GENERATIEVE AI

Overheidsbrede visie

Generatieve AI

Inhoudsopgave

1 Inleiding

Artificiële intelligentie (AI) heeft als systeemtechnologie grote impact op alle domeinen en sectoren van onze samenleving. Het raakt daarmee alle beleidsterreinen van de overheid.[1] Met de opkomst van *generatieve* AI-toepassingen is het gebruik van AI (nog meer) deel uit gaan maken van het dagelijks leven van veel mensen in Nederland, zowel privé als voor opleiding of werk.

Zo wordt generatieve AI niet alleen gebruikt door professionals zoals data-analisten, reclamemakers en journalisten, maar wordt het ook ingezet om bijvoorbeeld een (Sinterklaas) gedicht te schrijven of een gepersonaliseerd weekmenu samen te stellen. De toepassingsmogelijkheden van generatieve AI zijn daarmee veelzijdig.

Generatieve AI is een vorm van AI waarbij algoritmes worden ingezet om content te genereren.[2] Met een eenvoudige opdracht, een 'prompt', kunnen gebruikers in een handomdraai tekst, beeld, geluid of computercode genereren. Sinds de lancering van ChatGPT eind november 2022, steeg het aantal gebruikers binnen twee maanden naar ruim 100 miljoen wereldwijd. Eind 2023 zijn er circa 180 miljoen gebruikers actief. In Nederland maakten naar verwachting ruim anderhalf miljoen mensen al gebruik van generatieve AI.[3] Naast ChatGPT zijn er nog vele andere generatieve AI-toepassingen (openbaar) beschikbaar, zoals Midjourney, DALL-E en Google Bard. Daarbij wordt gebruikgemaakt van zogenoemde 'large language models' (LLM's).[4]

Generatieve AI is een krachtig verlengstuk van het analytisch en scheppend vermogen van mensen. In combinatie met aanverwante technologieën heeft het uitgebreide mogelijkheden om maatschappelijke en wetenschappelijke vraagstukken aan te pakken. Ook kan het arbeidsproductiviteit verhogen. Tegelijkertijd zijn er ook ingrijpende risico's en staat de verwezenlijking van verschillende publieke waarden en fundamentele rechten onder druk. Generatieve AI kan bijvoorbeeld worden ingezet voor de creatie en verspreiding van desinformatie; het kan discriminatoire dynamieken versterken en het kan sociaaleconomische ongelijkheden versterken.

1. Zie ook het rapport Opgave AI van de Wetenschappelijke Raad voor het Regeringsbeleid (WRR) uit 2021.
2. Een vorm van AI waarbij complexe algoritmes worden ingezet om nieuwe content te genereren zoals tekst, afbeeldingen, computercode of video's. Chatbot ChatGPT vormt hiervan de bekendste exponent.
3. autoriteitpersoonsgegevens.nl/actueel/ap-vraagt-om-opheldering-over-chatgpt
4. Een LLM is een gespecialiseerd type (generatieve) AI dat getraind is op grote hoeveelheden tekst om bestaande content te begrijpen en content te genereren.

De overheid heeft een belangrijke verantwoordelijkheid als het gaat om het in juiste banen leiden van de ontwikkeling, toepassing en inbedding van generatieve AI. De schaal, snelle ontwikkeling en de lage drempel voor gebruik van generatieve AI maken het noodzakelijk een visie op deze technologie te formuleren en om daar concrete acties aan te verbinden.

Het kabinet wil dat generatieve AI in dienst staat van het **vergroten van het menselijk welzijn en autonomie, duurzaamheid, welvaart, rechtvaardigheid en veiligheid**. Door specifiek in te zetten op verantwoorde toepassingen van generatieve AI, grijpen we de kansen die deze technologie biedt. Dit doen we voor alle sectoren. Door de nadruk te leggen op een verantwoorde en open toepassing van generatieve AI, kan de hele samenleving hiervan de vruchten plukken. Daarbij is het de ambitie om een sterk AI-ecosysteem in Nederland en de EU te realiseren, waarin volop geïnnoveerd kan worden met verantwoorde generatieve AI. Het kabinet wil **randvoorwaarden creëren voor de verantwoorde ontwikkeling en het gebruik van generatieve AI**, met behoud van onze digitale open strategische autonomie.

Het kabinet is zich ervan bewust dat de impact van generatieve AI – en AI in het algemeen – afhangt van een samenspel van technologische, economische, institutionele en maatschappelijke factoren. Het kabinet benadrukt dan ook dat het belangrijk is de ontwikkelingen en gevolgen van generatieve AI blijvend te monitoren en analyseren. Om te zorgen voor een goede maatschappelijke inbedding van generatieve AI is het essentieel om vroegtijdig rekening te houden met technologische ontwikkelingen en om daarbij een lerende en evaluerende aanpak te hanteren. Dit doen we nadrukkelijk samen met alle betrokken stakeholders in Nederland, maar zeker ook in een internationale context.

Het kabinet erkent de toegevoegde waarde van generatieve AI, mits het verantwoord wordt ontwikkeld en wordt toegepast. De AI-verordening is daarvoor een belangrijk wettelijk kader. Door verantwoord experimenteren en een lerende aanpak kunnen we in Nederland innovatief gebruik maken van generatieve AI en de mogelijkheden daarvan verkennen. Dit doet de overheid in nauwe samenwerking met bedrijven die over de nodige kennis en competenties beschikken. Samen met het bedrijfsleven wil de overheid actief onderzoeken wat de specifieke meerwaarde van generatieve AI is, bijvoorbeeld voor het bijdragen aan oplossingen voor maatschappelijke uitdagingen, zoals de energietransitie. Het kabinet zal de dialoog hierover met het bedrijfsleven intensiveren.

Met deze overheidsbrede visie benadrukt het kabinet het belang om actie te ondernemen op dit onderwerp, zowel op de korte als op de lange termijn. Daarmee sluit deze visie aan bij de ambities van de Werkagenda Waardengedreven Digitaliseren.[5] Het kabinet richt de aandacht met name op de impact van deze nieuwe digitale technologie op de samenleving. Dit staat in nauwe relatie tot de Strategie Digitale Economie, vooral waar het gaat om het scheppen van de juiste randvoorwaarden voor goed werkende digitale markten en diensten, het stimuleren van digitale innovatie, en het versterken van cybersecurity.[6] Met het presenteren van deze visie wordt ook invulling gegeven aan de motie Dekker-Abdulaziz en Rajkowski, die uw Kamer in april 2023 met een grote meerderheid aannam.[7]

Deze visie is het resultaat van een groot aantal sessies en gesprekken in verschillende domeinen en sectoren, zoals de zorg, de arbeidsmarkt, het onderwijs en het openbaar bestuur. Deze gesprekken zullen ook na de publicatie van deze visie worden voortgezet. Daarbij is actief de samenwerking gezocht met departementen, uitvoeringsorganisaties, medeoverheden, kennis- en hogeronderwijsinstellingen, ontwikkelaars en burgers.[8]

a Afbakening

Deze visie richt zich specifiek op generatieve AI. In tegenstelling tot taakspecifieke AI-systemen, die bijvoorbeeld gebruikt worden voor gezichtsherkenning op een smartphone, naast vele andere toepassingen, is generatieve AI in staat om zelfstandig content te creëren. Sommige generatieve AI-systemen (waaronder systemen met achterliggende grote taalmodellen (LLM's)) kunnen bovendien worden ingezet om een breed scala aan taken uit te voeren. Daarnaast is generatieve AI, via online tools zoals ChatGPT en de integratie in zoekmachines of applicaties als Microsoft Office, beschikbaar voor een breder publiek. De ontwikkeling van generatieve AI staat allerminst stil. Volgende generaties generatieve AI-systemen zullen hoogstwaarschijnlijk meerdere modaliteiten tegelijk kunnen verwerken en veel vaardiger zijn dan de producten die nu op de markt worden gebracht.

b Leeswijzer

Deze visie gaat allereerst in op de vraag om welke technologie het precies gaat en geeft aan wat de verwachte ontwikkelingen op de korte en lange termijn zijn op technologisch vlak. Vervolgens wordt er stilgestaan bij de (maatschappelijke) impact van generatieve AI. Aansluitend wordt het bestaande beleid en regelgeving uiteengezet, als kader waarbinnen de overheidsbrede visie op generatieve AI zal worden gepresenteerd. Hier wordt zowel ingegaan op de nationale, Europese als internationale context. Om ervoor te zorgen dat burgers en bedrijven in Nederland en Europa optimaal kunnen profiteren van deze technologie, en tegelijkertijd beschermd worden tegen de uitwassen, formuleert het kabinet daarbinnen vier uitgangspunten. Aan deze uitgangspunten zijn acties gekoppeld om deze visie in de komende jaren te kunnen realiseren.

5. Zie lijn 2.3 van de Werkagenda 'Anticiperen op nieuwe digitale technologie': rijksoverheid.nl/documenten/rapporten/2022/11/04/bijlage-1-werkagenda-waardengedreven-digitaliseren
6. rijksoverheid.nl/documenten/kamerstukken/2022/11/18/strategie-digitale-economie
7. Gewijzigde motie van de leden Dekker-Abdulaziz en Rajkowski van 4 april 2023 over 'Integrale visie op nieuwe AI-producten' (Kamerstukken 2022/23 26 643, nr. 1003).
8. In bijlage 1 is opgenomen hoe deze open aanpak precies is vormgegeven en welke inzichten dit heeft opgeleverd.

2 Generatieve AI

Dit hoofdstuk gaat nader in op wat er in deze visie vanuit technologisch perspectief wordt verstaan onder generatieve AI en welke ontwikkelingen dit de komende jaren (mogelijk) zal doormaken. Tot slot wordt het speelveld van deze technologie besproken.

a Wat is generatieve AI?

Generatieve AI is een vorm van AI[1] die in staat is om content zoals tekst, audio, afbeeldingen, computercode en video's te genereren. Het onderscheid tussen content gegenereerd door generatieve AI en content gemaakt door mensen is niet altijd direct door de mens te detecteren.

