. Sommige publieke waarden zijn nauw verwant aan mensenrechten, of al gecodificeerd in wettelijke kaders, zoals privacy, gelijke behandeling, eigenaarschap (recht op eigendom), werkgelegenheid (recht op werk, en goede arbeidsomstandigheden).

72 Citaat uit de Werkagenda, waarbij wordt verwezen naar het coalitieakkoord: "We hebben de plicht om grondrechten en publieke waarden (veiligheid, democratie, zelfeschikking, non-discriminatie, participatie, privacy en inclusiviteit) te beschermen en de taak om een gelijk economisch speelveld te creëren: met eerlijke concurrentie, consumentenbescherming en brede maatschappelijke samenwerking." In onderzoeken van het Rathenau Instituut over digitalisering komen ook waarden naar voren, zoals gezondheid, privacy, menselijke waardigheid, waarachtigheid, goed werk. Zie bijvoorbeeld Rathenau Instituut, 2017, 2019, 2020a, 2020b, 2020c, 2020d.

momenteel niet mogelijk is om persoonsgegevens op verzoek van de desbetreffende persoon te verwijderen of te corrigeren.[73]

Een GAI-systeem kan verder persoonlijke informatie afleiden uit de interactie met de gebruiker. De manier waarop iemand schrijft en communiceert, kan informatie geven over de politieke voorkeur of de gezondheid van die persoon.[74] De AVG biedt via de categorie ‘bijzondere persoonsgegevens’ personen extra bescherming aan dit soort intieme gegevens. Maar niet alle intieme informatie die de modellen kunnen verzamelen, vallen hieronder – zoals iemands emotie.

Ook wanneer een GAI-systeem meerdere databronnen combineert, kan het systeem uit die gecombineerde data intieme informatie afleiden. Zo liet een studie zien dat een model ‘door muren heen kon kijken’ door camerabeelden van mensen in een ruimte te combineren met Wifi-signalen van de elektronische apparaten van die mensen.[75] Het model had na training genoeg aan het detecteren van Wifi-signalen om het aantal mensen in de ruimte, en hun lichaamspose, gedetailleerd weer te geven.

GAI-modellen kunnen verder eenvoudig iemands stem, beeltenis of werkstijl nabootsen, zodanig dat de imitatie niet van echt te onderscheiden is. Zo’n ‘digitale kloon’ kan een individu schaden, zowel geestelijk als door identiteitsdiefstal. Dit risico is reëel, omdat het makkelijk is geworden dit te doen. Waar in 2021 circa twee minuten aan audiomateriaal genoeg was om een stem goed te klonen, kan dat nu met fragmenten van enkele seconden. Het is de vraag hoe je in zo’n digitale wereld je evenbeeld beschermt.

Voor de toekomst is de verwachting dat GAI-systemen nog meer intieme gegevens zullen verwerken. In de neurowetenschappen werken onderzoekers aan het interpreteren van fMRI-scans.[76] Onderzoekers zijn erin geslaagd om op basis van een hersenscan videobeelden of een foto te tonen van de gedachten van de proefpersoon. Deze onderzoeken bevinden zich nog in een vroege fase van ontwikkeling; ze werken bijvoorbeeld momenteel alleen voor de proefpersoon op wie het model is getraind. Toch roept deze ontwikkeling de vraag op in hoeverre iemands mentale privacy en vrijheid van gedachten in de toekomst kunnen worden beschermd. Wereldwijd wordt bediscussieerd hoe deze ‘neurorechten’ eruit zouden kunnen zien. Sommige landen hebben al varianten in de wet vastgelegd.[77] Een andere route is om te kijken naar bestaande een aangekondigde wetten, zoals de AVG en de AI-verordening.

---

73 Zie Norwegian Consumer Council, 2023.De privacy policy van OpenAI waarin dit geschreven stond, is op moment van schrijven niet beschikbaar via https://openai.com/nl/policies/privacy-policy.
74 Kaddour et al., 2023; Solaiman et al., 2023a; Weidinger et al., 2022
75 Geng et al., 2022
76 Chen et al., 2023; Takagi & Nishimoto, 2022
77 La Moncloa, 2021; UNESCO International Bioethics Committee, 2022

Non-discriminatie en inclusie

Sinds de lancering van ChatGPT en andere generatieve AI-systemen zijn er vele voorbeelden te vinden van onrechtvaardige, stigmatiserende, beledigende of anderszins niet-inclusieve resultaten. Dat heeft verschillende oorzaken. GAI-systemen zijn getraind met data, en reflecteren de daarin aanwezige vooroordelen.[78] Daarnaast bevatten de trainingsdata onvoldoende gegevens over specifieke groepen. Denk aan het feit dat er veel meer medische data over mannen beschikbaar zijn dan over vrouwen, of dat er relatief weinig internetpagina's zijn voor kleinere taalgebieden.

Die disbalans kan leiden tot onjuiste, onterechte en denigrerende content, en ongelijke behandeling van mensen of groepen. Juist degenen die historisch gezien het meest gemarginaliseerd zijn, lopen hier het meeste risico. Bias kan ook leiden tot fysieke schade, als dergelijke systemen in bijvoorbeeld de zorg worden ingezet.[79] Op de lange termijn kunnen bevooroordeelde generatieve AI-systemen discriminerende sociale normen bestendigen, zodanig dat deze normen bij de massale adoptie van de technologie lastig te veranderen zijn.[80] Bovendien kunnen de modellen politieke voorkeuren bevatten. Zo lieten Duitse onderzoekers zien dat ChatGPT een 'pro-milieu, links-libertaire' politieke oriëntatie bevatte. Andere onderzoekers vonden een rechts-autoritaire oriëntatie in ChatGPT4.[81] Juist omdat GAI-systemen voor allerlei taken ingezet kunnen worden, kan het vooroordeel in allerlei verschillende contexten terugkomen. Ontwikkelaars vragen daarom mensen, vaak in lagelonenlanden, de output van de modellen te beoordelen. Deze mensen moeten tegen een lage vergoeding nare content beoordelen.[82]

Het is de vraag of je het gevaar van discriminatie überhaupt kan wegnemen, of sterk kan verminderen. Wetenschappers geven namelijk aan dat de huidige technieken voor zorgvuldige datacuratie slechts mogelijk zijn bij kleinere datasets.[83] De hoop is dat op termijn nieuwe technieken beschikbaar zijn die datacuratie ook voor de enorme omvang van de trainingsdata voor GAI-systemen mogelijk maakt. Maar voorlopig kunnen GAI-modellen niet 'biasvrij' worden gemaakt.[84] Een garantie dat generatieve AI-systemen niet discrimineren is dus niet te geven.

Betrouwbaarheid

GAI-systemen kunnen op verschillende manieren onbetrouwbaar zijn. Enerzijds gaat het om gebreken van het systeem zelf, als het systeem bijvoorbeeld foutieve informatie bevat. Anderzijds gaat het om risico's die voortkomen uit de toepassing ervan, bijvoorbeeld wanneer kwaadwillende actoren het systeem inzetten om schade te

---

78 Abid et al., 2021; Bender et al., 2021; Kaddour et al., 2023; Sohail et al., 2023; Solaiman et al., 2023a; Weidinger et al., 2022

79 Zo kunnen bijvoorbeeld verkeerde medische adviezen gegeven worden. Zie bijvoorbeeld Rathenau Instituut, 2023a. Experts wijzen verder op het risico van teveel vertrouwen op de output van de systemen (overreliance) en confirmation bias – het systeem kan bevestigen wat mensen al denken.

80 Dit wordt ook wel *value lock* genoemd, bij massale adoptie kan het lastiger worden om culturele opvattingen te veranderen. Zie Bender et al., 2021; Weidinger et al., 2022.

81 Feng et al., 2023; Hartmann et al., 2023

82 Hao & Seetharaman, 2023; Perigo, 2023

83 Bender et al., 2021; Mittelstadt et al., 2023; Wachter et al., 2021

84 Het is uberhaupt de vraag of 'bias' geheel weggenomen kan worden. Zie voor toelichting Rathenau Instituut, 2022c.

veroorzaken. Een deel van het gevaar schuilt in een gebrek aan accuratesse: als een systeem verkeerde informatie geeft, kan dat leiden tot verkeerde medische diagnoses, een verkeerd advies over medicijngebruik, of programmeercode die kwetsbaarheden bevat.

Andere betrouwbaarheidskwesties ontstaan doordat mensen teveel gaan vertrouwen op het advies van een systeem, zodanig dat hun veiligheid in het geding komt (vergelijkbaar met de anekdotes dat mensen hun navigatiesysteem blind opvolgen, en zich zo in gevaarlijke verkeerssituaties manoeuvreren). De foutieve informatie (misinformatie) die GAI-systemen zelf produceren, kunnen gebruikers ook gaan verspreiden – als ze de informatie geloven. Een ander effect is dat ook authentieke content wordt betwijfeld.[85]

Naast het risico dat de systemen zelf foutieve informatie bevatten, is het mogelijk dat kwaadwillenden GAI-systemen bewust gebruiken om schade te veroorzaken.[86] Denk hierbij aan het laten schrijven van *malware*, of advies vragen over het maken van een bom, gevaarlijke chemicaliën of andere denkbare wapens. Ontwikkelaars proberen dit tegen te gaan door het systeem te leren deze informatie niet te geven. In de praktijk blijken gebruikers de aangebrachte vangrails vaak weer te kunnen omzeilen. Het is daarom denkbaar dat er informatie over gevaarlijke en ongewenste content gaat ronddwalen. Onderdeel daarvan is de verspreiding van desinformatie en deepfakes waar nu veel actuele voorbeelden online van te vinden zijn. Kwaadwillenden kunnen overtuigende valse berichten en filmpjes gaan maken, wat het publieke debat vervuilt en waarmee mensen persoonlijk aangevallen kunnen worden. Er worden al pornofilmpjes gemaakt waarbij de gezichten van andere mensen worden ingeplakt.[87] Je kunt met GAI-systemen iemands reputatie beschadigen en persoonlijke schade aanrichten. Maar verspreiding van desinformatie en deepfakes kan ook politiek gemotiveerd zijn en invloed hebben op het democratisch debat.

Ten derde kan het beperkte begrip onder ontwikkelaars, onderzoekers en gebruikers van de werking van de modellen leiden tot risico's. Het is onduidelijk, en onvoorspelbaar, hoe de systemen zich in bepaalde omstandigheden 'gedragen' en tot welke kwetsbaarheden dit leidt.[88] In deze discussie komt ook het eerder genoemde *alignment*-probleem terug: kunnen we er op vertrouwen dat GAI-systemen publieke waarden en wet- en regelgeving respecteren bij het uitvoeren van hun taken? Er zijn zorgwekkende voorbeelden van systemen die bereid zijn om tegen mensen te liegen

---

85 Dit gebeurde bijvoorbeeld in 2021 toen er een video online kwam waarin Viruswaarheidvoorman Willem Engel openlijk twijfelde aan de watersnoodramp in Limburg. Hij zette vraagtekens bij de overstromingen 'in ongeveer het enige gebied in Nederland dat boven zeeniveau ligt'. Op Twitter was men het er al snel over eens: deze uitspraken waren zo absurd, dit moest wel een deepfake zijn. De onduidelijkheid over wat authentiek is maakt het ook mogelijk om te ontkennen dat waargebeurde zaken ooit plaats hebben gevonden. Dit wordt de 'liar's dividend' genoemd.

86 Zie onder meer Doorenbosch, 2023; Europol Innovation Lab, 2023; Gupta et al., 2023; Weidinger et al., 2022; Yamin et al., 2021.

87 Zo werden onder andere Dionne Stax en Welmoed Sijtsma slachtoffer van deepfake pornografie, zie Van de Ven, 2023. Ook niet bekende personen kunnen hier mee te maken krijgen. In Spanje werden onlangs van tientallen meisjes AI-gegenereerde naaktfoto's verspreid op sociale media NOS, 2023.

88 Ananthaswamy, 2023; Anderljung et al., 2023; Bianchi & Hovy, 2021; Bommasani et al., 2022; Bowman, 2023

om een taak te volbrengen. Het lukte een GAI-systeem om een mens te vragen visuele taken uit te voeren door hem te misleiden. Het systeem beweerde iemand te zijn met een visuele beperking.[89] Er zijn nog meer van dit soort voorbeelden. Dat verklaart de angst van sommige partijen dat het veiligheidsrisico van GAI-systemen op de lange termijn niet te overzien is, en door hen existentieel wordt genoemd.

De relatieve geslotenheid van private ontwikkelaars over de ontwikkeling van hun taalmodellen maakt dat de samenleving beperkt zicht heeft op de capaciteiten van de modellen – en hoe deze zich ontwikkelen.[90] Het is daarom ook beperkt mogelijk om de claims en speculaties van ontwikkelaars te controleren, en te bepalen welke veiligheidsrisico's hun systemen met zich meebrengen. Tot slot is kennis over de werking van het systeem van belang met het oog op uitlegbaarheid en de mogelijkheid om verantwoording af te kunnen leggen (*accountability*). De mogelijkheid moet bestaan om als individu geïnformeerde keuzes te maken, om genoegdoening bij fouten te geven of te krijgen, en verantwoordelijkheden toe te wijzen voor het voorkomen van fouten. Dit geldt zeker in publieke sectoren, waaronder het gebruik van GAI door de openbaar bestuur, in de zorg, of in de rechtspraak.[91]

3.2. Mensgerichtheid

GAI-systemen oefenen invloed uit op hoe mensen zich ontwikkelen en samenleven. De risico's die hier spelen hebben impact op publieke waarden zoals menselijke waardigheid, autonomie, gezondheid, en pluriformiteit, maar zijn niet goed bij een enkele publieke waarde in te delen. We bespreken de risico's daarom aan de hand van drie aspecten van menselijke ontwikkeling: cognitief, sociaal en cultureel.

Cognitieve ontwikkeling

We noemden in hoofdstuk 2 dat generatieve AI kan worden gebruikt als leerinstrument. Daarbij spelen zorgen dat de vaardigheden van gebruikers juist kunnen afnemen: *deskilling*. De discussie draait met name om hogere cognitieve vaardigheden, zoals creativiteit, kritische reflectie en de vaardigheden om te leren.[92] GAI-systemen kunnen dit leerproces frustreren, bijvoorbeeld als leerlingen het systeem essays laten produceren, of bij ieder brainstormproces beginnen met de ideeën die het GAI-systeem aanreikt. Dit is wellicht minder een probleem als iemand deze vaardigheden al bezit, en een GAI-systeem handig kan bijsturen. Maar als een GAI-toepassing de taken die een mens nodig heeft voor zijn leerproces vervangt, hoe kan iemand dan nog een complexe taak onder de knie krijgen?[93] Soms is voor het verkrijgen van een hogere vaardigheid

---

89 Het gaat hier om een zogenaamde CAPTCHA-taak, gebruikt door websites om robots van mensen te onderscheiden. Hier vraagt het model een persoon een visuele CAPTCHA-taak uit te voeren. Als die persoon vraagt waarom (ben je soms een robot?), liegt het model door te stellen dat het een mens is met een visuele beperking. Zie Edwards, 2023.