Eén van de meest herkenbare toepassingen van generatieve AI zijn AI-chatbots. Deze digitale assistenten kunnen via tekst communiceren op een manier die sterk lijkt op menselijke interactie. Bekende voorbeelden zijn ChatGPT en Google Bard, beiden AI-chatbots die gebruikmaken van LLM's. De kracht en het groeiend succes van deze modellen ligt in hun veelzijdige toepasbaarheid, variërend van het schrijven van (computer) code tot het spelen van bordspellen.

Andere generatieve AI-systemen kunnen afbeeldingen of audio genereren zoals OpenAI's DALL-E 3 en Google's MusicLM. Door AI-gegenereerde afbeeldingen van personen, geen echte personen dus, komen steeds vaker voor in advertenties en op websites. Door generatieve AI gegenereerde audio is vooral vanaf 2023 in opkomst. Voorbeelden hiervan zijn het genereren van audio op basis van de muziek van bestaande artiesten of toepassing in de gezondheidszorg. Mensen met ALS kunnen op deze manier bijvoorbeeld met hun eigen stem blijven communiceren.

De totstandkoming van generatieve AI-modellen bestaat uit drie fases: pre-training, finetuning en de toepassing. In de pre-trainingsfase wordt het model gevoed met grote hoeveelheden data (zoals tekst, audio of afbeeldingen) van verschillende bronnen. Het model leert tijdens pre-training patronen te herkennen in de data. Dit vereist aanzienlijke computerkracht en wordt uitgevoerd op gespecialiseerde hardware. Na de pre-trainingsfase volgt de finetuningsfase. Hierin wordt het model getraind om instructies van gebruikers op te volgen, wordt eventueel specialistische kennis toegevoegd, en wordt het model mogelijk geoefend om sociaal aanvaardbare antwoorden te geven. Hierbij worden speciale technieken toegepast, zoals Reinforcement Learning from Human Feedback

1. We volgen de recent herziene definitie van 'AI-systeem' van de OESO (2023): een op machines gebaseerd systeem dat, voor expliciete of impliciete doelstellingen, afleidt, uit de input die het ontvangt, hoe de output zoals voorspellingen, inhoud, aanbevelingen of beslissingen moet genereren die fysieke of virtuele omgevingen kunnen beïnvloeden. Verschillende AI-systemen variëren in hun mate van autonomie en aanpassingsvermogen na de implementatie/toepassing (deployment) ervan.

(RLHF).[2] Na de finetuningsfase volgt de toepassingsfase, waarin het model beschikbaar wordt gesteld aan gebruikers.[3] Het model kan worden gedupliceerd om vervolgens via een consumenteninterface te kunnen worden ingezet door tienduizenden gebruikers tegelijk. Meer informatie over het technische ontwikkelproces van generatieve AI is te vinden in bijlage 2.

b Trends

In de ontwikkelingen rondom generatieve AI zijn vijf trends te onderscheiden:

1. *Modellen worden steeds vaardiger en toepasbaarder.* Dit omvat zowel de aanscherping van bestaande vaardigheden als de ontwikkeling van nieuwe vaardigheden. Huidige modellen kunnen bijvoorbeeld een gebruiker assisteren bij programmeertaken, terwijl de vorige generatie modellen daartoe nog amper in staat was.
2. *Modellen worden steeds vaker voorzien van 'vangrails'.* Vangrails (*guardrails*) zijn veiligheidsmaatregelen die de interactie tussen een AI-model en gebruiker voorschrijven en op basis waarvan gemonitord kan worden. Er is op dit vlak echter nog een lange weg te gaan. Huidige modellen genereren bijvoorbeeld regelmatig onjuiste uitkomsten ('hallucineren'). En in sommige gevallen kunnen ook beveiligingsmaatregelen en ethische kaders relatief gemakkelijk worden omzeild.
3. *Modellen worden multimodaal.* Waar AI-modellen eerst enkel tekst of audio of beeld konden verwerken, zijn er het afgelopen jaar nieuwe AI-systemen ontwikkeld die tegelijkertijd met deze contentvormen om kunnen gaan.
4. *Modellen worden zelfstandig(er).* Nieuwe AI-systemen zijn in staat zelfstandig digitale tools met elkaar in verbinding te brengen en deze vervolgens zelfstandig in een reeks te gebruiken. Ook het verzamelen van gegevens en het plannen en uitvoeren van taken kan zelfstandig worden uitgevoerd.
5. *Modellen worden kosteneffectiever.* Hoewel grotere AI-modellen een stuk vaardiger zijn geworden, vereisen ze ook aanzienlijk meer rekenkracht. Er wordt daarom actief gewerkt aan het creëren van compactere, betaalbaardere, en snellere modellen zonder significant prestatieverlies.

Monitoren van ontwikkelingen

We brengen jaarlijks een monitor generatieve AI uit om de ontwikkeling en het gebruik van generatieve AI voor en door overheden te volgen

c Ontwikkelsnelheid

De ontwikkelingen in generatieve AI gaan, ook na het beschikbaar worden voor het grote publiek eind 2022, onverminderd hard verder. De rekenkracht waarmee generatieve AI-modellen getraind worden neemt met een factor vier toe per jaar. Daarnaast neemt de algoritmische efficiëntie van AI-modellen toe met een factor 2,5 per jaar.[4] Deze gestapelde exponentiële groei heeft de afgelopen jaren geleid tot veel vaardigere generatieve AI-systemen. Zo zijn deze inmiddels in staat om, met een vergaand niveau van zelfstandigheid, complexe processen te automatiseren, ingewikkelde data-analyses uit te voeren, of een gebruiker op basis van een foto te laten zien hoe ze hun fiets moeten repareren. Generatieve AI-systemen kunnen bovendien worden gekoppeld aan het internet en instructies krijgen om allerlei handelingen uit te voeren, zoals het boeken van een vliegticket. Verbetering van AI-vaardigheden berust voor een aanzienlijk deel op schaalvergroting Ontwikkelaars kunnen een beter model trainen door simpelweg meer AI-chips en meer data in te zetten. Deze schaalvergroting zal zich de komende jaren naar verwachting voortzetten. Generatieve AI-modellen zullen dus nog veel vaardiger worden.

De opkomst van steeds vaardiger generatieve AI-systemen heeft de vraag doen rijzen of we hiermee op weg zijn naar *artificial general intelligence* (AGI). AGI refereert aan technologie die intelligentie vertoont over een breed scala aan domeinen, en die met deze vaardigheden op of boven het menselijke niveau presteert.[5] Tot op heden is er geen wetenschappelijke consensus dat we zouden kunnen spreken van AGI.[6] Wel staat vast dat grote geopolitieke mogendheden en techbedrijven aanzienlijke bedragen investeren in de ontwikkeling van geavanceerde AI-systemen. Dit heeft geleid tot een concurrentiestrijd tussen landen en bedrijven die zeggen bezig te zijn met de ontwikkeling van AGI. Omdat bedrijven in de EU

2. In het geval van RLHF wordt menselijke feedback opgenomen in het trainingsproces van AI-algoritmes om het leren van het AI-algoritme te sturen of te verbeteren. Deze feedback van mensen kan als effect hebben dat het algoritme sneller en effectiever kan leren. Het doel is vaak om menselijke expertise te benutten om AI-algoritmes een bepaalde gewenste richting op te sturen.
3. Generatieve AI trekt een divers scala aan gebruikers aan, variërend in expertise en doelstellingen.
4. Zie ook: AI Trends – Epoch (epochai.org)
5. Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data and Cognitive Computing, 7(1), 35.
6. Tredinnick, L., & Laybats, C. (2023). The dangers of generative artificial intelligence. Business Information Review.

en Nederland niet de middelen hebben om de concurrentie aan te gaan met deze spelers doen ze hier momenteel niet actief aan mee.

Verkennen van een AI-adviesorgaan op het hoogste niveau

We verkennen het inrichten van een AI adviesraad (of Rapid Response Team AI) op het hoogste niveau die het kabinet kan voorzien van (korte- en langetermijn) advies.

d Speelveld

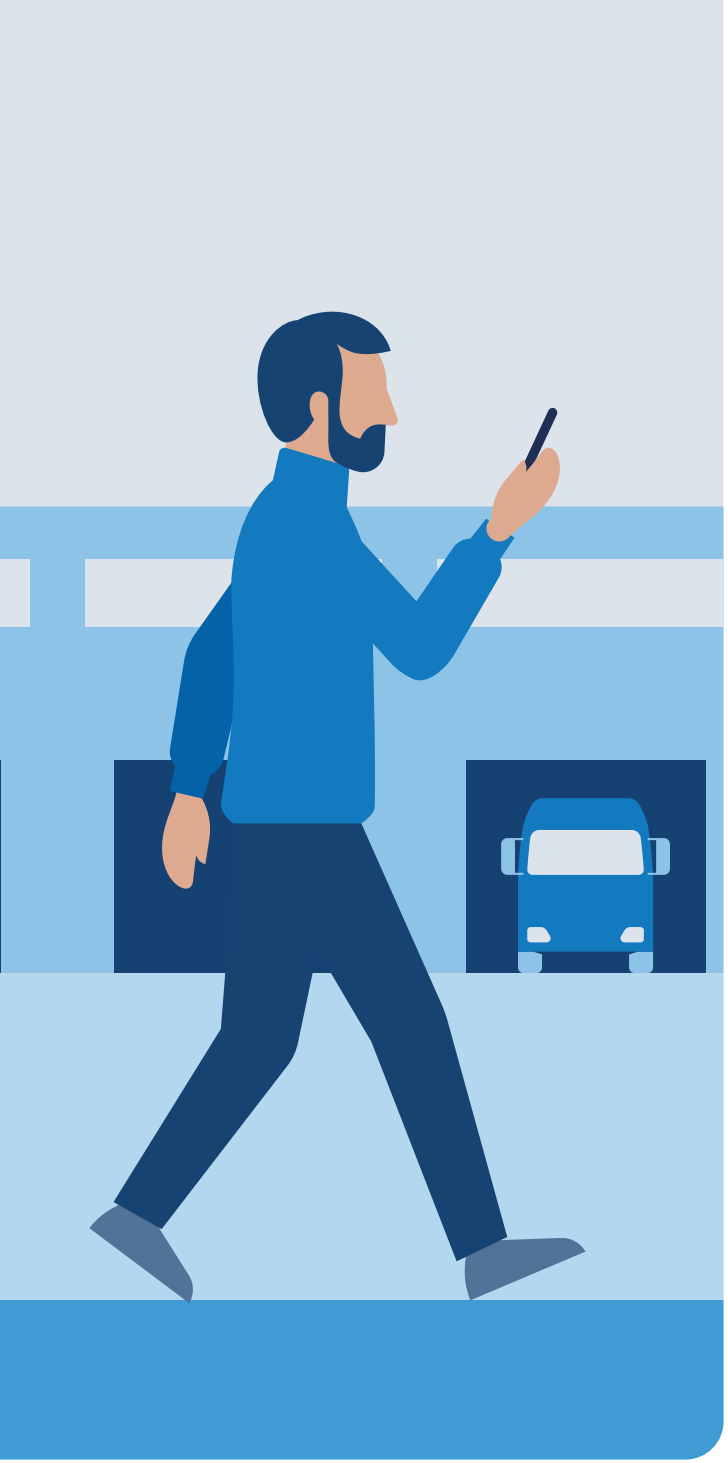
Om generatieve AI-modellen te ontwikkelen zijn grootschalige investeringen in computerinfrastructuur nodig. Dit heeft de ontwikkeling van generatieve AI richting commerciële partijen geduwd. Deze Amerikaanse AI-labs (gesteund door hun cloud-providers) domineren de ontwikkeling van state-of-the-art generatieve AI-modellen. Dit komt met name doordat zij over de rekenkracht, talent en data beschikken die nodig is voor het trainen en doorontwikkelen van generatieve AI. Deze winner-takes-all dynamiek draagt bij aan het verstevigen van de voorsprong van deze bedrijven. Dit heeft tot gevolg dat Europese organisaties en burgers in toenemende mate afhankelijk worden van een kleine groep ontwikkelaars van generatieve AI.[7] De grote investeringen die nu nodig zijn om voldoende computerkracht te kopen, en gebrek aan aantrekkelijke businessmodellen maken het voor Nederlandse organisaties vrijwel onmogelijk in deze dynamiek een plaats te veroveren. In andere lidstaten zijn er wel al enkele bedrijven die nieuwe generatieve AI modellen trainen, maar ook deze spelers lopen achter op grote Amerikaanse en Chinese concurrenten. Om een positie te verwerven in deze markt, is Europese samenwerking noodzakelijk.

Niet alle AI-labs vermarkten hun modellen op dezelfde manier. De meeste bedrijven verspreiden hun modellen alleen via een API of via een consumentenproduct. Andere bedrijven, zoals Meta, kiezen er bewust voor om de modelparameters van het AI-model openbaar te maken. Dit stelt gebruikers in staat het model verder te finetunen, waardoor het model flexibeler inzetbaar is. Een nadeel van deze aanpak is dat ook eventuele veiligheidsmaatregelen – bijvoorbeeld tegen racisme of illegaal gebruik – gemakkelijk uit het model gehaald kunnen worden. De ontwikkeling van generatieve AI-modellen is sterk afhankelijk van een geconcentreerde hardware-keten. Meer dan 75% van alle state-of-the-art AI-chips wordt ontworpen door het Amerikaanse NVIDIA en geproduceerd bij TSMC in Taiwan. Hierbij worden lithografiemachines gebruikt die vrijwel allemaal in Nederland gemaakt worden. Als thuisland van een sterk halfgeleiderecosysteem heeft Nederland een unieke positie in de AI-ontwikkelketen.

High Performance Computing

Nederland neemt deel aan het partnerschap EuroHPC. Nederlandse bedrijven en kennisinstellingen kunnen zo deelnemen aan Europese projecten op gebied van HPC en quantumcomputing. Met HPC kunnen complexe berekeningen op hoge snelheid worden uitgevoerd, waarmee een significante bijdrage kan worden geleverd aan complexe vraagstukken.

7. Zie ook Agenda DOSA: Agenda Digitale Open Strategische Autonomie | Rapport | Rijksoverheid.nl



3 De impact van generatieve AI

De ontwikkeling van generatieve AI, zoals geschetst in hoofdstuk 2, zal zich in de komende jaren intensiveren en waarschijnlijk een grote weerslag hebben op mens, maatschappij, werk en economie. Hieronder zal de verwachte impact op de (Nederlandse) samenleving worden opgedeeld in kansen en mogelijkheden enerzijds en risico's en uitdagingen anderzijds.

Met deze uiteenzetting wordt ook voor een gedeelte gehoor gegeven aan de motie Dassen van september 2023[1] en de motie Dekker-Abdulaziz en Rajkowski van april 2023.[2] De motie Dassen verzoekt het kabinet onder meer om per ministerie de impact van de toepassing van AI door de overheid in kaart te brengen. De motie Dekker-Abdulaziz en Rajkowski van maart 2023 verzoekt het kabinet om onder andere te inventariseren welke negatieve gevolgen en uitwassen het gebruik van AI met zich mee kan brengen voor de Nederlandse samenleving.

De gevolgen van generatieve AI zullen zich nog verder moeten manifesteren. Dit kan jaren duren. De uiteindelijke impact van generatieve AI hangt af van een samenspel van technologische, economische, institutionele en maatschappelijke factoren. Het kabinet benadrukt daarom het belang van het monitoren en analyseren van de ontwikkelingen en gevolgen van generatieve AI (zie ook onder 'acties' in hoofdstuk 5). In dit hoofdstuk wordt ingegaan op de voorziene kansen en risico's, geïdentificeerd op basis van de eerste wetenschappelijke bevindingen en voorspellingen van experts.

De impact van generatieve AI kan zowel positieve als negatieve kanten hebben. Zo biedt generatieve AI bijvoorbeeld kansen als het gaat om het genereren van informatie, maar kan het ook leiden tot desinformatie en onnavolgbaarheid. De risico's die verderop worden genoemd zijn dus veelal 'de andere kant van de medaille'. Of een bepaalde capaciteit van generatieve AI zich uiteindelijk als kans of als risico manifesteert, hangt af van de specifieke ontwikkeling, de toepassing van de technologie en de intenties dan wel expertise van de gebruiker. Het is daarom belangrijk om adequaat regie te voeren en toezicht te houden op generatieve AI. In hoofdstuk 4 en 5 wordt de invulling hiervan besproken. Dit hoofdstuk heeft als doel een overzicht te bieden van de kansen en risico's die generatieve

1. Kamerstuk 2023-2024, 36 410, nr. 53.
2. Kamerstuk 2022-2023, 26 643, nr. 1002.

AI opwerpt in verschillende sociale sferen en sectoren. Een deel van de genoemde gevolgen komt niet zozeer voort uit generatieve AI-modellen *an sich*, maar uit het gebruik en de maatschappelijke omarming (of juist niet omarming) van de technologie.

a Kansen en mogelijkheden

Generatieve AI opent een breed scala aan mogelijkheden, zoals het maken van gespreksverslagen, muziekcompositie, beeldsynthese en het ontdekken en ontwerpen van nieuwe moleculen en materialen. Daarnaast wordt er nog volop onderzoek gedaan naar de precieze kansen die generatieve AI biedt en hoe deze optimaal benut kunnen worden. Wat duidelijk is, is dat in vergelijking met kleinere en meer gespecialiseerde AI-modellen, een nieuwe generatie generatieve AI-modellen als 'basismodel' in verschillende domeinen kan worden ingezet voor uiteenlopende algemene doeleinden. Generatieve AI-modellen kunnen daardoor in tal van sectoren en domeinen ingezet worden om **processen te optimaliseren**, te automatiseren en te assisteren bij taken zoals het verzamelen, samenvatten en uitwerken van (grote hoeveelheden) informatie of het schrijven van computercode. Daarmee kan generatieve AI onder andere zorgen voor verbeterde efficiëntie, kostenbesparing, betere besluitvorming, betere dienstverlening en tal van innovatieve oplossingen.

Generatieve AI kan taken vanuit verschillende rollen vervullen: als **productietool, leerinstrument,** en als **probleemoplosser** (of een combinatie hiervan).[3] Dit biedt kansen voor individuen, bedrijven, de overheid en de samenleving als geheel. Gezien de algehele potentie van de technologie voor de samenleving en de economie, zet het kabinet in op het stimuleren van verantwoord experimenteren en verantwoord gebruik van generatieve AI-modellen en systemen in verschillende sectoren en domeinen.

Generatieve AI als productietool

Generatieve AI schept mogelijkheden voor de productie van allerlei soorten digitale content, zoals de beantwoording van vragen, het samenvatten van teksten, het maken van video's en het schrijven van teksten. Dit heeft nu al effect op het dagelijks leven van individuele burgers. Velen gebruiken AI-chatbots als ChatGPT voor allerlei taken in hun (persoonlijke) leven, zoals voor het bedenken van een recept, persoonlijke sporttrainingsschema's of voor het genereren van gedichten en sollicitatiebrieven.

Voor bedrijven en organisaties biedt generatieve AI als productietool ook **veelbelovende** mogelijkheden voor de efficiëntie en kwaliteit van allerlei **bedrijfsprocessen**. Zo wordt generatieve AI al ingezet om administratieve processen te ondersteunen en te versnellen, klantenservices te ondersteunen, computercode te schrijven en industriële processen te automatiseren.