90 Momenteel beschikken wetenschappers en ontwikkelaars over onvoldoende benchmarks/methoden om de capaciteiten/output van modellen precies te meten. Zie Bommasani et al., 2022; Kaddour et al., 2023; Urbina et al., 2023; Zhao et al., 2023

91 European Commission High Level Expert Group on AI, 2019; OECD.AI Policy Observatory, n.d.; Rudin, 2019; UNESCO, 2022

92 Blodgett & Madaio, 2021; Farrokhnia et al., 2023; Kasneci et al., 2023; Lodge et al., 2023

93 Malinka et al., 2023

eerst nodig veel simpeler werk te doen. Misschien kan je het beste een essay op niveau bijsturen, als je zelf al essays van begin tot eind hebt geschreven.[94]

Sociale ontwikkeling

GAI kan leiden tot een intieme interactie tussen mensen en chatbots. Het is bekend dat mensen zich kunnen hechten aan levenloze objecten, zoals knuffels, auto's en dus ook computers. Dit fenomeen wordt antropomorfisme genoemd, en kan ook optreden als een object of systeem geen menselijke trekjes nabootst.

De afgelopen jaren stelden ontwikkelaars zich ten doel om menselijk en sociaal gedrag zo nauwkeurig mogelijk na te bootsen,[95] en met de taalvaardigheid van GAI-systemen is een nieuwe, grote stap gezet – zeker nu daar ook levensechte stemmen en beeld aan worden toegevoegd.[96] Er zijn al bots die je kunt leren zich te gedragen als een overleden geliefde[97] en chatbots als Replika die je kunt instellen als vriend of vriendin. De app wordt onder meer door mensen gebruikt die verlegen of eenzaam zijn. Chatbots – in tegenstelling tot mensen altijd beschikbaar – reageren onmiddellijk en hebben een eindeloos geduld. Niet voor niets adverteert Replika met de boodschap: *"Replika is for anyone who wants a friend with no judgment, drama, or social anxiety involved."*[98]

Hoewel antropomorfisme positieve aspecten kan hebben, zijn er zorgen over hoe dit fenomeen sociale vorming en sociale omgang beïnvloedt. Wetenschappers buigen zich al langere tijd over deze kwestie. Wat als virtuele ontmoetingen zo verslavend zijn dat mensen de behoefte aan menselijke ontmoetingen verliezen? Ontleren mensen sociale omgangsvormen als ze de eigenschappen van een GAI-systeem gewend zijn?[99] Wat zijn de psychologische gevolgen als je leunt op een chatbot voor emotionele steun? Is het bevorderlijk voor de rouwverwerking wanneer mensen chatbots maken die hun overleden geliefde imiteren?

Het is belangrijk op deze vragen antwoord te krijgen, en ervoor te zorgen dat het menselijk samenzijn niet wordt ondermijnd door onze omgang met computers.[100] Extra voorzichtigheid bij kwetsbare groepen lijkt geboden, zoals bij kinderen – in de wetenschap dat diverse chatbots specifiek gericht zijn op kinderen. Zo bood de chatbot van Snapchat al aan om met kinderen in de echte wereld af te spreken.[101]

Culturele ontwikkeling

De opkomst van GAI-systemen kan ook invloed hebben op culturele ontwikkelingen: bots zullen betrokken zijn in culturele uitingen, of deze ook gaan maken. Dit roept de vraag op of de creatieve vermogens van mensen in de toekomst voldoende worden

---

94 Tegelijkertijd zal de komst van GAI ook leiden tot nieuwe vaardigheden, juist als het gaat om de omgang met het systeem: hoe kan de gebruiker er zoveel mogelijk uithalen?
95 Véliz, 2023
96 OpenAI, 2023
97 Fagone, 2021
98 Apple, n.d.
99 Turkle, 2015
100 Danaher, 2019, 2020
101 NOS Nieuws, 2023. Inmiddels is dit aangepast.

aangesproken en zich voldoende kunnen ontwikkelen. Als ieder schilderij begint met een voorzet van een GAI-systeem, heeft dit impact op de artistieke vermogens van kunstenaars. Is er nog tijd en ruimte voor ongedwongen en onverwachte creatieve processen, als je met een druk op de knop tal van beelden en teksten kan produceren? De output kan culturele en artistieke normen versterken, omdat de datasets historische voorbeelden bevatten. Dit mechanisme kan de pluriformiteit van culturele uitingen onder druk zetten. Het is de vraag hoe *nieuw* de output van de systemen in feite is. Bovendien zullen ideeën en stijlen die op sociale media aandacht krijgen veel gekopieerd worden, en kunnen bepaalde culturele opvattingen en keuzes oververtegenwoordigd raken in datasets.[102]

Bovenstaande ontwikkelingen roepen de vraag op naar ons menszijn in een wereld van robots. Welke menselijke activiteiten willen we niet uitbesteden? En welke menselijke vaardigheden willen we niet verliezen? Omdat GAI-systemen taal veel beter beheersen dan eerdere AI-toepassingen, kunnen ze ons aanspreken, vermaken en verleiden. Daardoor is het aantrekkelijk om steeds meer van deze technologie gebruik te maken, en onszelf er voortdurend mee te omgeven. Maar leiden al die individuele keuzes tot een maatschappij waar we goed in gedijen?

3.3. Eerlijkheid

Generatieve AI-systemen kunnen werk van mensen overnemen, en zo invloed hebben op de economie. Ook noemden we al dat dominante techbedrijven hun positie via GAI-systemen verder kunnen versterken. Verder zijn er vragen over de invloed van de training en het gebruik van de systemen op het milieu. We bespreken hier vier publieke waarden die door GAI-systemen negatief geraakt kunnen worden: eigenaarschap, duurzaamheid, werkgelegenheid en kwaliteit van werk. Deze aspecten hebben allemaal te maken met de manier waarop de baten en lasten van deze technologie in de samenleving verdeeld worden, en hoe rechtvaardig die verdeling is.

Eigenaarschap

Een belangrijke vraag die GAI-systemen oproepen, is of ontwikkelaars de data waarmee ze de taalmodellen trainen rechtmatig gebruiken. De data bestaan namelijk uit ontelbare bijdrages van kunstenaars, schrijvers, wetenschappers, vertalers en programmeurs die in de loop der jaren zijn gedigitaliseerd. Zonder die creatieve werken zouden de taalmodellen niet kunnen functioneren.

Het auteursrecht wordt als fundamenteel recht beschermd onder het recht op eigendom, en beschermt creaties van mensen zoals teksten, films en software, om het maken van kunst te erkennen, te belonen en aan te moedigen.[103] Makers maken zich door het massale gebruik van creatief werk in de modellen ernstig zorgen over hun toekomst. Inmiddels zijn er diverse rechtszaken tegen ontwikkelaars van GAI-systemen aangespannen. De rechter zal moeten verhelderen hoe het auteursrecht, en de wettelijke uitzonderingen voor tekst- en datamining, precies van toepassing zijn. Verder

[102] Epstein et al., 2023
[103] Visser, 2023

is het de vraag hoe we moeten omgaan met output van GAI-systemen die bestaand werk imiteren, er slechts een beetje van afwijken, en daardoor geen inbreuk maken op het auteursrecht. Is dit wenselijk?

Een ander element van eigenaarschap gaat over de vraag wie uiteindelijk de technologie bezit en er controle over heeft. Er bestaan zorgen over de grote rol die een handvol grote technologiebedrijven spelen, zoals Meta, Google en Microsoft. We bespreken die zorgen in paragraaf 3.4.

Duurzaamheid

In het eerste hoofdstuk bespraken we de benodigde computerkracht die grote techbedrijven gebruiken voor het trainen van grote taalmodellen. Die rekenkracht heeft impact op het milieu, in termen van energieverbruik, CO_2-uitstoot en waterverbruik. Ook zijn er schaarse grondstoffen nodig om de hardware te bouwen. Onderzoekers werken aan energiezuinigere oplossingen.[104] Tegelijkertijd wordt de digitale transitie, waaronder kunstmatige intelligentie en grote taalmodellen, als essentieel gezien om de duurzame transitie te realiseren (de zogenaamde *twin transition*).[105]

Momenteel bestaan er nog geen gestandaardiseerde methoden om de milieu-impact van AI (en generatieve AI) te meten; verschillende methoden meten verschillende aspecten.[106] Zo schatte een studie dat het integreren van taalmodellen in zoekmachines zou kunnen leiden tot 4 à 5 keer meer computerkracht per zoekopdracht, en daarmee aanzienlijke toename in energieverbruik en CO_2-uitstoot zou betekenen.[107] Naast de training heeft ook het gebruik van de systemen milieu-impact. Zo wordt geschat dat per chatconversatie van circa 20-50 antwoorden circa 500 ml koelwater nodig is – grofweg een flesje water per sessie.[108]

Wetenschappers en bedrijven werken vanwege de oplopende kosten aan alternatieven die efficiënter zijn, en daarmee ook zuiniger, zoals minder computergeheugen via (Q)LoRa, efficiëntere algoritmen, nieuwe technieken zoals quantumcomputing, kleinere modellen en meer kwalitatief goede data.[109] Technieken als (Q)LoRa maken het al mogelijk om ruwe modellen verder te verfijnen met consumentenhardware – waarmee het aanpassen van bestaande modellen voor meer partijen toegankelijker en zuiniger wordt.

Tegelijkertijd is de verwachting voor de gehele IT-infrastructuur dat technologische vergroening alleen niet voldoende zal zijn.[110] De groei van de digitale infrastructuur is ook afhankelijk van de eisen die gebruikers, samenleving en politiek hieraan stellen –

---

104 Lorenz et al., 2023; Patterson et al., 2021; Vipra & Myers West, 2023
105 European Commission, 2020; Muench et al., 2022
106 Lorenz et al., 2023; OECD, 2023b; Rathenau Instituut, 2022a; Solaiman et al., 2023b
107 Stokel-Walker, 2023
108 Li et al., 2023
109 Ananthaswamy, 2023; Dettmers et al., 2023; Vipra & Myers West, 2023
110 Rathenau Instituut, 2022a

en de keuzes die hierin worden gemaakt. Er blijft dus een uitdaging om de digitale economie te verduurzamen.

Werkgelegenheid en kwaliteit van werk

In hoofdstuk 2 beschreven we dat generatieve AI te gebruiken is als productietool. Een werkende kan daarmee productiever worden, maar het kan ook betekenen dat werk geautomatiseerd wordt. In het publieke debat gaat veel aandacht uit naar mogelijk banenverlies door generatieve AI. Vooral uit de techsector zelf komen alarmerende geluiden over de impact die GAI kan hebben op de arbeidsmarkt. Zo waarschuwde de CEO van OpenAI dat kunstmatige intelligentie onherroepelijk zal zorgen voor het verlies van banen, waarschijnlijk al binnen tien jaar.[111]

Uit de geschiedenis valt op te maken dat technologische ontwikkelingen banen doen verdwijnen, maar dat er tot nu toe meer nieuwe banen bijkwamen, en banen vooral door technologie zijn veranderd.[112] In het verleden zijn zorgen over massawerkloosheid niet uitgekomen, maar vroegen de nieuwe, en veranderende, banen wel om nieuwe vaardigheden van de beroepsbevolking. Via scholing en beleid is het gelukt aan deze veranderende vraag naar vaardigheden te voldoen en kwetsbare groepen – de mensen die hun baan verloren en voor wie omscholing lastig was – meer bescherming te bieden. Bij eerdere technologische revoluties bleken met name routinematige taken gevoelig voor automatisering. Sinds de opkomst van AI blijken ook diverse niet-routinematige taken geautomatiseerd te kunnen worden.

De afgelopen jaren zijn er studies gedaan naar de invloed van automatisering en AI op de arbeidsmarkt. Daarin wordt gewezen op mogelijke effecten zoals baan- en loonpolarisatie, kortdurige werkloosheid of een meer ongelijke inkomens- of vermogensverdeling.[113] De OESO rapporteert een verwachte toename van het aantal banen dat geraakt kan worden door AI van 14% in 2019 naar 27% in 2022. Dat kan met de opkomst van GAI toenemen.[114]

Een studie van de International Labour Organisation (ILO), de organisatie van de Verenigde Naties die zich bezighoudt met arbeidsvraagstukken, kijkt ook naar de invloed van GAI. Zij berekenen dat de hoogste mate van automatisering terecht zal komen bij kantoorbanen en kenniswerkers.[115] Volgens de ILO valt dit te verklaren door het feit dat GAI in tegenstelling tot eerdere technologieën in staat is om ook niet-

---

111 Andersen, 2023; Felsenthal & Perrigo, 2023 OpenAI deed eerder zelf onderzoek naar de mogelijke impact op de arbeidsmarkt Eloundou et al., 2023

112 Banen zijn bundels van taken, waarbij technologie een deel van deze taken kan overnemen, en het werk voor de mens verandert. Rathenau Instituut, 2015; Went et al., 2015

113 Bij baan- of loonpolarisatie stijgen de lonen aan de onder- en bovenkant van de arbeidsmarkt, en komt het middensegment onder druk te staan. Dit fenomeen is al langere tijd zichtbaar. Zie van den Berge & ter Weel, 2015 Bij frictiewerkloosheid is sprake van een tijdelijke stijging van werkloosheid door de mismatch tussen gevraagde en aanwezige vaardigheden bij de beroepsbevolking Koning, 2013 Ongelijkheid in inkomen kan toenemen doordat het aandeel loon van een nationaal inkomen daalt, omdat het inkomen verkregen uit kapitaal groeit Autor, 2022. Zie ook Brynjolfsson, 2022.

114 OECD, 2023a. In 2019 schatte de OESO dit percentage op circa 14% OECD, 2019.

115 Gmyrek et al., 2023

routinematige cognitieve taken uit te voeren. Toch concludeert de ILO dat het in de meeste gevallen zal gaan om een gedeelte van het takenpakket dat zal worden geautomatiseerd, waardoor de technologie werkenden waarschijnlijk vooral zal ondersteunen in plaats vervangen. Ook wetenschappers verwachten dat GAI taken van kenniswerkers zal overnemen, maar er nieuwe taken bij zullen komen.[116]

Naast de kwantiteit van werk, is ook de kwaliteit van werk onderdeel van de discussie over automatisering van werk. Is het werk dat door mensen gedaan wordt ook *goed* werk? Ook met de opkomst van GAI spelen deze zorgen. Zo zou een gedeeltelijke automatisering van het werk kunnen zorgen voor een verslechtering van de positie en arbeidsomstandigheden van werkenden. Het kan bijvoorbeeld leiden tot minder uitdagend werk.