In verschillende industriële processen kan generatieve AI snel een groot aantal ontwerpalternatieven produceren. Hierdoor kan onder andere in de **maakindustrie** het ontwerpproces van bijvoorbeeld machinebouw potentieel aanzienlijk versneld worden en kan onderhoud aan complexe machines makkelijker en goedkoper worden uitgevoerd. Daarnaast kunnen generatieve AI-modellen helpen bij proactieve besluitvorming en bij het verlagen van kosten die verband houden met overproductie of voorraadtekorten, door het simuleren van verschillende productiescenario's op basis van de voorspelde vraag van klanten.

In de **culturele sector** wordt generatieve AI ingezet ter ondersteuning van het creatieve proces, onder andere bij het schrijven van scripts door filmscenaristen of bij het genereren van beschrijvingen van kunstwerken. Ook wordt generatieve AI ingezet bij het creëren van simulaties voor trainingsdoeleinden, voor het uitwerken van toekomstscenario's of voor een virtuele representatie van een product of proces *(digital twins)*.

Innovatielabs met mkb

In 2024 zullen vanuit AiNed InnovatieLabs gestart worden. InnovatieLabs zijn publiek-private samenwerkingen gericht op de ontwikkeling van AI-innovaties, met een focus op mkb en start- en scale-ups.

Voor de **muziekindustrie** kan generatieve AI als een katalysator fungeren voor creativiteit en innovatie. Op dit gebied is het in staat om nieuwe composities te genereren, waardoor nieuwe melodieën, harmonieën en ritmes ontstaan die muzikanten kunnen inspireren of ondersteunen in hun werk.

Vanwege het vermogen van generatieve AI om computercode te schrijven op basis van 'prompts', biedt de technologie ook veel mogelijkheden in de **ICT-sector**. Generatieve AI heeft groot potentieel om het programmeren van software en applicaties te verbeteren en om IT'ers te ondersteunen in hun werk. Hierdoor kan het ontwikkelingsproces aanzienlijk worden versneld en houden IT'ers meer tijd over voor andere taken

Generatieve AI kan leiden tot algehele **productiviteitsgroei**.[4] In het verleden heeft productiviteitsgroei geleid tot een toename in de **materiële welvaart**, betere gezondheid en meer vrije tijd. Economen voorspellen een dergelijk productiviteitseffect,[5] niet alleen voor grote bedrijven, maar ook voor het midden- en

3. De indeling van generatieve-AI-capaciteiten in deze rollen is ontleend aan het rapport van het Rathenau Instituut over generatieve AI (2023): rathenau.nl/sites/default/files/2023-12/Scan_Generatieve_AI_Rathenau_Instituut.pdf
4. technologyreview.com/2021/06/10/1026008/the-coming-productivity-boom/
5. kentclarkcenter.org/surveys/ai-and-productivity-growth/

kleinbedrijf (mkb), omdat toepassingen van generatieve AI laagdrempelig zijn.[6] Generatieve AI kan bepaalde taken (zoals het uitvoeren van financiële analyses en juridische processen) namelijk ook op kostenefficiënte wijze, *inhouse*[7] beschikbaar maken voor het mkb.[8] Productiviteits- en welvaartgroei en de creatie van nieuwe taken als gevolg van de inzet van generatieve AI, kunnen leiden tot **nieuwe werkgelegenheid**.[9] Een direct effect is het ontstaan van nieuwe beroepsgroepen (zoals informatie-specialisten en ICT'ers) die over de vaardigheden beschikken om generatieve AI-toepassingen te implementeren.[10] Ook zal er vraag zijn naar werkenden die beschikken over de nodige competenties om de inzet van AI ter ondersteuning van hun werk verantwoord te laten plaatsvinden. Een indirect effect op werkgelegenheid loopt via de mogelijke groei van het besteedbare inkomen in de economie. Inkomensgroei (als gevolg van AI-geïnduceerde productiviteitsgroei) kan de vraag naar goederen en diensten doen toenemen, met de toename van werkgelegenheid als gevolg. De meeste economen verwachten dat het totaal aantal banen in de economie op de langere termijn niet zal afnemen als gevolg van AI-gestuurde automatisering.[11] Toch kunnen er verdelingsvraagstukken ontstaan, die in paragraaf 3b als mogelijke 'uitdaging of risico' worden besproken. Hierbij is het belangrijk om op te merken dat voor productiviteitsgroei moet worden voldaan aan specifieke randvoorwaarden, onder andere kennisopbouw bij organisaties.

Als productietool kan generatieve AI een positief effect hebben op de aard van het werk dat door mensen wordt uitgevoerd. Zo kan de inzet van AI routinematige taken (zoals het maken van notulen, het uitluisteren van audio of het beantwoorden van standaardvragen) van werknemers overnemen. Ook zijn er tekenen dat AI-technologie juist werknemers met minder kennis en ervaring helpt om bij te blijven bij werknemers met meer kennis en ervaring.[12] Dit vergroot het gevoel van professionele autonomie en competentie. Door de genoemde factoren kan de (gepercipieerde) **kwaliteit van werk** toenemen. Het omgekeerde kan ook gebeuren, zoals verderop besproken onder de noemer 'uitdagingen'.

Ook de overheid kan profiteren van generatieve AI als productietool. Het biedt overheden kansen om processen te verbeteren, het algemeen functioneren van de overheid te verbeteren en de **dienstverlening aan burgers te optimaliseren**. Bijvoorbeeld door bij te dragen aan het beter bereiken van de inwoner. Een andere mogelijkheid waar generatieve AI zich voor leent is het toegankelijker maken van overheidsinformatie voor iedereen, door in aanpassingen van het taalniveau te voorzien. Op deze manier kan de technologie een bijdrage leveren aan heldere en inclusieve communicatie met burgers. Tevens zou generatieve AI de efficiëntie van juridische en administratieve processen kunnen verhogen door het (deels) automatiseren van formulieren, zoals bij 'Legal Tech', waardoor er ruimte overblijft voor maatwerk. Voorwaarde hiervoor is wel dat de technologie ethisch verantwoord en goed gereguleerd wordt ingezet.

Door het snel analyseren van grote hoeveelheden data, het genereren van trainingsmateriaal en het simuleren van beleidsscenario's, kan generatieve ook een rol spelen in datagedreven beleidsvorming en -evaluatie, en biedt het kansen voor interne kennisontwikkeling.

Generatieve AI als leerinstrument

Generatieve AI-modellen kunnen gebruikt worden om enorme hoeveelheden data snel te analyseren. Generatieve AI kan daardoor bijvoorbeeld worden ingezet om complexe teksten uit te leggen, de belangrijkste onderwerpen te duiden en conclusies te verbinden. Gebruikers kunnen generatieve AI daarom ook benutten als leerinstrument.

Deze rol zien we bijvoorbeeld terug op het gebied van **taal en vertaling**, Generatieve AI-modellen zijn in staat met hoge accuraatheid grote hoeveelheden tekst te vertalen.[13] De technologie kan daardoor worden ingezet voor het vertalen van content zodat de betreffende informatie voor een breder publiek beschikbaar wordt. Dit kan bruikbaar zijn voor het leren van een nieuwe taal, maar ook bijvoorbeeld voor het vertalen van (overheids)websites of onderwijsmaterialen.

Generatieve AI kan dan ook een belangrijke rol spelen als **leerinstrument** in het **onderwijs**. Zo kan deze technologie

Generatieve AI pilots bij en met overheden

Via pilots bij de (rijks)overheid wordt beproefd hoe generatieve AI op een verantwoorde en veilige manier kan worden ingezet, bijvoorbeeld op het gebied van proactieve dienstverlening.

6. Mills, K. (2019). How AI could help small business. Harvard Business Review.
7. Bijvoorbeeld door middel van Artificial Intelligence as a Service (AIaaS).
8. OECD (3 februari 2021) the Digital Transformation of SMEs. Hoofdstuk 5. Artificial Intelligence, changing landscape for SMEs.
9. wsj.com/tech/ai/the-new-jobs-for-humans-in-the-ai-era-db7d8acd
10. OESO (2023). OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market.
11. Op de lange termijn houden de verdringing van banen en creatie van nieuwe taken elkaar ongeveer in balans. doordat vraag en aanbod via de prijzen naar een evenwicht bewegen. Zie: David H. Author (2015). 'Why Are There Still So Many Jobs? The History and Future of Workplace Automation.' Journal of Economic Perspectives 29(3): pp. 3-30.
12. Brynjolfsson, Li & Raymond (2023). 'Generative AI at Work.' National Bureau of Economic Research. Working paper no. 31161.
13. Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52-62.

leerlingen en studenten ondersteunen bij het maken van samenvattingen, het uitleggen van leerstof en het creëren van oefenvragen. Door het analyseren van de leerpatronen van de gebruiker, zijn generatieve AI-modellen in staat gepersonaliseerde feedback, aanbevelingen en interventies op te stellen, om zo onderwijs toe te spitsen op de persoonlijke leerbehoeften.[14] Ook docenten kunnen generatieve AI inzetten voor bijvoorbeeld het ontwerpen van lesvormen of het verbeteren van lesmateriaal. Daarnaast stelt generatieve AI hen in staat, op basis van studentdata uit het verleden, een voorspelling te maken over toekomstig presteren van studenten en daarmee studenten te identificeren die mogelijk meer ondersteuning nodig hebben.[15]

De rol van leerinstrument is bijvoorbeeld ook zichtbaar bij het gebruik van digitale zoekmachines, die in het dagelijks leven door vrijwel iedereen worden gebruikt. Generatieve AI kan de functie van zoekmachines aanzienlijk verbeteren en is reeds door bedrijven zoals Google en Microsoft daarin geïntegreerd.

Tot slot zijn generatieve AI-toepassingen zoals ChatGPT in staat concepten op allerlei manieren uit te leggen. Gebruikers kunnen dit inzetten voor studie of werk, maar ook voor alledaagse onderwerpen zoals uitleg over de werking van een stoppenkast of energiebesparing. Door het interactieve karakter van toepassingen zoals ChatGPT kan ook om een aangepaste of meer gedetailleerde uitleg worden gevraagd.