Kwaliteit van werk staat ook onder druk door de arbeidsomstandigheden van degenen die de data en output van de modellen beoordelen. Het is bekend dat het verwerken van haatzaaiende, discriminerende of anderszins illegale of ongewenste content psychische schade kan aanrichten, bijvoorbeeld omdat mensen herhaaldelijk worden blootgesteld aan obscene of schokkende uitingen.[117] Ook vindt dergelijk werk vaak plaats onder slechte arbeidsomstandigheden. Het werk wordt verricht door werknemers die weinig arbeidsbescherming genieten en in een precaire positie verkeren, zoals vluchtelingen. Dit leidt tot schadelijke praktijken, zoals onderbetaling of weigering van betaling. Daarnaast wordt dit werk beperkt gedocumenteerd, wat de ondoorzichtigheid van de werking van taalmodellen vergroot.

De OESO waarschuwt voor de bovenstaande risico's en roept landen daarom op medewerkers met een laag inkomen beter te ondersteunen, te investeren in veiligheidsmaatregelen en verantwoord gebruik van GAI op de werkvloer en te investeren in nieuwe vaardigheden.[118]

3.4. Reflectie: GAI zet druk op democratische samenleving

Uit het overzicht van publieke waarden en risico's komt een rode draad naar voren. De ontwikkeling en toepassing van GAI zet op twee manieren de democratische samenleving onder druk. Ten eerste zijn er zorgen over beïnvloeding van het publieke debat door de verspreiding van desinformatie en misinformatie, en de gevolgen die dit kan hebben voor het sociale en politieke vertrouwen, en voor de communicatie tussen burgers en volksvertegenwoordigers. GAI kan democratische processen ondermijnen. Ten tweede kan de groeiende machtspositie van enkele techbedrijven in steeds meer maatschappelijke domeinen de democratische sturing van digitale technologie bemoeilijken.

---

116 Autor, 2022
117 Hao & Seetharaman, 2023; Norwegian Consumer Council, 2023; Perigo, 2023
118 OECD, 2023b

Democratische processen

Er zijn al langer zorgen over de toename van online desinformatie, en generatieve AI maakt die zorgen groter. Het nieuws en het publieke debat zijn belangrijke informatiebronnen op basis waarvan burgers hun mening vormen over maatschappelijke thema's en politieke kwesties. Het is dan ook belangrijk dat iedereen toegang heeft tot betrouwbare en waarheidsgetrouwe informatie. Met behulp van GAI-systemen kan helaas op grote schaal foutieve en misleidende informatie worden gegenereerd. Zo trof de Amerikaanse nieuwswaakhond NewsGuard 37 sites aan die chatbots gebruiken om artikelen van media als CNN, The New York Times en Reuters over te nemen en aan te passen.[119] De sites maken gebruik van software die zonder tussenkomst van een mens artikelen kan vinden, herschrijven en publiceren.

Ook bestaat het risico dat mensen zo op GAI-systemen gaan vertrouwen, dat ze die systemen als een alwetend orakel gaan zien. De systemen kunnen in dat geval een grote invloed hebben op de opinies en het wereldbeeld van de gebruiker. Ze kunnen ook een sterke bemiddelende rol spelen bij verkiezingen, omdat ze beïnvloeden welke informatie een gebruiker tot zich neemt. Dat effect wordt versterkt doordat GAI kan worden ingezet voor *hyperpersonalisatie*: de gebruiker krijgt dan alleen nieuws te zien dat hem interesseert of zijn wereldbeeld bevestigt. Hierdoor ontbreekt het niet alleen aan een gedeelde waarheid of overeenstemming over feiten, maar ook aan gedeelde ervaringen.

Ten slotte zou GAI het vermogen van volksvertegenwoordigers om te reageren op signalen uit de samenleving kunnen verzwakken.[120] Zo kunnen GAI-systemen worden misbruikt in democratische processen, zoals publieksconsultaties, door op grote schaal inzendingen te versturen om zo te proberen de uitkomsten van de consultaties te beïnvloeden.[121] Ook via sociale media of de mailboxen van parlementariërs kunnen kwaadwillende personen een verwrongen beeld geven van wat leeft in de samenleving.

Democratische sturing op digitale technologie

Het tweede risico heeft te maken met de groeiende machtspositie van enkele techbedrijven. Inmiddels is een aanzienlijk deel van de taalmodellen en toepassingen in handen van techbedrijven als Microsoft, Google en Meta. Zij bepalen wie er toegang heeft tot de modellen, onder welke voorwaarden en voor welke prijs, hoe ze worden getraind, met welke data, en welke content geprioriteerd of juist gefilterd wordt. Met andere woorden: de techbedrijven bepalen grotendeels de waarden waar de GAI-toepassing rekening mee houdt.

De bedrijven hebben middels hun socialemediaplatformen ook invloed op de manier waarop nieuws en politieke ideeën zich verspreiden, en hoe het publieke debat verloopt. De zorgen over de machtspositie van techbedrijven beperken zich dus niet tot marktmacht, maar betreffen ook de invloed op publieke domeinen zoals onderwijs,

---

119 Brewster et al., 2023
120 Kreps & Kriner, 2023
121 Bell et al., 2023

zorg, journalistiek en de rechtstaat. Binnen deze sectoren worden publieke instellingen steeds afhankelijker van de diensten van enkele technologiebedrijven. Dit zet het vermogen onder druk om gezamenlijk, met burgers, maatschappelijke organisaties, publieke instellingen en bedrijven, te sturen hoe digitale technologie ontwikkeld en toegepast wordt in de samenleving.[122]

Een nieuwe kwestie hierbij is de invloed van enkele techbedrijven op wetenschappelijke kennisontwikkeling. Het trainen van geavanceerde taalmodellen vraagt dermate veel rekenkracht, dat het daaraan vasthangende prijskaartje voor veel academische instellingen onbetaalbaar is. Dit maakt ze afhankelijk van de modellen van de techbedrijven. Bovendien is het lastig om de kennisclaims van techbedrijven te controleren, omdat publieke kennisinstellingen doorgaans beperkte toegang tot de onderliggende trainingsdata en transparantie van de werking van de taalmodellen van technologiereuzen ontbreekt. Ten slotte waarschuwen diverse wetenschappers al langer dat een te grote nadruk op data en statistiek de wetenschap kan verarmen, bijvoorbeeld wanneer correlaties gelijk worden gesteld aan het creëren van wetenschappelijke kennis en bijbehorende methoden.[123] Daarom is er discussie over de vraag in hoeverre, en onder welke condities, wetenschappers grote taalmodellen moeten gebruiken.[124] Sommigen stellen dat maatschappelijk verantwoorde wetenschap alleen gebruik maakt van modellen die helder te interpreteren zijn.

3.5. Conclusie

Het vorige hoofdstuk liet zien dat GAI tal van kansen biedt, maar voorlopig niet goed genoeg is om te gebruiken in kritieke processen in de gezondheidszorg of defensie. Dit hoofdstuk liet zien dat GAI-technologie gepaard gaat met tal van risico's in relatie tot publieke waarden. Een deel van deze kwesties is binnen de digitale samenleving niet nieuw, zoals privacy, discriminatie, desinformatie, duurzaamheid, veiligheid en de machtspositie van techbedrijven. GAI versterkt en compliceert bestaande problemen.

Zo was het voorkomen van discriminatie in lerende systemen al niet gemakkelijk en is dit bij GAI voorlopig onmogelijk. De milieu-impact van de digitale infrastructuur was al een zorg, en GAI doet daar nu een significante schep bovenop. Het verspreiden van desinformatie was al lastig tegen te gaan, en nu maakt GAI-technologie de productie ervan aanzienlijk eenvoudiger. Lerende AI-systemen waren al een black box, en de complexiteit van GAI maakt een heldere interpretatie van de werking vrijwel onmogelijk. Technologiebedrijven waren al machtig, en met GAI kunnen zij hun economische positie en hun rol in de informatievoorziening van de samenleving significant versterken. Bedenk bovendien dat deze taalmodellen de basis kunnen vormen van vele andere toepassingen.

Een deel van de risico's van GAI is nieuw; ze speelden niet of nauwelijks een rol in het digitaliserings- en AI-beleid van de afgelopen vier jaar. Zo was er de afgelopen

---

122 Kerssens & van Dijck, 2023; Nemitz, 2018; Passchier, 2021; Sharon, 2016; van Dijck et al., 2018
123 Anderson, 2008; Haggart, 2023
124 Bender et al., 2021; Birhane et al., 2023; Rudin, 2019

kabinetsperiode weinig discussie over het auteursrecht, of over de gevolgen van automatisering op de werkvloer. En hoewel de machtspositie van technologiebedrijven op de politieke agenda stond, komt daar met hun invloed in de wetenschap een zorg bij. Academische en publieke kennisinstellingen hebben doorgaans beperkt toegang tot de onderliggende data en broncode van de modellen. De claims van de ontwikkelaars, onder meer met betrekking tot de werking, nieuwe functies, veiligheid en andere risico's, zijn daardoor lastig te controleren.

Al met al tellen de bestaande en nieuwe risico's van GAI op tot een zorgelijk beeld, dat de democratie raakt en zich ten dele ook al manifesteert. Dagelijks rapporteren gebruikers over foutieve informatie, vooroordelen en deepfakes. Het is daarom noodzaak de risico's van GAI het hoofd te bieden. In het volgende hoofdstuk verkennen we hoe beleidsmakers, politici, bedrijven, maatschappelijke organisaties en burgers dat kunnen doen.

4. Welke beleidskeuzes liggen voor?

In het vorige hoofdstuk constateerden we dat GAI-technologie risico's in de digitale samenleving versterkt en nieuwe risico's met zich meebrengt. De bedrijven die GAI-toepassingen ontwikkelen, lijken het adagium *'move fast and break things'* weer te volgen: ze introduceren toepassingen in de samenleving en zien vervolgens wel wat er precies mee gebeurt. Dit is riskant. Zonder mitigerend beleid kunnen de schadelijke effecten van GAI-toepassingen de overhand nemen.

Een deel van dat beleid is al gemaakt. Het afgelopen decennium is veel werk verzet om digitalisering en in het bijzonder AI in goede banen te leiden. Op internationaal niveau zijn uitgangspunten geformuleerd voor maatschappelijk verantwoorde toepassingen.[125] De Europese Unie tracht met bestaande en nieuwe wetten deze uitgangspunten juridisch te borgen, zoals de Algemene Verordening Gegevensbescherming, de Wet inzake Digitale Diensten, de Wet inzake Digitale Markten en de aankomende AI-verordening. In Nederland zijn publieke waarden en maatschappelijke uitdagingen ook steeds centraler komen te staan in het digitaliseringsbeleid.[126]

De hamvraag is of deze inspanningen genoeg zullen zijn. Het is een reële mogelijkheid dat het huidige en voorgenomen beleid niet opgewassen zijn tegen de impact van generatieve AI-systemen, onder meer op het terrein van non-discriminatie, veiligheid, mededinging, desinformatie en de uitbuiting van werknemers. Het is daarom zaak dat het kabinet een strategie uitwerkt om de grip van de samenleving op deze technologie te versterken. Dat begint met het evalueren en tijdig aanpassen van het Nederlandse en Europees beleid, van wetgeving tot financiële en communicatieve instrumenten.[127] Gelet op het brede palet van publieke waarden dat daarbij in het geding is, is dit een even complexe als urgente opgave.[128] Het is aan de overheid om deze taak op te pakken, en samen met het bedrijfsleven en maatschappelijke partijen op verantwoorde wijze vorm te geven.

Het kabinet heeft al aangegeven te kijken naar een *rapid response team*, een groep experts van binnen en buiten de overheid die kan adviseren over generatieve AI. Ook is aangegeven dat het kabinet wil kijken naar een team dat 'niet alleen maar adviseert, maar ook de mogelijkheid heeft om in te grijpen op een snelle manier (…)'. Het is goed om te verkennen hoe bestaande expertise kan worden benut. Tegelijkertijd roept het voornemen om "in te grijpen" de vraag op wat daarmee precies wordt bedoeld, en hoe het team zich verhoudt tot het werk van toezichthouders en de democratische verantwoording van beleid.

---

125 European Commission High Level Expert Group on AI, 2019; OECD.AI Policy Observatory, n.d.; UNESCO, 2022

126 Ministerie van Economische Zaken en Klimaat, 2019; Rijksoverheid, 2022b

127 Zie hier de instrumenten genoemd in het beleidskompas (Kenniscentrum voor beleid en regelgeving, n.d.), en de publicatie *Stand van digitaal nederland* (Rathenau Instituut, 2021).

128 Men kan hierbij voortbouwen op de onderzoeken die adviesraden en kennisinstellingen naar de brede governance van AI en digitalisering al hebben gedaan. Zie Cyber Security Raad, 2021; European Parliamentary Research Service et al., 2021; Onderwijsraad, 2022; Raad voor de leefomgeving en infrastructuur, 2021; Raad voor het Openbaar Bestuur, 2021; Rathenau Instituut, 2021, 2022a, 2022b; Wetenschappelijke Raad voor het Regeringsbeleid, 2021.

Gezien de breedte van het relevante beleid kan een integrale beleidsanalyse niet binnen de scope van deze scan uitgevoerd worden. Maar op basis van literatuurstudie, gesprekken met experts en een werksessie met beleidsambtenaren, geeft het Rathenau Instituut het kabinet vijf handelingsopties mee:

1. Creëer het vermogen om schadelijke GAI-toepassingen van de markt te halen;
2. Zorg voor toekomstbestendige juridische kaders;
3. Investeer in internationaal AI-beleid, om mondiale innovatieprocessen van technologiebedrijven bij te sturen;
4. Stel een ambitieuze agenda op voor maatschappelijk verantwoorde GAI;
5. Stimuleer maatschappelijk debat over de wenselijkheid van GAI.

4.1. Creëer het vermogen om schadelijke GAI-toepassingen van de markt te halen

Zorg voor een rem op de toepassing en uitrol van GAI-producten

Het is voorstelbaar dat een GAI-toepassing dermate veel schade aanricht, dat het van de markt gehaald moet worden. De samenleving moet dan beschikken over een rem waaraan getrokken kan worden. Het is de vraag of hier nu een afdoende juridisch mechanisme voor bestaat. Er wordt op dit punt met name gekeken naar de aankomende AI-verordening, waarover nu onderhandeld wordt. We bespreken wat er op het moment van schrijven bekend is.