Generatieve AI als probleemoplosser

Generatieve AI kan een belangrijke rol spelen als **probleemoplosser**. Dit is onder andere terug te zien in het **wetenschapsdomein**, zoals bij de ontwikkeling van nieuwe medicijnen. Het proces van medicijnontwikkeling kent vaak een lang, complex en kostbaar verloop. Generatieve AI heeft al veelbelovende resultaten laten zien in het versnellen en het verbeteren van het proces van medicijnontwikkeling,[16] waardoor tijd kan worden gewonnen en mogelijk kosten verlaagd. Deze voordelen beperken zich niet tot medicijnontwikkeling en verspreiden zich verder over het wetenschapsdomein.

Ook voor het ontwikkelen en verbeteren van **materialen** biedt generatieve AI mogelijkheden. Onderzoek naar nieuwe materialen voor bijvoorbeeld batterijen of microchips kan maanden, al dan niet jaren in beslag nemen. Generatieve AI-modellen zijn in staat, veel sneller dan mensen, nieuwe chemicaliën, moleculen en materialen te genereren en dragen bij aan een sneller en efficiënter proces.[17]

De rol van probleemoplosser wordt al ingezet binnen de **gezondheidszorg**, waar bijvoorbeeld geëxperimenteerd wordt met generatieve AI als adviseur bij kankerbehandeling.[18] Er zijn daarbij kansen bij het analyseren van data binnen klinisch onderzoek, om zo beter te voorspellen welke patiënten baat hebben bij een nieuwe behandeling. Generatieve AI kan op deze manier op termijn een rol spelen bij het verminderen van late of incorrecte diagnoses. Daarnaast kan generatieve AI in de gezondheidszorg onder andere worden ingezet voor de uitvoering van repetitieve en administratieve taken, zoals het samenvatten van patiëntgesprekken en het aanvullen van patiëntdossiers. Medische professionals hebben zo meer tijd voor inhoudelijke werkzaamheden. Dit kan de druk op de gezondheidszorg verminderen en de kwaliteit van de gezondheidszorg verbeteren.

Generatieve AI kan daarmee als probleemoplosser bijdragen aan het **oplossen van grote maatschappelijke problemen**. Ondanks zorgen over het energiegebruik van (generatieve) AI-technologie, staat daartegenover potentiële bijdrage die (generatieve) AI kan leveren aan de **duurzaamheidstransitie**. Generatieve AI kan bijvoorbeeld worden ingezet voor de analyse van natuurlijke ecosystemen of het voorspellen van klimaattrends[19]. Er bestaan daarnaast al generatieve AI-toepassingen die scheepvaartbedrijven hun emissies in de gaten laten houden of die voor industrieën operationele strategieën genereren voor verduurzaming.[20] De rol van (generatieve) AI als probleemoplosser voor maatschappelijke uitdagingen valt ook terug te zien in het **militaire domein**. Voorbeelden hiervan zijn modelling en simulatie (*wargaming*)[21] en de inzet bij operationeel-tactische planning door middel van toegankelijke big data analytics.[22] Ook in het **cybersecuritydomein** liggen er kansen. AI-toepassingen maken het bijvoorbeeld mogelijk voor organisaties om automatisch aanvallen te detecteren via geconstateerde anomalieën in hun netwerk. Met generatieve AI kunnen daarna automatisch analyses gegenereerd worden, zodat op basis daarvan actie kan worden genomen, merkt de Cyber Security Raad (CSR) in het najaar van 2023 op.[23]

14. Abunaseer, H. The Use of Generative AI in Education: Applications, and Impact. Technology and the Curriculum: Summer 2023.
15. Wang, T., Lund, B. D., Marengo, A., Pagano, A., Mannuru, N. R., Teel, Z. A., & Pange, J. (2023). Exploring the Potential Impact of Artificial Intelligence (AI) on International Students in Higher Education: Generative AI, Chatbots, Analytics, and International Student Success. Applied Sciences, 13(11), 6716.
16. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science, 12(5), e1608.
17. Liu, Y., Yang, Z., Yu, Z., Liu, Z., Liu, D., Lin, H., ... & Shi, S. (2023). Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. Journal of Materiomics.
18. Sorin, V., Klang, E., Sklair-Levy, M., Cohen, I., Zippel, D. B., Balint Lahat, N., ... & Barash, Y. (2023). Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer, 9(1), 44.
19. abnamro.nl/nl/media/rapport-generatieve-ai-pakt-rol-in-de-duurzaamheidstransitie-december-2023_tcm16-216530.pdf
20. bearing.ai/
21. magazines.defensie.nl/defensiekrant/2019/23/06_wargaming_23
22. open.overheid.nl/documenten/d49f42ca-181b-4e2f-9986-b412de40f2f5/file
23. Zie ook: cybersecurityraad.nl/actueel/nieuws/2023/12/22/csr-brief-over-ai-en-cybersecurity

Kansen en mogelijkheden

Generatieve AI als
Productietool

Efficiëntie en kwaliteit bedrijfsprocessen

Creatieve proces

ICT Sector

Materiële welvaart

Nieuwe werkgelegenheid

Kwaliteit van werk

Overheid

Juridisch

Generatieve AI als
Leerinstrument

Taal en vertaling

Onderwijs

Zoekmachines

Interactief hulpmiddel

Generatieve AI als
Probleemoplosser

Wetenschap, medicijnen

Materialen o.a. batterijen

Gezondheidszorg

Maatschappelijke problemen o.a. duurzaamheid

Militaire domein

Cybersecurity

Uitdagingen en risico's

Invloed op individuele burgers

Bias/discriminatie

Privacy, gegevensbescherming gebruikersautonomie

Cognitieve ontwikkeling
Sociale ontwikkeling

Auteurs-, nabuur- en databankrecht portretrecht

Afhankelijkheid en marktmacht

Groeiende afhankelijkheid Amerikaanse techbedrijven
Strategische afhankelijkheden

Machtsconcentratie
Toetredingsdrempels

Arbeid en arbeidsmarkt

Werkgelegenheid
Inkomensverdeling.
Werkloosheid/loondaling

Kwaliteit van werk

Verdeling inkomen/ werkzekerheid
Polarisatie arbeidsmarkt

Invloed op de maatschappij

Superstar firms
Toenemende sociale en economische ongelijkheid

Groot energieverbruik
Klimaatimpact

Aantasting informatie-ecosysteem
Mis- en desinformatie

Onzekere betrouwbaarheid/ automation bias

Militaire veiligheid
Systemische veiligheidsrisico's

Opzettelijk misbruik generatieve AI-modellen
Hate speech

b Uitdagingen en risico's

Generatieve AI biedt niet alleen kansen, maar gaat ook gepaard met risico's en uitdagingen, uitdagingen die soms voortvloeien uit de kansen voor de toepassingen van de technologie. Hieronder maken we onderscheid tussen impact op individuele burgers, marktinrichting, arbeid en inkomen en de maatschappij als geheel.

Stimuleren van (taal) modellen voor talen als Fries en Papiamento

We stimuleren actief het ontwikkelen en verbeteren van (open en publieke) taalmodellen die getraind zijn op talen zoals bijvoorbeeld Fries, Papiaments of gebarentaal.

Invloed op individuele burgers

Er zijn risico's verbonden aan het gebruik van generatieve AI. De eerste uitdaging is dat door bestaande **bias** (vooringenomenheid of selectiviteit verankerd in trainingsdata en modelparameters[24]) **discriminatoire dynamieken** versterkt kunnen worden.[25] Deze bias wordt mogelijk versterkt door het feit dat veelgebruikte AI-modellen van grote ontwikkelaars door een selecte groep mensen wordt gemaakt met vaak een eenzijdig perspectief. Bias heeft negatieve gevolgen voor de maatschappelijke erkenning en representatie van personen die generatieve AI- gebruiken of erdoor worden beïnvloed. **Gelijke behandeling en non-discriminatie** staan daarmee onder druk. Door gebrekkige transparantie, uitlegbaarheid en complexiteit van AI-modellen kunnen bias en discriminatoire effecten lang verborgen blijven.

Een tweede uitdaging betreft de mogelijke schendingen van **rechten** op het gebied van **privacy, gegevensbescherming** en **auteursrecht en de daaraan verwante rechten**. Zo kan de trainingsdata, veelal verkregen via grootschalige (web)scraping[26] van openbare bronnen op internet of andere digitale bronnen, (bijzondere) persoonsgegevens bevatten.[27] Het ontbreekt vaak aan transparantie welke data worden gebruikt en hoe. De gegenereerde inhoud kan onnauwkeurig zijn, verouderd, onjuist, ongepast, beledigend, of aanstootgevend en kan een eigen leven gaan leiden.[28] Het Rathenau Instituut constateert daarbij dat generatieve AI het mogelijk maakt om zeer persoonlijke informatie, zoals iemands stemming of gedachten, af te leiden uit de interactie met de gebruiker.[29] Dit kan leiden tot ongewenste sturing en manipulatie via hyperpersoonlijke content en zogenoemde *dark patterns*, die inspelen op onze verlangens en (onbewuste) cognitieve processen. Dit kan een mogelijke inperking van **gebruikersautonomie** zijn.[30]

Verantwoorde generatieve AI-tools via het Rijks AI-validatieteam

Via het Rijks AI-validatieteam komen we tot (publiek beschikbare) vangrails en tooling voor generatieve AI-modellen. Om verdere kennis en ervaring op te doen met de validatie van AI, heeft het kabinet een (Rijks)-AI-validatieteam opgericht. Dit team buigt zich onder meer over het meetbaar maken van risico's en kansen van generatieve AI. Het team bestaat uit software engineers die samen met beleidsmakers gaan werken aan concrete hulpmiddelen om (generatieve) AI te valideren.

24. ChatGPT Replicates Gender Bias in Recommendation Letters | Scientific American
25. Humans Absorb Bias from AI--And Keep It after They Stop Using the Algorithm - Scientific American
26. Webscraping is het gebruik van software om informatie van webpagina's te extraheren om deze vervolgens te analyseren.
27. Zie ook: Roundtable of G7 Data Protection and Privacy Authorities Statement on Generative AI (21 juni 2023), online via: Roundtable of G7 Data Protection and Privacy Authorities Statement on Generative AI -Personal Information Protection Commission- (ppc.go.jp).
28. Aanhangsel Handelingen II 2022/23, nr. 3381.
29. Rathenau Instituut (2023) Generatieve AI: p. 21.
30. wired.com/story/ai-chatbots-can-guess-your-personal-information/

Een andere uitdaging is de invloed van generieke AI-systemen op de **cognitieve ontwikkeling** van mensen die deze systemen gebruiken.[31] Er worden zorgen geuit over het verlies van kennis en kunde van mensen,[32] met name op het gebied van creativiteit, kritische reflectie en begripsvermogen. Als generatieve AI steeds meer cognitieve vaardigheden overneemt, is er de kans dat mensen deze vaardigheden verliezen. Ook kan de **sociale ontwikkeling** negatief worden beïnvloed als generatieve AI-systemen steeds meer (intieme) menselijke interacties vervangen.[33]

Tot slot kan er bij scraping voor het trainen van een generatief AI-model ook sprake zijn van het gebruik van **auteurs-, nabuur- en databankrechtelijk beschermde werken, andere materialen respectievelijk databanken** en kan de output van generatieve AI inbreuk maken op zowel het **portretrecht** als de bovengenoemde rechten.[34] Zo zijn er in de VS bijvoorbeeld al meerdere rechtszaken aangespannen met betrekking tot auteursrecht en generatieve AI.[35] In hoofdstuk 4 wordt nader ingegaan op de huidige en toekomstige wettelijke kaders (zoals privacy- en gegevensbeschermingswetgeving, de Grondwet, het auteursrecht en de aankomende Europese AI-verordening).

Afhankelijkheid en marktmacht

Er is een **groeiende afhankelijkheid** van een selecte groep techbedrijven. De meest gebruikte generatieve AI-modellen en diensten in Nederland zijn afkomstig van een klein aantal **Amerikaanse techbedrijven**. Deze bedrijven beschikken over grote hoeveelheden data, computerkracht en ontwikkelcapaciteit.[36] Een ander groeiend aandeel van de wereldwijde markt voor generatieve AI-modellen ligt ook buiten Europa, **hoofdzakelijk in China**. De betreffende bedrijven hebben dan ook een betere uitgangspositie dan Europese bedrijven bij het ontwikkelen van generatieve AI-modellen en bepalen daarbij grotendeels in welke richting de technologie zich verder ontwikkelt. Gezien het belang van generatieve AI voor de innovatiekracht en het lange termijn verdienvermogen van Nederland kan dit leiden tot **strategische afhankelijkheden**.[37]

De ontwikkeling van generatieve AI versterkt de **machtsconcentratie** in digitale markten en vergroot hiermee het risico op machtsmisbruik. Schaalvoordelen worden steeds belangrijker, onder andere door het belang van data voor het ontwikkelen van generatieve AI. Dit versterkt de *winner-takes-all-dynamiek*.[38] Dit geldt vooral voor de markt van modellen die gebruikt worden als technologische basis voor toepassingen van generatieve AI.[39] Op deze markt is een selecte groep techbedrijven in onderlinge concurrentie. Ook vindt er integratie plaats van generatieve AI met verschillende (bestaande) diensten (ecosysteemvorming). Voor andere bedrijven werpt dit hoge **toetredingsdrempels** op om hierop ook te concurreren. Innovatieve nieuwkomers krijgen minder kans en ontwikkelaars van specifieke toepassingen kunnen ingesloten raken, wat een productieve marktwerking kan belemmeren. Zo kan dit bijvoorbeeld leiden tot **oneerlijke handelspraktijken**, hogere prijzen voor toegang en gebruik van AI-toepassingen en AI-infrastructuur, en minder keuzemogelijkheden voor consumenten.[40]

De ontwikkeling van generatieve AI is afhankelijk van een handvol grote bedrijven. Dit kan leiden tot ongelijke toegang tot de technologie en onevenredige kansen om deze te benutten. Kleinere bedrijven, onderwijsinstellingen, leraren of leerlingen en (sociaaleconomisch) achtergestelde groepen kunnen worden benadeeld, waardoor sociale verschillen binnen een samenleving als ook tussen samenlevingen wereldwijd kunnen toenemen. Tot slot vormt marktmacht vaak de opmaat voor het uitoefenen van maatschappelijke en politieke invloed.[41]

Nederlands eigen open taalmodellen

Wij stimuleren de ontwikkeling van (open) Nederlandse en Europese LLM's in lijn met publieke waarden. De financiering van GPT-NL is hiervan een voorbeeld. Ook verkennen we in Europees-verband deelname aan o.a. een Europees programma voor taaltechnologie (ALT-EDIC).

31. Rathenau Instituut (2023), Generatieve AI: pp. 24-25.
32. J. Pitt (2023), "ChatSh*t and Other Conversations (That We Should Be Having, But Mostly Are Not), IEEE Technology and Society Magazine, vol. 42, no. 3, pp. 7-13.
33. Danaher, J. (2019) The rise of the robots and the crisis of moral patience. AI & Society 34, 129–136. Rathenau Instituut (2023). Generatieve AI: p. 25.
34. Zie ook: open.overheid.nl/documenten/dpc-c82f1b6b5ce7c6826069b7b857983536oab041ea/pdf
35. nytimes.com/2023/08/21/arts/design/copyright-ai-artwork.htm
36. economist.com/business/2023/09/18/could-openai-be-the-next-tech-giant
37. Agenda Digitale Open Strategische Autonomie | Rapport | Rijksoverheid.nl
38. Marktdynamiek waarbij één of enkele bedrijven zo dominant zijn dat concurrentie bijna niet meer mogelijk is. De markt beweegt dan toe naar een situatie van (quasi-)monopolie (ook wel tipping genoemd).
39. De bedreigingen en kansen voor concurrentie verschillen per marktniveau. Voor ontwikkelaars van toepassingen van generatieve AI liggen er juist vooral veel kansen, bijvoorbeeld in de vorm van alternatieve verdienmodellen en nieuwe vormen van concurrentie binnen bestaande marktstructuren.
40. ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns
41. Rathenau Instituut (2023), Generatieve AI: pp. 30-31.

Arbeid en arbeidsmarkt

Ten aanzien van werk zijn er zorgen over de impact van generatieve AI op **werkgelegenheid**, **de kwaliteit van werk** en **inkomensverdeling.**

De meeste economen verwachten dat de totale **werkgelegenheid** niet zal afnemen als gevolg van (generatieve) AI. Toch kunnen de gevolgen voor werkgelegenheid ongelijk zijn en op korte termijn verstorend werken. De laagdrempelige beschikbaarheid en schaalbaarheid van generatieve AI zullen de snelheid van implementatie van deze automatisering aanzienlijk beïnvloeden. In voor (generatieve) AI kwetsbare beroepsgroepen en sectoren (zoals creatief werk, data-analyse, juridische werkzaamheden en kantoorondersteuning[42]) kan dit leiden tot verandering in taken en/of baanverdringing. Omdat het tijd kost om bij te scholen en de juiste baan te vinden, kan dit op korte termijn gepaard gaan met **toename van werkloosheid**. Als banen snel verdwijnen, kan dit ook leiden tot een **tijdelijke loondaling**.[43] Op middellange termijn kan het langdurige werkloosheid veroorzaken bij werknemers die niet kunnen of willen overstappen naar nieuwe banen, waardoor ze afstand tot de arbeidsmarkt opbouwen (dit heet '*scarring*').

Generatieve AI heeft ook invloed op de kwaliteit van werk. Hoewel het op de werkvloer kansen biedt om de **kwaliteit van werk** te vergroten, is er ook een risico dat taken verschralen doordat het complexe taken overneemt. Dit kan leiden tot beperking van de (ervaren) autonomie van werknemers en druk uitoefenen op de menselijkheid van professionele relaties.[44]

Een andere uitdaging schuilt in de **verdeling van inkomen en werkzekerheid** op de langere termijn. Versnelde en makkelijk toegankelijke automatisering door generatieve AI kan oneven-redig nadelige gevolgen hebben voor werknemers die niet over de vaardigheden beschikken die in nieuwe banen worden gevraagd. Mensen met adequate toegang en vaardigheden en taken die sterk bijdragen aan een individuele productiviteitsverhoging, kunnen uitgroeien tot een kleine groep 'superstar werknemers'.[45] Aan de andere kant kunnen werknemers van wie de vaardigheden worden geautomatiseerd juist aan productiviteit verliezen. Hierdoor kan generatieve AI bijdragen aan **polarisatie op de arbeidsmarkt** en van toenemende inkomensongelijkheid[46], zoals bij eerdere automatiseringsgolven het geval was.[47] Welke groepen (relatief meer) geraakt worden door deze ongelijkheidseffecten is nog ongewis. In tegenstelling tot eerdere technologische disrupties kan generatieve AI een grote impact hebben op relatief hooggeschoolde en hoogbetaalde functies, zowel in positieve als in negatieve zin. Een deel van de taken die bij deze functies horen, die cognitief doch routinematig van aard zijn) lopen een groot risico om geautomatiseerd te worden. Aan de andere kant kunnen hooggeschoolde werkers naar verwachting het meest profiteren van de inzet van generatieve AI.[48] De uiteindelijke ongelijkheidseffecten hangen samen met de inzet van generatieve AI op de werkvloer (als aanvulling of als vervanger), met hoe taken worden verdeeld over functies en werkers en met de kwaliteit en effectiviteit van om- en bijscholing.[49]

De SER onderzoekt AI en de arbeidsmarkt

Het kabinet heeft de Sociaal-Economische Raad (SER) gevraagd om de impact van AI (waaronder generatieve AI) op de arbeidsproductiviteit, kwantiteit en kwaliteit van werk in kaart te brengen.