Waarschijnlijk zal de AI-verordening vereisen dat alle *foundation models* pas de markt op mogen, zodra bepaalde aspecten geëvalueerd zijn, en het ontwikkelproces gedocumenteerd is. Is een GAI-toepassing eveneens te kwalificeren als een ‘hoogrisico-AI-systeem’, dan dient de ontwikkelaar een ‘conformiteitsbeoordeling’ uit te voeren. Een dergelijke beoordeling is ook onderdeel van Europese regels voor productveiligheid. Voordat het GAI-systeem aangeboden wordt, moet gecontroleerd worden of het systeem voldoet aan allerlei eisen ten aanzien van datakwaliteit, cybersecurity en documentatie.

Afhankelijk van het type systeem, mag de ontwikkelaar de beoordeling zelf uitvoeren, of moet een onafhankelijke derde partij dit doen. Voor welke *foundation models* dat laatste zou moeten gelden, en of er aanvullende eisen gesteld zouden moeten worden aan sommige van die modellen en GAI-toepassingen, wordt op het moment van schrijven nog bediscussieerd door de EU-beleidsmakers.[129]

---

[129] Op moment van schrijven zijn de triloogonderhandelingen over de wettekst nog gaande. In een document van de EU Raad van Ministers van oktober 2023, zijn de meest recente voorstellen te vinden. De Raad wil onafhankelijke beoordeling en strengere regels voor zeer capabele basismodellen’ of ‘hoog impact modellen’ (*foundation models*), zoals GPT-4. De criteria voor zulke modellen zijn nog niet gespecificeerd, maar er zou bijvoorbeeld gekeken kunnen worden naar rekenkracht, omvang van de trainingsdata en economische middelen van de aanbieder. Omdat zulke basismodellen ‘systeemrisico's’ (een DSA-term) met zich meebrengen, overweegt de wetgever de handhaving te centraliseren, bij de nog op te richten European Artificial Intelligence Board (EAIB). Naast de extra regels voor zeer capabele basismodellen, zijn aanvullende eisen voorgesteld voor AI-*toepassingen* op basis van foundation models, en die worden gebruikt door meer dan 10.000 zakelijke gebruikers of 45 miljoen particulieren. Het voorgaande zou dus een benadering inhouden die meer lijkt op de DSA en DMA. Zie Bertuzzi, 2023a, 2023b.

Echter, landen als Duitsland, Frankrijk en Italië hebben in november de regels voor *foundation models* bekritiseerd: die

Anders dan de productregelgeving, omvat de AI-conformiteitsbeoordeling ook mensenrechtelijke bescherming. Zo moet de documentatie de risico's voor fundamentele rechten beschrijven, en moeten allerlei maatregelen genomen worden om vooroordelen in trainingsdata te voorkomen. In het voorstel van het Europees Parlement is in aanvulling op de conformiteitsbeoordeling van aanbieders, een verplichte Fundamental Rights Impact Assessment (FRIA) opgenomen voor *exploitanten* van hoogrisico-(G)AI-systemen. Uit toepassingen kunnen immers grondrechtelijke risico's voortvloeien die bij de ontwikkelingsfase niet te voorzien zijn, zo redeneert het Europees Parlement.[130]

Naast de interne en externe controles voorafgaand aan de lancering van GAI-toepassingen, zal achteraf handhaving en controle door toezichthouders plaatsvinden. Zij beschikken over dezelfde instrumenten als onder productregelgeving. Als de toezichthouder onderzoek besluit te doen naar een conformiteitsbeoordeling en constateert dat de toepassing niet voldoet aan de gestelde normen, kan deze van de markt afgehaald worden. De introductie van de FRIA zou het mogelijk maken om toepassingen te verbieden waarvan de grondrechtelijke risico's onvoldoende geregistreerd en ingeperkt zijn door een exploitant.

In theorie zou de AI-verordening de samenleving dus de benodigde rem kunnen bieden. Maar hier zitten de nodige haken en ogen aan. Ten eerste zullen de meeste aanbieders en exploitanten van GAI-systemen nog niet de benodigde mensenrechtelijke expertise bezitten om hun producten aan de gestelde normen te laten voldoen, en dat zelf te toetsen. Ten tweede vereist een rem adequate handhaving. Naar verwachting zullen het grote aantal AI-aanbieders op de markt, en de vereiste coördinatie tussen toezichtsvelden (gegevensbescherming, marktregulering, productregulering, etc.) leiden tot een aanzienlijke verzwaring van de handhavingslast.

Ten derde zijn mensenrechtelijke eisen op verschillende manieren te interpreteren. Dit is een fundamenteler probleem. Aan een gasfles kunnen concrete en controleerbare veiligheidseisen gesteld worden, maar dit ligt bij abstractere mensenrechten ingewikkelder. Wanneer is bijvoorbeeld het risico op discriminatie tot een 'acceptabel' niveau teruggebracht? En, acceptabel voor wie? Er bestaat discussie over welke methoden gebruikt moet worden om discriminatie op te sporen, omdat die methoden samenhangen met verschillende ideeën over wat eerlijk is.[131] In de praktijk zal het antwoord gegeven worden door standaardisatieorganisaties als ISO en het Forum Standaardisatie, en door rechtszaken die waarschijnlijk zullen volgen.

Het is daarom zaak dat de politiek en beleidsmakers het vizier openhouden en gaandeweg afwegen of bovenstaande voorstellen de beste rem bieden. Alternatieven zijn denkbaar. Zo is het mogelijk om een vergunningsstelsel in te voeren, waarbij een

---

zouden start-ups teveel hinderen. De discussie tussen de EU-instituten over de regulering van GAI-systemen, loopt dan ook nog steeds. Als in december geen consensus bereikt wordt, kan de hele wet in gevaar komen, omdat het Europees Parlement ontbonden zal worden voor nieuwe EU-verkiezingen volgend jaar. Zie Bertuzzi, 2023c.

[130] Artikel 29 bis en overweging 58 bis AI-Verordening (EP-versie).

[131] Rathenau Instituut, 2022c

toezichthouder vooraf goedkeuring verleent. Ook zou de eis gesteld kunnen worden dat bij zeer impactvolle toepassingen de toezichthouder wordt geraadpleegd, naar voorbeeld van de al bestaande 'voorafgaande raadpleging' van de AVG. Sommigen zien meer in een DSA-achtige benadering voor grote GAI-modellen met veel gebruikers, waarbij jaarlijks onafhankelijke beoordelingen uitgevoerd worden, (mensenrechtelijke) risico's continu gemonitord worden door aanbieders en notice-and-action-mechanismes voor gebruikers verplicht zijn.[132] Tenslotte is het van belang normontwikkeling niet uitsluitend in handen te leggen van juridische processen en standaardiseringsorganisaties. Ook politiek en burgers kunnen zich uitspreken over de mensenrechtelijke eisen waaraan GAI zou moeten voldoen.

4.2. Zorg voor toekomstbestendige juridische kaders

Naast de AI-verordening zijn er ook een aantal andere juridische kaders die om aandacht vragen. Sommige kwesties gaan over een mogelijke verheldering of aanpassing van bestaande kaders. Andere vragen zijn fundamenteler: past het gekozen stelseltype nog bij de mogelijke impact van GAI op de samenleving? We vatten hieronder de belangrijkste kwesties samen, die uit onze verkenning naar voren kwamen.

Dataprotectie

De Algemene Verordening Gegevensbescherming is van toepassing als GAI-systemen persoonsgegevens verwerken. De vraag is nog wel of intieme gegevens voldoende beschermd worden. Zo kunnen GAI-systemen informatie verzamelen over iemands stemming ('sentimentanalyse'). Dit soort informatie valt doorgaans niet onder de categorie 'bijzondere persoonsgegevens' die in de AVG extra bescherming genieten. Een overweging kan zijn om deze categorie uit te breiden met sentiment, spraak, gezichts- en lichaamsgegevens, zodat de gegevens van individuen beter zijn beschermd. De verwachting is dat GAI-systemen steeds meer typen gevoelige data kunnen verwerken, waaronder informatie uit hersenscans. Daarom is er een discussie gestart over de vraag hoe vrijheid van gedachten (mentale privacy) beschermd kan worden.[133]

Discriminatie

Volgens Nederlands en Europees recht is discriminatie verboden. Sociale media staan echter vol met voorbeelden van GAI-systemen die vooroordelen uiten en stereotypen bevestigen. Dat kan leiden tot ongelijke behandeling. Ontwikkelaars moeten daarom bias terugdringen in hun systemen.[134] Ook de aankomende AI Act zal eisen stellen aan ontwikkelaars op het gebied van de kwaliteit van data en data governance. Het is de vraag of deze maatregelen discriminatie sterk kunnen verminderen. Wetenschappers geven aan dat de huidige technieken voor zorgvuldige datacuratie slechts mogelijk zijn bij kleinere datasets. Dat maakt het voorlopig lastig om GAI-modellen 'biasvrij' te maken. De overheid zal gezien deze technische beperkingen dus goed moeten nagaan

---

132 Hacker et al., 2023; Helberger & Diakopoulos, 2023

133 Zie ook de discussie over neurorechten in de Rathenau Scan over immersieve technologieën Rathenau Instituut, 2023b.

134 College voor de Rechten van de Mens, 2021; Van Bekkum & Zuiderveen Borgesius, 2023

welke anti-discriminatie-eisen ze stelt, en wat er moet gebeuren als GAI-systemen niet aan deze eisen kunnen voldoen.

Veiligheid
Generatieve AI-systemen kennen meerdere veiligheidsrisico's. Zo kunnen de systemen foutieve of misleidende informatie verspreiden, onvoorspelbare resultaten produceren, en misbruikt worden om desinformatie te verspreiden of cyberaanvallen uit te voeren. De afgelopen jaren is er in Nederland en Europa al veel geïnvesteerd in cyberveiligheid, onder meer door aangescherpte wet- en regelgeving en een wettelijk kader voor productveiligheid. Ook de aankomende AI-verordening zal naar verwachting eisen stellen aan de prestaties, interpreteerbaarheid en cybersecurity-aspecten van systemen. De vraag is of dit voldoende is om misbruik tegen te gaan. Verder betekent de relatieve geslotenheid van private ontwikkelaars dat wetenschappers en de samenleving weinig zicht hebben op hoe de taalmodellen ontwikkeld worden en welke vermogens ze verwerven. Welke mate van inzicht en controle acht de politiek noodzakelijk?

Desinformatie
Belangrijke juridische kaders om desinformatie tegen te gaan zijn o.a. de Wet inzake Digitale Diensten en de aankomende AI-verordening. Er bestaat discussie over de vraag of deze kaders voldoende zijn, omdat GAI-systemen het eenvoudiger maken overtuigende desinformatie te creëren – van geloofwaardige e-mails tot deepfakes.[135] Zo onderzoekt het Openbaar Ministerie de mogelijkheden om met het Wetboek van Strafrecht *deep nudes* aan te pakken, gemanipuleerde video's waarin mensen ongewild ontbloot zijn en seksuele handelingen verrichten. Een alternatieve optie is het verbieden van specifieke toepassingen van deepfakes, zoals in een pornografische context of het reguleren van deepfakes als hoogrisicotoepassing in de AI-verordening. De AI-verordening bevat nu alleen transparantieverplichtingen voor partijen die deepfakes genereren.

Auteursrecht
Het auteursrecht beschermt als onderdeel van het fundamenteel recht op eigendom makers van werken, met als uitgangspunt dat alleen zij hun werk mogen verveelvoudigen of beschikbaar kunnen maken voor het publiek. Er zijn vragen over auteursrechtelijk beschermd materiaal met betrekking tot de *input* (trainingsdata) en de *output* (de content die de systemen creëren). Met betrekking tot input is de kans groot dat de trainingsdata van taalmodellen auteursrechtelijk beschermd materiaal bevatten. Inmiddels lopen er diverse rechtszaken om te verhelderen of dat materiaal rechtmatig is verkregen. Een van de problemen is daarbij dat het auteursrecht zoals we dat nu kennen veel verantwoordelijkheid legt bij rechthebbenden: zij moeten zelf voorbehouden maken. Rechthebbenden kunnen echter niet altijd gemakkelijk achterhalen of hun werk gebruikt is om GAI-modellen te trainen. Creëert dit een

[135] We merken hierbij op dat het beleid om desinformatie tegen te gaan uit meer bestaat dan juridische kaders, zoals het investeren in kwaliteitsmedia en bewustwording.

omgeving die het maken van kunst en andere werken bevordert en eigendom voldoende beschermt?[136]

Het is de vraag of ontwikkelaars in het huidige juridische kader voldoende worden ontmoedigd om auteursrechtelijk beschermd werk illegaal te gebruiken. Bedenk dat het kwaad al snel is geschied, omdat modellen eenmaal geleerde kwaliteiten en stijlen niet zomaar 'ontleren'.[137] Het is onduidelijk hoe het auteursrecht zich zou moeten verhouden tot GAI-systemen die bestaand werk imiteren, maar voldoende afwijken om niet onder het auteursrecht te vallen. De facto gaan makers zo concurreren met GAI-systemen. Het is zaak dat beleidsmakers en politici zich afvragen wat voor auteursrecht ze willen: een AI-vriendelijk auteursrecht, of een juridisch kader dat de makers bescherming biedt?

Mededingingsrecht

Diverse grote technologiebedrijven hebben de afgelopen jaren een sterke positie opgebouwd op het gebied van taalmodellen, maar ook op het gebied van sociale media, zoekmachines en cloudinfrastructuur. Wetenschappers vragen zich af of het huidige mededingingsrecht voldoende is toegerust om de economische macht van deze conglomeraten in te dammen en of de recent ingevoerde Wet inzake Digitale Markten (DMA) daar daadwerkelijk verandering in weet brengen.[138] Een fundamentelere kwestie is dat wetenschappers aangeven dat technologiebedrijven niet alleen over economische macht beschikken, maar ook invloed kunnen uitoefenen op sociale en politieke kwesties. Ook publieke domeinen raken afhankelijker van de diensten die door deze bedrijven worden aangeboden. Deze problematiek vergt wellicht meer dan het creëren van condities voor 'eerlijke markten' – de tot nu toe gevolgde beleidsstrategie.[139] In Nederland is daarom een onderzoeksgroep gestart die onderzoekt in hoeverre dit vraagstuk vanuit staats- en bestuursrecht kan worden benaderd.[140]

Tot slot: de rol van toezichthouders

De doorlichting van de hierboven beschreven kaders zal tijd kosten, en daarom is het van belang dat in de tussentijd wettelijke kaders die al van kracht zijn, goed worden nageleefd. Toezichthouders spelen hierin met handhaving een cruciale rol, en het is belangrijk dat zij zich ten opzichte van GAI assertief opstellen. Beleidsmakers en politici hebben op hun beurt een rol in het creëren van de juiste voorwaarden, met name door toezichthouders van voldoende middelen te voorzien, en na te gaan hoe de expertise van toezichthouders verder versterkt kan worden.[141]

---

136 Visser, 2023

137 Wong, 2023

138 Zo grijpt DMA beperkt in op de onderliggende machtsbronnen, waaronder computerkracht, early mover voordelen, databronnen en geïntegreerde systemen. Ook zitten onder de nu aangemerkte poortwachters en 'kernplatformdiensten' geen GAI-diensten, behalve waar GAI geïntegreerd is in zoekmachines. De DMA bevat bepalingen die – mits van toepassing verklaard op GAI-diensten of ontwikkelaars – een eerlijke markt kunnen bevorderen. Yasar et al., 2023.

139 Gerbrandy & Phoa, 2022; Nemitz, 2018; Passchier, 2021; Sharon & Gellert, 2023

140 Jak & Lokin, 2023

141 Een Europese werkgroep van toezichthouders bereidt zich nu al voor op de AI-verordening. De werkgroep wordt voorgezeten door de Nederlandse Rijksinspectie Digitale Infrastructuur.

4.3. Investeer in internationaal AI-beleid, om mondiale innovatieprocessen van technologiebedrijven bij te sturen

Vanwege het grensoverschrijdende karakter van generatieve AI en haar ontwikkelaars, is het belangrijk dat Nederland internationale samenwerking bevordert als onderdeel van een strategie om grip op de technologie te krijgen. De hoop is dat de AI-verordening, evenals andere Europese wetten, internationaal doorwerken, waarbij actoren in het buitenland EU-wetgeving als uitgangspunt hanteren. In dat verband wordt ook wel gesproken van het 'Brussels Effect'.[142] Dit effect zou op kunnen gaan voor het AI-verdrag dat nu gefinaliseerd wordt bij de Raad van Europa. Dat verdrag is bedoeld om burgers meer rechtsbescherming bieden, bijvoorbeeld door burgers die geraakt worden door AI-beslissingen bezwaar- en verhaalsmogelijkheden te geven. Rechterlijke uitspraken over generatieve AI kunnen eveneens een mondiale werking hebben, zie de uitspraak van het Europees Hof van Justitie in *Schrems I*.[143]

Het is op dit moment te vroeg om te zeggen of er van deze Europese afspraken inderdaad een dergelijke internationale werking uit zal gaan. Wel is duidelijk dat ook de Verenigde Staten en China zich buigen over regelgeving. Zo hebben de VS onlangs besloten dat GAI-toepassingen die door de overheid gebruikt worden aan strengere voorwaarden moeten voldoen.[144]

Maar zelfs als andere landen zich laten inspireren door Europese wetgeving, is het belangrijk te komen tot internationale afspraken, zodat bedrijven overal ter wereld aangesproken kunnen worden op hun verantwoordelijkheid. De WRR adviseerde eerder dat Nederland werk zou moeten maken van 'AI-diplomatie'.[145] Het afgelopen jaar zijn met name vrijwillige codes opgesteld. De OESO heeft het *Global Partnership on AI* (GPIA) opgestart, alsook het *Partnership on AI* waarin industrie, wetenschapers, ngo's en mediaorganisaties samenkomen. Ook heeft de OESO een inhoudelijk sterk kader op het gebied van maatschappelijk verantwoord ondernemen opgesteld, dat kwesties als uitbuiting en milieu-impact adresseert. Onlangs heeft de G7 een vrijwillige gedragscode afgesproken voor de ontwikkelaars van GAI, en heeft de Europese Commissie aangegeven samen met de VS tot vrijwillige gedragscodes te willen komen voor de bedrijven achter GAI-modellen.[146]

Er is ook een begin gemaakt met bindende afspraken. UNESCO heeft het mondiale kader *Ethics of Artificial Intelligence* opgesteld, waar landen die het hebben ondertekend aan gehouden zijn.[147] Verschillende landen introduceren daarnaast wetgeving op het gebied van maatschappelijk verantwoord ondernemen. In Europa gaat het om de Richtlijn *zorgvuldigheid in het bedrijfsleven op het gebied van duurzaamheid* (*Corporate sustainability due diligence*). Hoe meer die wetgeving in lijn is

142 Deze term is afkomstig van de Fins Amerikaanse hoogleraar rechten Anu Bradford, zie Bradford, 2020.
143 *Maximillian Schrems v Data Protection Commissioner*, 2015
144 Lima & Zakrzewski, 2023
145 Dit kan betrekking hebben op vijf domeinen: fundamenteel onderzoek, commerciële toepassingen, regulering, ethische richtlijnen en standaarden
146 Zubascu, n.d.
147 Sabzalieva & Valentini, 2023

met de vastgestelde OESO richtlijn op dit vlak, hoe meer bedrijven en risico's worden afgedekt.

4.4. Stel een ambitieuze agenda op voor maatschappelijk verantwoorde GAI

Om grip te krijgen op generatieve AI en wenselijke toepassingen te realiseren, zal het niet voldoende zijn om juridische kaders te versterken. De overheid zal innovatie voor maatschappelijk verantwoorde AI ook met andere beleidsinstrumenten actief moeten aanmoedigen. Hiervoor is een ambitieuze agenda nodig, die tenminste twee elementen omvat:

Investeer in alternatieve technologie

In hoofdstuk 3 lieten we zien dat het risicovol is als de ontwikkeling van GAI-technologie louter in handen ligt van enkele grote technologiebedrijven.[148] Er zijn verschillende manieren om andere partijen een betekenisvolle rol te geven, zodanig dat er een meer democratische sturing op technologie mogelijk wordt. Zo zouden overheden zowel op nationaal als Europees niveau de ontwikkeling van GAI-technologie mede kunnen financieren met maatschappelijke partners.[149] Een recent voorbeeld is het aangekondigde project GPT-NL, waarbij TNO, het Nederlands Forensisch Instituut en ICT-Coöperatie SURF samen optrekken.[150] Het ministerie van Economische Zaken en Klimaat (EZK) heeft 13,5 miljoen euro beschikbaar gesteld. Dit model zou geschikt moeten zijn voor gebruik binnen de academische wereld en de overheid. Als geldverstrekker kan de overheid er makkelijker op toezien dat juridische en ethische standaarden bij de ontwikkeling gerespecteerd worden. Men kan op dit punt de vergelijking trekken met andere publiek gefinancierde technologische trajecten, zoals CERN, dat onderzoek doet naar elementaire deeltjes, of Europese publiek private samenwerkingen in ruimtevaartprogramma's – waarin Europa grote ambities heeft weten te realiseren.

Ook kan ingezet worden op *open source* ontwikkeling van GAI-technologie. Duitse en Franse overheden werken al samen met het platform *Nextcloud*. Het is zaak hierbij kritisch te zijn op de mate waarin producten daadwerkelijk *open source* zijn. Zo constateerden we al dat het taalmodel Llama 2 slechts een beperkte mate van openheid geeft. De overheid kan in eigen aanbestedingstrajecten de voorkeur geven aan *open source*, en duidelijk maken wat daar in de context van GAI precies mee wordt bedoeld.[151]

Het is van belang om als overheid te investeren in wetenschap die er voor zorgt dat GAI-technologie vanuit een publieke waardenperspectief vorm kan krijgen. Denk hierbij aan het stimuleren van onderzoek naar taalmodellen voor kleinere taalgebieden en dialecten, veiligheid, synthetische data, biasdetectie bij grotere datasets en duurzame

---

148 Europa heeft zich voorgenomen om "strategische digitale autonomie" op te bouwen, zowel ten opzichte van landen als China en de Verenigde Staten als mondiale technologiereuzen, zie European Parliament, 2022; European Union External Action, 2020. Het opbouwen van publieke GAI-alternatieven zou hierbij aansluiten.

149 Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2023. Over Gaia project klinken diverse zorgen of de oorspronkelijk doelstelling (autonomie) niet verwaterd is geraakt, zie bijvoorbeeld Goujard & Cerulus, 2021

150 Digitale Overheid, 2023

151 Rijksoverheid, 2022a

GAI-technologie. Ook is het zaak fundamenteel onderzoek naar de interpreteerbaarheid van modellen te organiseren. Tot slot is het belangrijk om technologische tunnelvisie te voorkomen, door te onderzoeken of en hoe andere AI-technologieën de tekortkomingen van GAI kunnen adresseren.

Monitor en begeleid ontwikkelingen per sector
Gezien de brede maatschappelijke impact van GAI is het zaak die impact in de praktijk te monitoren en op ongewenste effecten te anticiperen. Hoe verandert deze technologie de dynamiek in het klaslokaal of de administratie van een ziekenhuis? Welke aanpassingen vraagt dat bijvoorbeeld in het curriculum of het beleid van onderwijsinstellingen? Hoe zwaar is de milieubelasting van GAI in de praktijk? En, welke veranderingen vinden plaats op de arbeidsmarkt, bijvoorbeeld met betrekking tot loonpolarisatie of inkomens- en vermogensverdeling? Elke sector zal zich moeten gaan buigen over hoe GAI echt kan bijdragen aan het realiseren van de wensen en ambities van professionals. En net zo belangrijk: als generatieve AI-toepassingen de gewenste resultaten niet leveren, is het zaak dat mensen dit tijdig signaleren en bij beleidsmakers aankaarten.

4.5. Stimuleer maatschappelijk debat over de wenselijkheid van GAI

Het is cruciaal om naast maatregelen op het gebied van wetgeving en stimulering het maatschappelijk debat over GAI te entameren. Gebruikers zien GAI vaak als een handige, onschuldige technologie waar iedereen mee kan experimenteren. Maar het gebruik van GAI-toepassingen gaat gepaard met risico's.

Die risico's vragen om technologisch burgerschap: mensen moeten zich bewust zijn van de gevaren, in staat zijn om met risico's om te gaan en mee kunnen doen aan de democratische besluitvorming over GAI-technologie.[152] Dit burgerschap begint met voorlichting: het is voor iedere burger en organisatie van belang om te weten wat voor technologie GAI eigenlijk is, wat je er wel en niet van kan verwachten en met welke risico's het gepaard gaat.

Vervolgens zal iedere organisatie, ieder bedrijf, iedere wetenschapper en iedere gebruiker zich de vraag moeten stellen onder welke voorwaarden zij GAI-technologie in willen zetten, en of dat nu maatschappelijk verantwoord kan. De AI-verordening is er nog niet, toezichthouders zijn nog bezig hun rol vorm te geven en er zijn veel onopgeloste kwesties, onder meer ten aanzien van veiligheid, discriminatie, uitlegbaarheid, dataprotectie, uitbuiting en duurzaamheid. Er zijn daarom al diverse oproepen geweest om de technologie niet te gebruiken, bijvoorbeeld van wetenschappers die vinden dat het gebruik van GAI niet te rijmen valt met hun maatschappelijke verantwoordelijkheid.[153] Ook heeft UNESCO opgeroepen om GAI-technologie in het onderwijs voor kinderen onder de 13 niet te gebruiken.[154]

---

152 Zie voor nadere uitwerking van technologisch burgerschap Est, 2020.
153 Rooij, 2023
154 UNESCO, 2023

Bovendien is het van belang dat in het debat over GAI alle relevante vragen en kwesties aan bod komen. Denk bijvoorbeeld aan mogelijke liefdesrelaties tussen mensen en chatbots of interacties tussen kinderen en chatbots: wat vinden we daarbij wenselijk? Het is illustratief om GAI te vergelijken met de opkomst van mobiele telefoons. De apparaten zijn multifunctioneel en om verschillende redenen zo aantrekkelijk dat ze vrijwel onontkoombaar zijn geworden. Maar inmiddels bestaan er zorgen over de impact van mobiele telefoons op onze fysieke en mentale gezondheid. Als je in de trein eenieder op zijn telefoon ziet turen, besef je dat technologie ook iets kostbaars van mensen af kan nemen. Nu computers met ons kunnen spreken, ons op ons gemak kunnen stellen en ons intellectueel kunnen prikkelen, is het des te belangrijker om te doordenken hoe we als samenleving onze tijd met en zonder AI-systemen door willen brengen.

Scholen, ziekenhuizen, bibliotheken, kunstcollectieven en maatschappelijke organisaties zijn allemaal aan zet om technologisch burgerschap te vergroten en het debat over GAI verder op gang te brengen. De overheid kan hierbij een stimulerende rol spelen. Het Rathenau Instituut zal de komende vier jaar een dialoogprogramma organiseren waarbij burgers in gesprek gaan over de toekomst van de digitale samenleving. GAI zal binnen dit programma één van de gespreksonderwerpen zijn.

4.6 Conclusie

Gedurende dit onderzoek merkten we op dat de opvattingen onder experts over de mogelijkheden en gevaren van GAI-technologie sterk uiteenlopen. De ene expert ziet GAI-technologie als een *game changer* die de vermogens van AI-systemen razendsnel zal vergroten. De ander wijst op de inherente beperkingen van de technologie, en stelt dat GAI-systemen ‘een verkeerd paradigma’ van AI inluiden.[155] Deze uiteenlopende opvattingen zie je ook terug in de discussie over de risico’s van GAI: draait het om meer klassieke kwesties rond privacy, discriminatie en veiligheid, of vormen deze systemen een existentieel risico voor de mensheid?

Wij trekken de volgende conclusies. GAI-technologie vormt wel degelijk een doorbraak in het vermogen van AI-systeem om taaltaken uit te voeren en verschillende modaliteiten met elkaar te combineren. Het is tegelijkertijd nog zeer de vraag hoe goed de toepassingen van deze technologie nou precies zijn, en in hoeverre mensen taken uit handen kunnen geven. Daarnaast versterkt GAI de reeds bekende risico’s van de digitale samenleving en voegt het nieuwe toe, zoals verlies van eigenaarschap en de verstoring van menselijke ontwikkeling. Deze risico’s lijken niet eenvoudig op te lossen. Het is een reële mogelijkheid dat het bestaande en aangekondigde beleid onvoldoende zijn opgewassen tegen deze risico’s.

Actie is daarom nodig. De overheid en politici moeten toetsen waar beleid aanscherping behoeft. In de tussentijd dienen zij toezicht maximaal te ondersteunen, afspraken te maken met ontwikkelaars en de samenleving te waarschuwen voor de risico’s van GAI.