Invloed op de maatschappij

Maatschappelijke uitdagingen als gevolg van de inzet van generatieve AI doen zich op een aantal fronten voor. Allereerst heeft generatieve AI een potentieel ontwrichtende invloed op het sociale en maatschappelijke domein. Het brede gebruik van deze technologie kan ertoe leiden dat slechts een beperkt aantal mensen en bedrijven (de zogenoemde '*superstar firms*'[50]) ervan profiteert, wat kan leiden tot **toenemende sociale en economische ongelijkheid**.

Een tweede maatschappelijke uitdaging heeft te maken met de **grote energie** die nodig is voor het trainen en gebruiken van generatieve AI-modellen. Als deze energie niet afkomstig is van hernieuwbare bronnen kan generatieve AI een onwenselijk effect hebben op **klimaatverandering**. Zelfs bij gebruik van hernieuwbare bronnen kan generatieve AI ongunstig bijdragen aan klimaatverandering, doordat het energiebronnen opslokt die anders door bestaande sectoren zouden worden gebruikt. Momenteel is de klimaatimpact van generatieve AI-modellen nog relatief beperkt: het trainen van OpenAI's GPT-3 zorgde volgens het Massachusetts Institute of Technology (MIT) voor

42. mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#business-and-society
43. Greenhouse, S. (8 februari 2023) US experts warn AI likely to kill off jobs – and widen wealth inequality. The Guardian.
44. Summary - Artificial intelligence for worker management: an overview | Safety and health at work EU-OSHA (europa.eu)
45. Benzell, S. & Brynjolfsson, E. (2019). 'Digital Abundance and Scarce Genius: Implications for Wages, Interest Rates, and Growth.' National Bureau of Economic Research.
46. AI and the Labor Market - Clark Center Forum (kentclarkcenter.org)
47. Acemogly, Koster & Ozgen (2023). 'Robots and Workers: Evidence from the Netherlands.' National Bureau of Economic Research, working paper no. 31009.
48. Pizzinelli et al. (2023). 'Labor Market Exposure to AI: Cross-Country Differences and Distributional Implications.' International Monetary Fund Working Paper No. 2023/216.
49. technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like
50. Autor, D. et al. (2020). The Fall of the Labor Share and the Rise of Superstar Firms, The Quarterly Journal of Economics.

Onderzoek naar duurzamere inzet generatieve AI

We onderzoeken het duurzaamheidsaspect bij de ontwikkeling en het gebruik van generatieve AI (door de overheid) en waar mogelijk nemen we maatregelen om de negatieve gevolgen te reduceren.

circa 500 ton CO_2-uitstoot.[51] Dit is vergelijkbaar met duizend auto's die duizend kilometer rijden.[52] Nieuwe, grotere modellen vereisen fors meer energie voor zowel trainen als gebruik waardoor de uitstoot zou kunnen oplopen tot tientallen, mogelijk honderden megatonnen CO_2.[53] Naast de training brengt het gebruik van generatieve AI-systemen ook klimaatimpact met zich mee. Uit een studie blijkt dat voor een chatconversatie van 20 tot 50 antwoorden, circa 50 milligram koelwater nodig is. Dit komt overeen met een flesje water per sessie.[54]

Een derde uitdaging betreft de **aantasting van ons informatie-ecosysteem**. Generatieve AI heeft al bewezen dat sneller en op grotere schaal kan bijdragen aan **de creatie en verspreiding van mis- en desinformatie**.[55] Het was al enige jaren mogelijk om bijvoorbeeld deepfakes te genereren. Met generatieve AI-tools als Midjourney, Synthesia en D-ID is de schaal, eenvoud en 'echtheid' waarop dit plaatsvindt aanzienlijk toegenomen.[56] Er zijn diverse campagnes bekend waarbij overheden en anderen, desinformatie verspreiden met behulp van AI-gegenereerde nieuwsuitzendingen die nauwelijks van echt te onderscheiden zijn. Onderzoek toont aan dat mensen niet alleen moeilijk kunnen onderscheiden of een gezicht echt of synthetisch is, maar ook meer vertrouwen hebben in nepgezichten.[57] De ondermijnende uitwerking van desinformatie wordt versterkt doordat (generatieve) AI-systemen ook op elkaar gaan reageren en de content prioriteren in *newsfeeds* en tijdlijnen op sociale media. Hierdoor kan desinformatie nog effectiever verspreid worden en de impact ervan vergroot.[58] Voor een goed functionerende democratie zijn goed geïnformeerde burgers en een gemeenschappelijke visie op de realiteit echter belangrijke randvoorwaarden, evenals een brede steun voor democratische instituties. Onafhankelijke en kwalitatieve nieuws- en informatievoorziening is in dat verband van groot belang. Generatieve AI-toepassingen brengen dit mogelijk in gevaar.[59] Tot nu toe lijken traditionele factcheckmethoden, het informeren en opleiden van gebruikers, en detectietools minder effectief te zijn voor content die is gegenereerd op basis van generatieve AI.[60] Dit betekent dat ook in democratische samenlevingen de verspreiding van desinformatie het maatschappelijk debat en kernprocessen van de democratie, zoals verkiezingen kan destabiliseren of ondermijnen.[61]

Ook de **onzekere betrouwbaarheid** van veel generatieve AI-modellen en toepassingen heeft negatieve invloed op de kwaliteit van ons informatie-ecosysteem. De modellen zijn gebaseerd op kansberekeningen en missen begrip. Er zijn veel voorbeelden waarbij generatieve AI-programma's onbedoeld informatie genereren die niet waar is. Bijvoorbeeld, chatbots als ChatGPT regelmatig naar niet-bestaande (wetenschappelijke) bronnen en 'verzinnen' gegevens die worden gepresenteerd als feiten.[62] Dit zogenoemde 'hallucineren' brengt grote risico's met zich mee als het gaat om waarheidsvinding, vooral omdat veel mensen de neiging hebben tot '*automation bias*' waarbij ze te veel vertrouwen hebben in de uitkomsten van geautomatiseerde systemen.[63]

In het licht van de toenemende mogelijkheden voor manipulatie en desinformatie, is het cruciaal om te begrijpen hoe verschillende regeringen deze technologieën inzetten en reguleren. Toepassingen van generatieve AI bieden autoritaire regimes de mogelijkheid om informatie op een ongekende schaal te controleren, dissidentie te smoren en burgers intensiever te surveilleren, met ernstige gevolgen voor de mensenrechten, waaronder schendingen van privacy en vrijheid van meningsuiting.[64] Dit brengt aanzienlijke uitdagingen met zich mee voor fundamentele vrijheden en mensenrechten wereldwijd.

51. We're getting a better idea of AI's true carbon footprint | MIT Technology Review
52. tudelft.nl/stories/articles/duurzame-kunstmatige-intelligentie-van-chatgpt-naar-groene-ai
53. De Vries, A. (2023). The growing energy footprint of artificial intelligence, Joule (2023).
54. 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). En zie ook: Rathenau Instituut (2023), Generatieve AI.
55. Bontridder & Poullet (2021). The Role of Artificial Intelligence in Disinformation. Data & Policy 3(E32).
56. rijksoverheid.nl/documenten/kamerstukken/2023/06/16/tk-beleidsreactie-op-de-wodc-onderzoeken-naar-de-regulering-van-deepfakes-en-immersieve-technologieen
57. pnas.org/doi/10.1073/pnas.2120481119
58. Dit was bijvoorbeeld het geval toen de oprichter van de open source onderzoeksorganisatie Bellingcat, Eliot Higgins, in maart 2023 op Twitter foto's had gepost van de arrestatie van Donald Trump. Alhoewel Higgins had vermeld dat hij de foto's met de generatieve AI tool Midjourney had gecreëerd, werden de foto's, mede door nieuwskanalen, duizenden keren gedeeld: nytimes.com/2023/04/08/business/media/ai-generated-images.html
59. 'Dat zijn toch gewoon ál onze artikelen?' – De Groene Amsterdammer.
60. OESO (2023). As language models and generative AI take the world by storm, the OECD is tracking the policy implications - OECD.AI
61. Kamerstuk 2023-2024, 35165 Nr. 46.
62. Beutel, Geerits & Kielstein (2023). Artificial Hallucination: GPT on LSD? Critical Care 27(148).Hallucinations Could Blunt ChatGPT's Success - IEEE Spectrum.
63. Goddard et al. (2012), Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators, Journal of the American Medical Informatics Association 19(1): 121-127.
64. The Repressive Power of Artificial Intelligence | Freedom House

Een vierde type maatschappelijke uitdaging betreft **militaire veiligheid**, vooral waar het gaat om de effecten die generatieve AI zal hebben op het internationale veiligheidsdomein. **Systemische veiligheidsrisico's** kunnen voortkomen uit de versteviging van bestaande ongelijkheden door de inzet van generatieve AI, snelle en grootschalige veranderingen op de arbeidsmarkt of door verschuivingen in economische en militaire verhoudingen als gevolg van de inzet van geavanceerde generatieve AI met mogelijke gevolgen voor de geopolitieke verhoudingen.