---

155 Marcus, 2020

Wereldwijd worden de risico’s van GAI-systemen serieus genomen; dat zou iedere Nederlandse burger en instantie ook moeten doen.

Bijlage: Geraadpleegde beleidsmakers en experts

Geïnterviewde experts

1. Lambèr Royakkers, hoogleraar Ethics of the Digital Society aan de Technische Universiteit Eindhoven
2. Pim Haselager, hoogleraar Artificiële Intelligentie aan de Radboud Universiteit
3. Haroon Sheikh, senior onderzoeker bij de WRR en bijzonder hoogleraar Strategic Governance of Global Technologies aan de Vrije Universiteit Amsterdam
4. Anna Gerbrandy, hoogleraar Mededingingsrecht aan de Universiteit Utrecht en kroonlid van de Sociaal-Economische Raad.
5. Naomi Appelman, PhD onderzoeker bij het Institute for Information Law van de Universiteit van Amsterdam
6. Catelijne Muller, voorzitter van de Alliantie voor AI en lid van de High Level Expert Group on AI van de Europese Commissie
7. Lokke Moerel, hoogleraar Global ICT Law aan Tilburg University en Senior of Counsel bij Morrison & Foerster, lid van de Cyber Security Raad
8. Jeroen van den Hoven, universiteitshoogleraar en hoogleraar Ethiek en Technologie aan TU Delft en hoofdredacteur Ethiek en Informatietechnologie

Deelnemers werksessie met beleidsmakers

1. Francisca Wals, ministerie van Binnenlandse Zaken en Koninkrijksrelaties
2. Jasper Kars, ministerie van Binnenlandse Zaken en Koninkrijksrelaties
3. Haye Hazenberg, ministerie van Binnenlandse Zaken en Koninkrijksrelaties
4. Elja Daae, ministerie van Binnenlandse Zaken en Koninkrijksrelaties
5. David van Es, ministerie van Binnenlandse Zaken en Koninkrijksrelaties
6. Anne Thier, ministerie van Binnenlandse Zaken en Koninkrijksrelaties
7. Luca Kuiper, ministerie van Buitenlandse Zaken
8. Mijntje Jansen, ministerie van Economische Zaken en Klimaat
9. Gelijn Werner, ministerie van Economische Zaken en Klimaat
10. Noud Louwers, ministerie van Economische Zaken en Klimaat
11. Vincent Pot, ministerie van Onderwijs, Cultuur en Wetenschap
12. Cyril van der Net, Ministerie van Justitie en Veiligheid

Geraadpleegde technologie-experts (meegelezen met hoofdstuk 1)

1. Eric Postma, hoogleraar Artificiële Intelligentie, Tilburg University
2. Frank van Harmelen, hoogleraar Knowledge Representation & Reasoning, Vrije Universiteit Amsterdam

Verder wil het projectteam Joost Gerritsen zeer bedanken voor het uitgevoerde juridische feitenonderzoek en het tegenlezen van hoofdstukken drie en vier.

Literatuurlijst

Abid, A., Farooqi, M., & Zou, J. (2021). *Persistent Anti-Muslim Bias in Large Language Models* (arXiv:2101.05783). arXiv. https://doi.org/10.48550/arXiv.2101.05783

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, *28*(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Al-Hawawreh, M., Aljuhani, A., & Jararweh, Y. (2023). Chatgpt for cybersecurity: practical applications, challenges, and future directions. *Cluster Computing*, *26*(6), 3421–3436. https://doi.org/10.1007/s10586-023-04124-5

Alshurafat, H. (2023). *The usefulness and challenges of chatbots for accounting professionals: application on ChatGPT* (SSRN Scholarly Paper No. 4345921). https://doi.org/10.2139/ssrn.4345921

Ananthaswamy, A. (2023). In AI, is bigger always better? *Nature*, *615*(7951), 202–205. https://doi.org/10.1038/d41586-023-00641-w

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI regulation. Managing emerging risks to public safety* (arXiv:2307.03718). arXiv. https://doi.org/10.48550/arXiv.2307.03718

Andersen, R. (2023, 24 July). Does Sam Altman know what he's creating? The OpenAI CEO's ambitious, ingenious, terrifying quest to create a new form of intelligence. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2023/09/sam-altman-openai-chatgpt-gpt-4/674764/

Anderson, C. (2008, 23 June). The end of theory. The data deluge makes the scientific method obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Anthropic. (2022). *Constitutional AI. Harmlessness from AI feedback*. https://www.anthropic.com/index/constitutional-ai-harmlessness-from-ai-feedback

Apple. (n.d.). *Replika. Virtual AI companion on the App Store.* Retrieved November 14, 2023, from https://apps.apple.com/us/app/replika-virtual-ai-companion/id1158555867

Autor, D. (2022). *The labor market impacts of technological change. From unbridled enthusiasm to qualified optimism to vast uncertainty* (Working Paper No. 30074). National Bureau of Economic Research. https://doi.org/10.3386/w30074

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and Artificial Intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*. https://doi.org/10.1001/jamainternmed.2023.1838

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI. Harmlessness from AI feedback* (arXiv:2212.08073). arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative AI. A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260. https://doi.org/10.3390/fi15080260

Baughman, J. (2023). *China's ChatGPT war*. China Aerospace Studies Institute. https://www.airuniversity.af.edu/Portals/10/CASI/documents/Research/Cyber/2023-08-21%20China's%20ChatGPT%20War.pdf

Bell, G., Burgess, J., Thomas, J., & Sadiq, S. (2023). *Rapid response information report. Generative AI*. Australia's Chief Scientist. https://www.chiefscientist.gov.au/sites/default/files/2023-

06/Rapid%20Response%20Information%20Report%20-%20Generative%20AI%20v1_1.pdf
Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
Bender, E. M., & Koller, A. (2020). Climbing towards NLU. On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463
Bertuzzi, L. (2023a, 17 October). *AI Act. EU countries headed to tiered approach on foundation models amid broader compromise*. EURACTIV. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-countries-headed-to-tiered-approach-on-foundation-models-amid-broader-compromise/
Bertuzzi, L. (2023b, 25 October). EU policymakers enter the last mile for Artificial Intelligence rulebook. *EURACTIV*. https://www.euractiv.com/section/artificial-intelligence/news/eu-policymakers-enter-the-last-mile-for-artificial-intelligence-rulebook/
Bertuzzi, L. (2023c, 10 November). *EU's AI Act negotiations hit the brakes over foundation models*. Www.Euractiv.Com. https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/
Bianchi, F., & Hovy, D. (2021). On the gap between adoption and understanding in NLP. *Findings of the Association for Computational Linguistics 2021*, 3895–3901. https://doi.org/10.18653/v1/2021.findings-acl.340
Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, *5*(5), 277–280. https://doi.org/10.1038/s42254-023-00581-4
Blodgett, S. L., & Madaio, M. (2021). *Risks of AI foundation models in education* (arXiv:2110.10024). arXiv. https://doi.org/10.48550/arXiv.2110.10024
Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the opportunities and risks of foundation models* (arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258
Bowman, S. R. (2023). *Eight things to know about Large Language Models* (arXiv:2304.00612). arXiv. https://doi.org/10.48550/arXiv.2304.00612
Bozeman, B. (2007). *Public values and public interest. Counterbalancing economic individualism*. Georgetown University Press. https://www.jstor.org/stable/j.ctt2tt37c
Bradford, A. (2020). *The Brussels effect. How the European Union rules the world*. Oxford University Press. https://scholarship.law.columbia.edu/books/232
Brewster, J., Wang, M., & Palmer, C. (2023, 24 August). Plagiarism-bot? How low-quality websites are using ai to deceptively rewrite content from mainstream news outlets. *NewsGuard*. https://www.newsguardtech.com/misinformation-monitor/august-2023
Bronzwaer, S. (2023, 9 January). Van reclameteksten schrijven tot computercode checken. Waar wordt ChatGPT al voor gebruikt? *NRC*. https://www.nrc.nl/nieuws/2023/01/09/van-reclameteksten-schrijven-tot-computercode-checken-waar-wordt-chatgpt-al-voor-gebruikt-a4153682
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., …

Amodei, D. (2020). *Language models are few-shot learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bruijn, H. D., & Dicke, W. (2006). Strategies for safeguarding public values in liberalized utility sectors. *Public Administration*, *84*(3), 717–735. https://doi.org/10.1111/j.1467-9299.2006.00609.x

Brynjolfsson, E. (2022, 12 January). The Turing trap. The promise & peril of human-like artificial intelligence. *Stanford Digital Economy Lab*. https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-artificial-intelligence/

Brynjolfsson, E., Li, D., & Raymond, L. (2023). *Generative AI at work* (SSRN Scholarly Paper No. 4426942). https://papers.ssrn.com/abstract=4426942

Callaway, E. (2022). Scientists are using AI to dream up revolutionary new proteins. *Nature*, *609*(7928), 661–662. https://doi.org/10.1038/d41586-022-02947-7

Cardon, P., Fleischmann, C., Aritz, J., Logemann, M., & Heidewald, J. (2023). The challenges and opportunities of AI-assisted Writing. Developing AI literacy for the AI age. *Business and Professional Communication Quarterly*, *86*(3), 257–295. https://doi.org/10.1177/23294906231176517

Chen, Z., Qing, J., Xiang, T., Yue, W. L., & Zhou, J. H. (2023). *Seeing beyond the brain. Conditional diffusion model with sparse masked modeling for vision decoding* (arXiv:2211.06956). arXiv. https://doi.org/10.48550/arXiv.2211.06956

Chohan, U. W. (2023). *Generative AI, ChatGPT, and the future of jobs* (SSRN Scholarly Paper No. 4411068). https://doi.org/10.2139/ssrn.4411068

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). *ChatGPT goes to law school* (SSRN Scholarly Paper No. 4335905). https://doi.org/10.2139/ssrn.4335905

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). *Economic potential of generative AI. The next productivity frontier*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

College voor de Rechten van de Mens. (2021). *Discriminatie door risicoprofielen. Een mensenrechtelijk toetsingskader*. https://publicaties.mensenrechten.nl/publicatie/61a734e65d726f72c45f9dce

Coscarelli, J. (2023, 19 April). An A.I. hit of fake 'Drake' and 'The Weeknd' rattles the music world. *The New York Times*. https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html

Cyber Security Raad. (2021). *Nederlandse digitale autonomie en cybersecurity. Hoe verminderen we onze digitale afhankelijkheden met behoud van een open economie?* (Beleidsnota No. 3; CSR Advies 2021). Ministerie van Justitie en Veiligheid. https://www.cybersecurityraad.nl/documenten/adviezen/2021/05/14/csr-advies-nederlandse-digitale-autonomie-en-cybersecurity---csr-advies-2021-nr.-3

Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & Society*, *34*(1), 129–136. https://doi.org/10.1007/s00146-017-0773-9

Danaher, J. (2020). Robot betrayal. A guide to the ethics of robotic deception. *Ethics and Information Technology*, *22*(2), 117–128. https://doi.org/10.1007/s10676-019-09520-3

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA. Efficient finetuning of quantized LLMs* (arXiv:2305.14314; Version 1). arXiv. http://arxiv.org/abs/2305.14314

Digitale Overheid. (2023, 6 November). *Nederland bouwt eigen open taalmodel GPT-NL*. https://www.digitaleoverheid.nl/nieuws/nederland-bouwt-eigen-open-taalmodel-gpt-nl/

Doorenbosch, T. (2023, 1 September). ChatGPT opent nieuwe wereld voor scriptkiddies en meer kwaadwillenden. *AG Connect*. https://www.agconnect.nl/artikel/chatgpt-opent-nieuwe-wereld-voor-scriptkiddies-en-meer-kwaadwillenden

Edwards, B. (2023, 15 March). OpenAI checked to see whether GPT-4 could take over the world. *Ars Technica*. https://arstechnica.com/information-technology/2023/03/openai-checked-to-see-whether-gpt-4-could-take-over-the-world/

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs. An early look at the labor market impact potential of Large Language Models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

Epstein, Z., Hertzmann, A., Herman, L., Mahari, R., Frank, M. R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., Farid, H., Leach, N., Pentland, A., & Russakovsky, O. (2023). Art and the science of generative AI. A deeper dive. *Science*, *380*(6650), 1110–1111. https://doi.org/10.1126/science.adh4451

Est, R. van. (2020). Technologisch burgerschap als dé democratische uitdaging van de eenentwintigste eeuw. *Christen Democratische Verkenningen*, *2016*(3). https://www.tijdschriftcdv.nl/inhoud/tijdschrift_artikel

European Commission. (2020). *Commission Work Programme 2020*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0037&from=ES

European Commission High Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament. (2022). *EU strategic autonomy 2013-2023. From concept to capacity*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733589/EPRS_BRI(2022)733589_EN.pdf

European Parliamentary Research Service, Scientific Foresight Unit, & Panel for the Future of Science and Technology. (2021). *Tackling deepfakes in European policy European Parliament*. (auteurs: Van Huijstee, M., Van Boheemen, P., Das, D., Nierling, L.,Jahnel, J., Karaboga & M., Fatun, M.). https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)690039

European Union External Action. (2020, 8 June). *For a united, resilient and sovereign Europe (with Thierry Breton)*. https://www.eeas.europa.eu/eeas/united-resilient-and-sovereign-europe-thierry-breton_und_en

Europol Innovation Lab. (2023). *ChatGPT. The impact of Large Language Models on Law Enforcement*. https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement#downloads

Fagone, J. (2021, 23 July). *He couldn't get over his fiancee's death. So he brought her back as an A.I. chatbot*. The San Francisco Chronicle. https://www.sfchronicle.com/projects/2021/jessica-simulation-artificial-intelligence/

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT. Implications for educational practice and research. *Innovations in Education and Teaching International*, 1–15. https://doi.org/10.1080/14703297.2023.2195846

Felsenthal, E., & Perrigo, B. (2023, 21 June). OpenAI CEO Sam Altman is pushing past doubts on Artificial Intelligence. *TIME*. https://time.com/collection/time100-companies-2023/6284870/openai-disrupters/

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). *From pretraining data to language models to downstream tasks. Tracking the trails of political biases leading to*

unfair NLP Models (arXiv:2305.08283). arXiv. https://doi.org/10.48550/arXiv.2305.08283

Field, H. (2023, 29 June). The first fully A.I.-generated drug enters clinical trials in human patients. *CNBC*. https://www.cnbc.com/2023/06/29/ai-generated-drug-begins-clinical-trials-in-human-patients.html

Floridi, L. (2023). AI as agency without intelligence. On ChatGPT, Large Language Models, and other generative models. *Philosophy & Technology*, *36*(1), 15. https://doi.org/10.1007/s13347-023-00621-y

Geng, J., Huang, D., & De la Torre, F. (2022). *DensePose from WiFi* (arXiv:2301.00250). arXiv. https://doi.org/10.48550/arXiv.2301.00250

Gerbrandy, A., & Phoa, P. (2022). The Power of Big Tech Corporations as Modern Bigness and a Vocabulary for Shaping Competition Law as Counter-Power. In H. Brouwer, M. Bennett, & R. Claassen, *Wealth and Power* (1st ed., pp. 166–185). Routledge. https://doi.org/10.4324/9781003173632-11

Gmyrek, P., Berg, J., & Bescond, D. (2023). *Generative AI and jobs. A global analysis of potential effects on job quantity and quality*. International Labour Organization. https://www.ilo.org/global/publications/working-papers/WCMS_890761/lang--en/index.htm

Goujard, C., & Cerulus, L. (2021, 26 October). Inside Gaia-X. How chaos and infighting are killing Europe's grand cloud project. *POLITICO*. https://www.politico.eu/article/chaos-and-infighting-are-killing-europes-grand-cloud-project/

Gownder, J. P., & O'Grady, M. (2023). *Forrester's 2023 generative AI jobs impact forecast, US*. Forrester. https://www.forrester.com/report/forresters-2023-generative-ai-jobs-impact-forecast-us/RES179790

Grünebaum, A., Chervenak, J., Pollet, S. L., Katz, A., & Chervenak, F. A. (2023). The exciting potential for ChatGPT in obstetrics and gynecology. *American Journal of Obstetrics and Gynecology*, *228*(6), 696–705. https://doi.org/10.1016/j.ajog.2023.03.009

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT. Impact of generative AI in cybersecurity and privacy. *IEEE Access*, *11*, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Gurnee, W., & Tegmark, M. (2023). *Language models represent space and time* (arXiv:2310.02207). arXiv. https://doi.org/10.48550/arXiv.2310.02207

Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other Large Generative AI Models* (arXiv:2302.02337). arXiv. https://doi.org/10.48550/arXiv.2302.02337

Haggart, B. (2023, 23 January). ChatGPT strikes at the heart of the scientific world view. *Centre for International Governance Innovation*. https://www.cigionline.org/articles/chatgpt-strikes-at-the-heart-of-the-scientific-world-view/

Hao, K., & Seetharaman, D. (2023, 24 July). Cleaning up ChatGPT takes heavy toll on human workers. *Wall Street Journal*. https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483

Harrer, S. (2023). Attention is not all you need. The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, *90*. https://doi.org/10.1016/j.ebiom.2023.104512

Hartmann, J., Schwenzow, J., & Witte, M. (2023). *The political ideology of conversational AI. Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation* (arXiv:2301.01768). arXiv. https://doi.org/10.48550/arXiv.2301.01768

Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1682

Hugenholtz, P. B., & Quintais, J. P. (2021). Copyright and artificial creation. Does EU copyright law protect AI-assisted output? *International Review of Intellectual Property and Competition Law*, *52*, 1190–1216. https://doi.org/10.1007/s40319-021-01115-0

Ingram, D. (2023, 14 January). AI Chat used by mental health tech company in experiment on real users. *NBC News*. https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110

Jak, N., & Lokin, M. (2023, 1 May). Big Tech als publiek belang vraagt om publieke waarborgen. *iBestuur*. https://ibestuur.nl/artikel/big-tech-als-publiek-belang-vraagt-om-publieke-waarborgen/

Jakkal, V. (2023, 19 October). *Microsoft Security Copilot Early Access Program is now available*. Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2023/10/19/microsoft-security-copilot-early-access-program-harnessing-generative-ai-to-empower-security-teams/

Janssen, B. V., Kazemier, G., & Besselink, M. G. (2023). The use of ChatGPT and other large language models in surgical science. *BJS Open*, *7*(2), zrad032. https://doi.org/10.1093/bjsopen/zrad032

Jeon, J., & Lee, S. (2023). Large language models in education. A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-11834-1

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and applications of Large Language Models* (arXiv:2307.10169). arXiv. https://doi.org/10.48550/arXiv.2307.10169

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* (arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kelsey, T. (2023a). *Dream recording through non-invasive brain-machine interfaces and generative AI-assisted multimodal software* (arXiv:2304.09858). arXiv. https://doi.org/10.48550/arXiv.2304.09858

Kelsey, T. (2023b, 6 May). *Brain Interface > GenAI > Recording Dreams?* OpenAI Developer Forum. https://community.openai.com/t/brain-interface-genai-recording-dreams/194795

Kenniscentrum voor beleid en regelgeving. (n.d.). *Beleidskompas*. Retrieved November 14, 2023, from https://www.kcbr.nl/beleid-en-regelgeving-ontwikkelen/beleidskompas

Kerssens, N., & van Dijck, J. (2023). Transgressing local, national, global spheres: the blackboxed dynamics of platformization and infrastructuralization of primary education. *Information, Communication & Society*, *0*(0), 1–17. https://doi.org/10.1080/1369118X.2023.2257293

Knight, W. (2023). The last AI boom didn't kill jobs. Feel better? *Wired*. https://www.wired.com/story/fast-forward-the-last-ai-boom-didnt-kill-jobs/

Koncz, A. (2023, 7 February). 5 AI tools that could help huild a medical practice. *The Medical Futurist*. https://medicalfuturist.com/5-ai-tools-that-could-help-build-a-medical-practice/

Koning, P. (2013). Activerend arbeidsmarktbeleid. Een beknopte handleiding. *TPEdigitaal*, *7*(2), 60–66.

https://www.tpedigitaal.nl/sites/default/files/bestand/activerend_arbeidsmarktbeleid_een_beknopte_handleiding.pdf
Korteweg, N. (2023, 2 May). ChatGPT verslaat artsen bij het geven van antwoord op medische vragen. *NRC*. https://www.nrc.nl/nieuws/2023/05/02/chatgpt-verslaat-artsen-bij-het-geven-van-antwoord-op-medische-vragen-a4163538
Kreps, S., & Kriner, D. (2023, 21 March). How generative AI impacts democratic engagement. *Brookings*. https://www.brookings.edu/techstream/how-generative-ai-impacts-democratic-engagement/
Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE. Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198
La Moncloa. (2021, 14 July). *The government adopts the Digital Rights Charter to articulate a reference framework to guarantee citizens rights in the new digital age*. La Moncloa. https://www.lamoncloa.gob.es/lang/en/gobierno/news/Paginas/2021/20210713_rights-charter.aspx
Lang, O., Yaya-Stupp, D., Traynis, I., Cole-Lewis, H., Bennett, C. R., Lyles, C., Lau, C., Semturs, C., Webster, D. R., Corrado, G. S., Hassidim, A., Matias, Y., Liu, Y., Hammel, N., & Babenko, B. (2023). *Using generative AI to investigate medical imagery models and datasets* (arXiv:2306.00985). arXiv. https://doi.org/10.48550/arXiv.2306.00985
Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI less "thirsty". Uncovering and addressing the secret water footprint of AI models* (arXiv:2304.03271). arXiv. http://arxiv.org/abs/2304.03271
Lima, C., & Zakrzewski, C. (2023, 30 October). Biden signs AI executive order, the most ambitious U.S. regulation yet. *The Washington Post*. https://www.washingtonpost.com/technology/2023/10/30/biden-artificial-intelligence-executive-order/
Lo, C. K. (2023). What Is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410
Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. *Australasian Journal of Educational Technology*, *39*(1), 1–8. https://doi.org/10.14742/ajet.8695
Loh, E. (2023). ChatGPT and generative AI chatbots. Challenges and opportunities for science, medicine and medical leaders. *BMJ Leader*, leader. https://doi.org/10.1136/leader-2023-000797
Lorach, H., Galvez, A., Spagnolo, V., Martel, F., Karakas, S., Intering, N., Vat, M., Faivre, O., Harte, C., Komi, S., Ravier, J., Collin, T., Coquoz, L., Sakr, I., Baaklini, E., Hernandez-Charpak, S. D., Dumont, G., Buschman, R., Buse, N., … Courtine, G. (2023). Walking naturally after spinal cord injury using a brain–spine interface. *Nature*, *618*(7963), 126–133. https://doi.org/10.1038/s41586-023-06094-5
Lorenz, P., Perset, K., & Berryhill, J. (2023). *Initial policy considerations for generative artificial intelligence*. OECD. https://doi.org/10.1787/fae2d1e6-en
Malinka, K., Perešíni, M., Firc, A., Hujňák, O., & Januš, F. (2023). *On the educational impact of ChatGPT. Is Artificial Intelligence ready to obtain a university degree?* (arXiv:2303.11146). arXiv. https://doi.org/10.48550/arXiv.2303.11146
Marcus, G. (2020). *The next decade in AI. Four steps towards robust Artificial Intelligence* (arXiv:2002.06177). arXiv. https://doi.org/10.48550/arXiv.2002.06177

Marr, B. (2023, 2 March). Revolutionizing healthcare. The top 14 uses of ChatGPT in medicine and wellness. *Forbes*. https://www.forbes.com/sites/bernardmarr/2023/03/02/revolutionizing-healthcare-the-top-14-uses-of-chatgpt-in-medicine-and-wellness/

Maximillian Schrems v Data Protection Commissioner, Case C-362/14 (ECJ October 6, 2015). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62014CJ0362

Meta. (2023, 27 September). *Introducing new AI experiences across our family of apps and devices*. https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2023, 24 January). *Antwoorden op Kamervragen over Rijksbreed cloudbeleid 2022* [Brief]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/brieven/2023/01/24/antwoorden-op-kamervragen-over-rijksbreed-cloudbeleid-2022

Ministerie van Economische Zaken en Klimaat. (2019). *Strategisch Actieplan voor Artificiële Intelligentie*. https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie

Mittelstadt, B., Wachter, S., & Russell, C. (2023). The unfairness of fair machine learning. Levelling down and strict egalitarianism by default. *Michigan Technology Law Review*. https://ora.ox.ac.uk/objects/uuid:09debd0c-7f13-4042-a37e-76381a389362

Mok, A. (2023, 20 April). *ChatGPT could cost over $700,000 per day to operate. Microsoft is reportedly trying to make it cheaper*. Business Insider. https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4

Muench, S., Stoermer, E., Jensen, K., Asikainen, T., Salvi, M., & Scapolo, F. (2022). *Towards a green & digital future*. Publications Office of the European Union. https://doi.org/10.2760/977331

Nabatchi, T. (2018). Public values frames in administration and governance. *Perspectives on Public Management and Governance*, *1*(1), 59–72. https://doi.org/10.1093/ppmgov/gvx009

Nature editorial. (2023). For chemists, the AI revolution has yet to happen. *Nature*, *617*(7961), 438–438. https://doi.org/10.1038/d41586-023-01612-x

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A*, *376*, 1–14. https://doi.org/10.1098/rsta.2018.0089

Nógrádi, B., Polgár, T. F., Meszlényi, V., Kádár, Z., Hertelendy, P., Csáti, A., Szpisjak, L., Halmi, D., Erdélyi-Furka, B., Tóth, M., Molnár, F., Tóth, D., Bősze, Z., Klivényi, P., Siklós, L., & Patai, R. (2023). *ChatGPT M.D.. Is there any room for generative AI in neurology and other medical areas?* (SSRN Scholarly Paper No. 4372965). https://doi.org/10.2139/ssrn.4372965

Norwegian Consumer Council. (2023). *Ghost in the machine. Adressing the consumer harms of generative AI*. https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf

NOS. (2023, 24 September). *Met AI gemaakte naaktfoto's van tientallen meisjes uit Spaans dorp verspreid*. https://nos.nl/artikel/2491701-met-ai-gemaakte-naaktfoto-s-van-tientallen-meisjes-uit-spaans-dorp-verspreid

NOS Nieuws. (2023, 19 April). *Een chatbot die in het echt wil afspreken? Ook Snapchat vindt het wat ver gaan*. https://nos.nl/artikel/2472026-een-chatbot-die-in-het-echt-wil-afspreken-ook-snapchat-vindt-het-wat-ver-gaan

Noy, S., & Zhang, W. (2023). *Experimental evidence on the productivity effects of generative Artificial Intelligence* (SSRN Scholarly Paper No. 4375283). https://doi.org/10.2139/ssrn.4375283

OECD. (2019). *OECD Skills Outlook 2019. Thriving in a Digital World*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/education/oecd-skills-outlook-2019_df80bc12-en

OECD. (2023a). *OECD Employment Outlook 2023. Artificial Intelligence and the Labour Market*. https://www.oecd-ilibrary.org/employment/oecd-employment-outlook-2023_08785bba-en

OECD. (2023b). *AI Language Models. Technological, socio-economic and policy considerations.* (No. 352). OECD Publishing. https://www.oecd-ilibrary.org/docserver/13d38f92-en.pdf?expires=1682422393&id=id&accname=guest&checksum=9528D65C1CA15C4A9A93A1689F2F507E

OECD.AI Policy Observatory. (n.d.). *OECD AI Principles overview*. Retrieved November 15, 2023, from https://oecd.ai/en/ai-principles

Onderwijsraad. (2022). *Inzet van intelligente technologie* [Publicatie]. Ministerie van Onderwijs, Cultuur en Wetenschap. https://www.onderwijsraad.nl/publicaties/adviezen/2022/09/28/inzet-van-intelligente-technologie

OpenAI. (2023, 25 September). *ChatGPT can now see, hear, and speak*. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak?utm_source=substack&utm_medium=email

Pandey, S., & Sharma, S. (2023). A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. *Healthcare Analytics*, *3*, 100198. https://doi.org/10.1016/j.health.2023.100198

Passchier, R. (2021). *Artificiële intelligentie en de rechtsstaat Over verschuivende overheidsmacht, Big Tech en de noodzaak van constitutioneel onderhoud*. Boom.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon emissions and Large Neural Network training* (arXiv:2104.10350). arXiv. https://doi.org/10.48550/arXiv.2104.10350

Perigo, B. (2023, 18 January). The $2 Per hour workers who made ChatGPT safer. *TIME*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Philippidis, A. (2023, 10 February). Insilico gains FDA's first orphan drug designation for AI candidate. *Genetic Engineering and Biotechnology News*. https://www.genengnews.com/news/insilico-gains-fdas-first-orphan-drug-designation-for-ai-candidate/

Qi, X., Zhu, Z., & Wu, B. (2023). The promise and peril of ChatGPT in geriatric nursing education. What we know and do not know. *Journal of Translational Medicine*, *3*(2), 100136. https://doi.org/10.1016/j.ahr.2023.100136

Quekel, S., & Hoijtink, D. (2023, 13 July). 'Briljante' Google-chatbot Bard blundert volop. 'Hugo de Jonge in race voor VVD.' *Algemeen Dagblad*. https://www.ad.nl/binnenland/briljante-google-chatbot-bard-blundert-volop-hugo-de-jonge-in-race-voor-vvd~aebf0eb3/

Raad voor de leefomgeving en infrastructuur. (2021). *Digitaal duurzaam*. https://www.rli.nl/publicaties/2021/advies/digitaal-duurzaam

Raad voor het Openbaar Bestuur. (2021). *Sturen of gestuurd worden? Over de legitimiteit van sturen met data* [Publicatie]. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. https://www.raadopenbaarbestuur.nl/documenten/publicaties/2021/05/25/advies-sturen-of-gestuurd-worden

Rahman, M., & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Applied Sciences*, *13*(9), 5783. https://doi.org/10.3390/app13095783

Raso, F., Hilligloss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial Intelligence & human rights. Opportunities & risks*. Berkman Klein Center for Internet & Society at Harvard University. https://dash.harvard.edu/bitstream/handle/1/38021439/2018-09_AIHumanRights.pdf

Rathenau Instituut. (2015). *Werken aan de robotsamenleving. Visies en inzichten uit de wetenschap over de relatie technologie en werkgelegenheid*. (auteurs: Kool, L. & Van Est, R.). https://www.rathenau.nl/nl/digitalisering/werken-aan-de-robotsamenleving

Rathenau Instituut. (2017). *Opwaarderen. Borgen van publieke waarden in de digitale samenleving*. (auteurs: Kool, L. & Van Est, R.). https://www.rathenau.nl/nl/digitalisering/opwaarderen

Rathenau Instituut. (2019). *Verantwoord virtueel. Bescherm consumenten in virtual reality*. (auteurs: Kool, L. & Van Est, R.). https://www.rathenau.nl/nl/digitalisering/verantwoord-virtueel

Rathenau Instituut. (2020a). *Hoor wie het zegt. Hanvatten voor het verantwoorde gebruik van spraaktechnologie*. (auteurs: Hamer, J. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/hoor-wie-het-zegt

Rathenau Instituut. (2020b). *Nep echt. Verrijk de wereld met augmented reality*. (auteur: Van Est, R.). https://www.rathenau.nl/nl/digitalisering/nep-echt

Rathenau Instituut. (2020c). *Rathenau Manifest. Stel nu 10 ontwerpeisen aan de digitale samenleving van morgen*. https://www.rathenau.nl/nl/digitalisering/rathenau-manifest-stel-nu-10-ontwerpeisen-aan-de-digitale-samenleving-van-morgen

Rathenau Instituut. (2020d). *Werken op waarde geschat. Grenzen aan digitale monitoring op de werkvloer door middel van data, algoritmen en AI*. (auteurs: Das, D. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/werken-op-waarde-geschat

Rathenau Instituut. (2021). *De stand van digitaal Nederland. Naar zeggenschap en vertrouwen in de digitale samenleving*. (auteurs: Hamer, J. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/de-stand-van-digitaal-nederland

Rathenau Instituut. (2022a). *Beter beslissen over datacentra. De noodzaak van een breed publiek perspectief op de digitale infrastructuur*. (auteur: Van Est, R.). https://www.rathenau.nl/nl/digitalisering/beter-beslissen-over-datacentra

Rathenau Instituut. (2022b). *Naar hoogwaardig digitaal onderwijs*. (auteurs: Karstens, B. & Kool, L.). https://www.rathenau.nl/nl/digitalisering/naar-hoogwaardig-digitaal-onderwijs

Rathenau Instituut. (2022c). *Non-discriminatie bij algoritmes*. https://www.rathenau.nl/nl/berichten-aan-het-parlement/non-discriminatie-bij-algoritmes

Rathenau Instituut. (2023a). *Gezondheidstechnologie speciaal voor vrouwen. FemTech en de gezondheidskloof*. (auteurs: Pison, I. & Edelenbosch, R.). https://www.rathenau.nl/nl/gezondheid/gezondheidstechnologie-speciaal-voor-vrouwen

Rathenau Instituut. (2023b). *Immersieve technologieën*. (auteurs: Ex, L., Nieuwenhuizen, W., Hijstek, B., Roolvink, S. & Huijstee, M.). https://www.rathenau.nl/nl/digitalisering/immersieve-technologieen

Ravindran, A. (2023, 8 May). *Brain-computer interfaces and AI language models. A new frontier*. LinkedIn. https://www.linkedin.com/pulse/brain-computer-interfaces-ai-language-models-new-aniruddh-ravindran/

Riemens, R., Nast, C., Pelzer, P., & van den Hurk, M. (2021). An assessment framework for safeguarding public values on mobility platforms. *Urban Transformations*, *3*(1), 7. https://doi.org/10.1186/s42854-021-00023-3

Rijksoverheid. (2022a, 29 August). *Werken 'in de cloud' wordt mogelijk voor Rijksoverheid* [Nieuwsbericht]. https://www.rijksoverheid.nl/actueel/nieuws/2022/08/29/werken-in-de-cloud-wordt-mogelijk-voor-rijksoverheid

Rijksoverheid. (2022b). *Werkagenda Waardengedrageven Digitaliseren*. http://www.digitaleoverheid.nl/werkagenda

Rooij, I. V. (2023, 14 January). Stop feeding the hype and start resisting. *Iris van Rooij*. https://irisvanrooijcogsci.com/2023/01/14/stop-feeding-the-hype-and-start-resisting/

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots. Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.23

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education. Quick start guide*. UNESCO International Institute for Higher Education in Latin America and the Caribbean. https://unesdoc.unesco.org/ark:/48223/pf0000385146

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). *Are emergent abilities of large language models a mirage?* (arXiv:2304.15004). arXiv. http://arxiv.org/abs/2304.15004

Sharon, T. (2016). The Googlization of health research: from disruptive innovation to disruptive ethics. *Personalized Medicine*, *13*(6), 563–574. https://doi.org/10.2217/pme-2016-0057

Sharon, T., & Gellert, R. (2023). Regulating Big Tech expansionism? Sphere transgressions and the limits of Europe's digital regulatory strategy. *Information, Communication & Society*, *0*(0), 1–18. https://doi.org/10.1080/1369118X.2023.2246526

Simon, Julien. (n.d.). *Large Language Models: A New Moore's Law?* Retrieved September 20, 2023, from https://huggingface.co/blog/large-language-models

Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., Atalla, S., & Mansoor, W. (2023). *The future of GPT. A taxonomy of existing ChatGPT research, current challenges, and possible future directions* (SSRN Scholarly Paper No. 4413921). https://doi.org/10.2139/ssrn.4413921

Solaiman, I. (2023, 24 May). Generative AI systems aren't just open or closed source. *Wired*. https://www.wired.com/story/generative-ai-systems-arent-just-open-or-closed-source/

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S., Daumé, I., Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A., Lusoli, A., Mitchell, M., Newman, J., Png, M.-T., Strait, A., & Vassilev, A. (2023a). *Evaluating the Social Impact of Generative AI Systems in Systems and Society*.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Daumé, H., Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A. S., Lusoli, A., Mitchell, M., Newman, J., Png, M.-T., Strait, A., & Vassilev, A. (2023b). *Evaluating the Social Impact of Generative AI Systems in Systems and Society* (arXiv:2306.05949). arXiv. https://doi.org/10.48550/arXiv.2306.05949

Sparnaaij, K., Van Eekeren, P., & Vasseur, J. (2023). *AI Monitor Ziekenhuizen. Editie 2023*. M&I Partners. https://mxi.nl/uploads/files/publication/ai-monitor-ziekenhuizen-2023.pdf

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, *613*(7945), 620–621. https://doi.org/10.1038/d41586-023-00107-z

Sweeney, C. (2021, 20 November). Artificial intelligence. Good or bad for public health? *World Economic Forum*. https://www.weforum.org/agenda/2021/11/artificial-intelligence-technology-public-health/

Takagi, Y., & Nishimoto, S. (2022). *High-resolution image reconstruction with latent diffusion models from human brain activity* (p. 2022.11.18.517004). bioRxiv. https://doi.org/10.1101/2022.11.18.517004

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*(5), 858–866. https://doi.org/10.1038/s41593-023-01304-9

Turkle, S. (2015). *Reclaiming conversation. The power of talk in a digital age.* Penguin Press.

UNESCO. (2022). *Recommendation on the ethics of Artificial Intelligence*. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

UNESCO. (2023). *Guidance for generative AI in education and research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

UNESCO International Bioethics Committee. (2022). *Ethical issues of neurotechnology*. https://unesdoc.unesco.org/ark:/48223/pf0000383559

United Nations. (2023). *Our common agenda policy brief 9. A new agenda for peace.* [Policy brief]. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf

Unlearn.AI. (n.d.). *AI-powered digital twins of individuals patients*. Retrieved August 2, 2023, from https://www.unlearn.ai/technology

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2023). AI in drug discovery. A wake-up call. *Drug Discovery Today*, *28*(1), 103410. https://doi.org/10.1016/j.drudis.2022.103410

U.S. Department of Defense. (2023, 10 August). *DOD announces establishment of generative AI task force*. https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/

Van Bekkum, M., & Zuiderveen Borgesius, F. (2023). De spanning tussen het non-discriminatierecht en het gegevensbeschermingsrecht. Heeft de AVG een nieuwe uitzondering nodig om discriminatie door kunstmatige intelligentie tegen te gaan? *Nederlands Juristenblad*, *27*, 2271–2280.

Van Buchem, M., Boosman, H., Bauer, M., Kant, I., Cammel, S., & Steyerberg, E. (2021). The digital scribe in clinical practice. A scoping review and research agenda. *Npj Digital Medicine*, *4*. https://doi.org/10.1038/s41746-021-00432-5

Van de Ven, L. (2023, 28 July). OM gaat maker deepfakes van Welmoed Sijtsma vervolgen. *NRC*. https://www.nrc.nl/nieuws/2023/07/28/om-gaat-maker-deepfakes-van-welmoed-sijtsma-vervolgen-a4170774

van den Berge, W., & ter Weel, B. (2015). *Baanpolarisatie in Nederland*. Centraal Planbureau. https://www.cpb.nl/publicatie/baanpolarisatie-in-nederland

van Dijck, J., Poell, T., & de Waal, M. (2018). *The Platform Society*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems*. Conference on Neural Information Processing Systems, Long Beach. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Véliz, C. (2023). Chatbots shouldn't use emojis. *Nature*, *615*(7952), 375. https://doi.org/10.1038/d41586-023-00758-y

Villasenor, J., & West, D. M. (2023, 10 April). Will generative AI kill jobs? *Brookings*. https://www.brookings.edu/blog/techtank/2023/04/10/will-generative-ai-kill-jobs/

Vipra, J., & Myers West, S. (2023). *Computational power and AI*. AI Now Institute. https://ainowinstitute.org/publication/policy/compute-and-ai

Visser, D. (2023). Robotkunst en auteursrecht. *Nederlands Juristenblad*, *7*, 504–515. https://www.ipmc.nl/wp-content/uploads/2023/06/Robotkunst-en-auteursrecht-1.pdf

Vogt, M. (2023). Exploring chemical space. Generative models and their evaluation. *Artificial Intelligence in the Life Sciences*, *3*, 100064. https://doi.org/10.1016/j.ailsci.2023.100064

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated. Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567. https://doi.org/10.1016/j.clsr.2021.105567

Wang, J., Mao, J., Wang, M., Le, X., & Wang, Y. (2023). Explore drug-like space with deep generative models. *Methods*, *210*, 52–59. https://doi.org/10.1016/j.ymeth.2023.01.004

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by Language Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Went, R., Kremer, M., & Knottnerus, A. (red.). (2015). *De robot de baas. De toekomst van werk in het tweede machinetijdperk (WRR-Verkenning nr. 31)*. Amsterdam University Press. https://www.wrr.nl/publicaties/verkenningen/2015/12/08/de-robot-de-baas

Wetenschappelijke Raad voor het Regeringsbeleid. (2021). *Opgave AI. De nieuwe systeemtechnologie* (Rapport No. 105). Ministerie van Algemene Zaken. https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

Whang, O. (2023a, 1 May). A.I. is getting better at mind-reading. *The New York Times*. https://www.nytimes.com/2023/05/01/science/ai-speech-language.html

Whang, O. (2023b, 24 May). Brain implants allow paralyzed man to walk using his thoughts. *The New York Times*. https://www.nytimes.com/2023/05/24/science/paralysis-brain-implants-ai.html

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., & Henderson, J. M. (2023). A high-performance speech neuroprosthesis. *Nature*, 1–6. https://doi.org/10.1038/s41586-023-06377-x

Wong, M. (2023, 2 October). Artists are losing the war against AI. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/10/openai-dall-e-3-artists-work/675519/

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for cyber attacks. *Psychology & Marketing*, *57*, 102722. https://doi.org/10.1016/j.jisa.2020.102722

Yasar, A. G., Chong, A., Dong, E., Gilbert, T. K., Hladikova, S., Maio, R., Mougan, C., Shen, X., Singh, S., Stoica, A.-A., Thais, S., & Zilka, M. (2023). *AI and the EU Digital Markets Act. Addressing the risks of bigness in generative AI* (arXiv:2308.02033). arXiv. https://doi.org/10.48550/arXiv.2308.02033

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A survey of Large Language Models* (arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zubascu, F. (n.d.). EU and US hatch transatlantic plan to rein in ChatGPT. *Science|Business*. Retrieved June 19, 2023, from https://sciencebusiness.net/news/AI/eu-and-us-hatch-transatlantic-plan-rein-chatgpt

Auteurs

Jurriën Hamer, Linda Kool, Bo Hijstek, Quirine van Eeden en Djurre Das

Illustraties
Rathenau Instituut

Foto omslag
Stock-Asso / Shutterstock

Bij voorkeur citeren als:
Rathenau Instituut (2023). *Generatieve AI.* Den Haag. Auteurs: Hamer, J., L. Kool, B. Hijstek, Q. van Eeden en D. Das

Contactgegevens
Anna van Saksenlaan 51
Postbus 95366
2509 CJ Den Haag
070-342 15 42
info@rathenau.nl
www.rathenau.nl

Het Rathenau Instituut stimuleert de publieke en politieke meningsvorming over de maatschappelijke aspecten van wetenschap en technologie. We doen onderzoek en organiseren het debat over wetenschap, innovatie en nieuwe technologieën.

Rathenau Instituut