Een vijfde uitdaging heeft te maken met het opzettelijk **misbruiken van generatieve AI-modellen**. Naarmate deze modellen vaardiger worden (door toegenomen rekenkracht, beschikbaarheid en capaciteiten) nemen ook de mogelijkheden voor risicovol **misbruik** verder toe.[65] Bijvoorbeeld, het gebruik van generatieve AI-modellen om kwetsbaarheden te ontdekken in de computercode en vervolgens volledig autonome en grootschalige cyberaanvallen uit te voeren, zonder menselijke tussenkomst. Generatieve AI zou ook kwaadwillenden kunnen helpen bij het creëren van nieuwe virussen.[66] Zoals al eerder aangeduid zou de technologie ook ingezet kunnen worden om op grote schaal strafbare content te creëren en online te verspreiden , waaronder bedreigingen of *hate speech*. Online zien we al dat bedreigingen tegen politici, bestuurders, journalisten, columnisten en wetenschappers toenemen. Dit brengt het risico met zich mee dat de bereidheid om dergelijke belangrijke functies te vervullen in de **democratische rechtsstaat** afneemt, en dat het democratisch debat niet vrijelijk gevoerd kan worden.[67]

Een zesde maatschappelijke uitdaging betreft het risico op ongelukken. Naarmate generatieve AI-modellen vaardiger worden zullen ze vaker ingezet worden in complexe, maatschappelijke processen. Dit vergroot de kans op én de impact van **ongelukken**, bijvoorbeeld als AI-modellen foutieve resultaten genereren bij cruciale processen. Een andere mogelijkheid is dat AI-systemen min of meer autonoom doelen gaan nastreven op een manier die schade berokkent.[68] Het (nog) ondoorgrondelijke (black box-)karakter van generatieve AI-modellen maakt het moeilijk om dit soort ongelukken te voorkomen. Het is echter ook mogelijk dat generatieve AI aanzienlijk minder fouten leert te maken dan mensen, waardoor het wellicht niet langer wenselijk is dat dergelijke taken door mensen worden uitgevoerd.

65. Generatieve AI kan nu al worden misbruikt - zo circuleert op het dark web een AI-systeem genaamd 'WormGPT' dat automatisch gepersonaliseerde phishing mails kan genereren. Er is ook het risico op datapoisoning: Forcing Generative Models to Degenerate Ones: The Power of Data Poisoning Attacks for NeurIPS 2023 | IBM Research
66. De hiervoor vereiste biotechnologie bestaat al. Generatieve AI kan dit soort technologie echter toegankelijk maken voor een veel groter aantal (kwaadwillende) actoren door kennis te verschaffen of te adviseren over planning en uitvoering van bioterroristische aanvallen. Zie: Anthropic \ Frontier Threats Red Teaming for AI Safety.
67. "Koester de Democratie! Een dringende oproep om de democratische rechtsorde weer voor iedereen te laten werken." Eindrapportage Adviescommissie Versterken Weerbaarheid en Democratische Rechtsorde, 2-11-2023.
68. newscientist.com/article/mg25834382-000-what-is-the-ai-alignment-problem-and-how-can-it-be-solved/

Visie Generatieve AI

Waarom?

De samenleving kan het volle potentieel van generatieve AI benutten als de overheid actief bijdraagt aan veilige en rechtvaardige ontwikkeling en gebruik van generatieve AI, aan generatieve AI die het menselijk welzijn en autonomie dient, en duurzaamheid en onze welvaart vergroot.

Ambitie

Nederland is koploper in Europa op het gebied van veilige en verantwoorde generatieve AI-toepassingen en heeft een sterk landelijk en internationaal ecosysteem waarin volop geïnnoveerd wordt met verantwoorde generatieve AI.

Lerende aanpak

De razendsnelle ontwikkelingen vereisen een iteratieve en lerende aanpak waarbij we van elkaar leren en kennis vergaren over hoe we generatieve AI verantwoord kunnen ontwikkelen en gebruiken. Coördineren en aanjagen zijn hierbij de kernwoorden om te zorgen dat de opgedane kennis en inzichten vanuit de actielijnen bij elkaar komen en effectief benut kunnen worden voor verantwoorde inzet van generatieve AI.

Actielijnen

- A Samenwerken
- B Nauwgezet volgen van alle ontwikkelingen
- C Vormgeven en toepassen wet- en regelgeving
- D Kennis/kunde vergroten
- E Innoveren met generatieve AI
- F Sterk en helder toezicht houden en handhaven

4 Visie op generatieve AI

Dit hoofdstuk presenteert de overheidsbrede visie op generatieve AI. Het vertrekpunt hierbij is een waardengedreven benadering, aansluitend bij de Werkagenda Waardengedreven Digitaliseren,[1] de Agenda Coalities voor de Digitale Samenleving,[2] de Strategie Digitale Economie[3] en het gecoördineerde plan inzake AI van de EU.[4]

Deze benadering is de basis onder vier **centrale uitgangspunten** en die leidend zijn bij de ontwikkeling, toepassing en inbedding van generatieve AI in onze samenleving. Zoals nader toegelicht zal worden, streeft het kabinet naar generatieve AI die **veilig** en **rechtvaardig** is en bijdraagt aan **menselijk welzijn, duurzaamheid** en **welvaart**. Nederland heeft met haar waardengedreven aanpak de kans om een leidende rol te spelen in Europa en op wereldniveau. Om deze visie te realiseren, zijn **nieuwe acties** vanuit de Nederlandse overheid vereist. Hoofdstuk 5 biedt een overzicht hiervan.

Voorafgaand aan de totstandkoming van de visie is gekeken naar **bestaand beleid en de geldende wet- en regelgeving** die van toepassing is op (generatieve) AI. Dit is geen eenmalige activiteit. Gezien de snelle ontwikkeling van generatieve AI is het van belang om te blijven monitoren welke publieke waarden onder druk staan, welke uitdagingen en kansen generatieve AI met zich meebrengt (zie voorgaand hoofdstuk), en in hoeverre beleid hier invulling aan geeft. Het eerste deel van dit hoofdstuk is aan deze wetgevings- en beleidsanalyse gewijd. Het tweede deel gaat in op de vier uitgangspunten voor de overheidsbrede visie op generatieve AI. In hoofdstuk 5 wordt beschreven met welke acties het kabinet deze visie concretiseert.

a Huidige wet- en regelgeving en beleid

Deze paragraaf bespreekt het bestaande beleid en de wet- en regelgeving die van toepassing is op zowel de ontwikkeling als het gebruik van generatieve AI. Daarnaast wordt ingegaan op de Europese AI-verordening. Tot slot wordt er aandacht besteed aan internationale ontwikkelingen op generatieve AI-gebied.

i AI-beleid

De Nederlandse overheid werkt al langere tijd aan beleid op het gebied van AI. Voor een groot deel is het ingezette beleid (deels opgesteld voor traditionelere of '*narrow* AI') ook passend voor de uitdagingen en kansen die generatieve AI biedt. De brede beschikbaarheid van generatieve AI, de schaal en het tempo waarop het zich momenteel ontwikkelt, vraagt echter om een toekomstbestendige visie met een daaraan verbonden nieuwe acties om daarmee de toegenomen risico's optimaal het hoofd te bieden of mogelijkheden te benutten.

AI heeft hoge prioriteit in het Nederlands beleid voor de digitale samenleving. Het is een sleuteltechnologie waarmee we willen meedoen met de koplopers in Europa.[5] Sinds 2019 is het Strategisch Actieplan voor AI (SAPAI) in werking dat beoogt de kansen van AI te verzilveren en de publieke belangen bij AI te borgen. De Werkagenda Waardengedreven Digitalisering noemt de beleidsprioriteiten rond het beschermen van publieke waarden bij AI. Op het gebied van AI gaat dit onder andere over het eerlijk en transparant maken van de toepassing van algoritmen. Daarbij is het van belang dat iedereen kan deelnemen aan het digitale tijdperk, de digitale wereld kan vertrouwen en de regie over zijn digitale leven heeft. De Strategie Digitale Economie gaat onder andere in op het verzilveren van kansen en stroomlijnen van de AI-markt.

1. rijksoverheid.nl/documenten/rapporten/2022/11/04/bijlage-1-werkagenda-waardengedreven-digitaliseren
2. open.overheid.nl/documenten/10c88500-cdb5-4815-bd00-c915a5242ea3/fil
3. open.overheid.nl/documenten/ronl-c6a3495a523bef54ca41011f629b77b7b611045f/pdf
4. digital-strategy.ec.europa.eu/nl/library/coordinated-plan-artificial-intelligence-2021-review
5. Zie ook: open.overheid.nl/documenten/ronl-e14cdcee-690c-4995-9870-fa4141319d6f/pdf

Zoals ook benoemd in de verzamelbrief Algoritmes reguleren[6] van juli 2023, neemt het kabinet verschillende stappen om grip te krijgen op AI. Dit zijn daarbij belangrijke initiatieven:

- Inzetten op **verantwoorde AI-toepassingen**. Via de **Nederlandse AI-Coalitie (NL-AIC)** werken overheid, bedrijfsleven, onderwijs- en onderzoeksinstellingen en maatschappelijke organisaties samen aan maatschappelijk verantwoorde AI-toepassingen. Onder andere via labs waarin door wetenschappers, ondernemers en publieke instellingen – de zogenoemde ELSA-labs – **onderzoek** wordt gedaan naar de **ethische, juridische en sociale aspecten van AI.** De NL-AIC heeft een **AI-cursus** ontwikkeld die voor iedereen gratis beschikbaar is.
- **Het AiNed programma** is een publiek-privaat meerjarenprogramma binnen het Nationaal Groeifonds. Het programma heeft de ambitie Nederland blijvend in de kopgroep van AI-landen te brengen en wil een bijdrage leveren aan economisch herstel en groei, structurele versterking van de economische basis in Nederland, én aan een mensgerichte en verantwoorde toepassing van AI. Via AiNed is er de afgelopen jaren o.a. geïnvesteerd in het aantrekken van uitzonderlijk AI-talent en is het aantal Nederlandse partijen dat kan deelnemen aan AI-onderzoek- en innovatieprojecten met Europese samenwerking vergroot.
- Het is belangrijk dat de overheid een ondersteuningsstructuur faciliteert die de **ontwikkeling van AI voor het onderwijs** in goede banen leidt. Met financiering van het Nationaal Groeifonds investeren de ministeries van EZK en OCW voor een periode van tien jaar substantieel in **het Nationaal Onderwijslab AI (NOLAI).** Hierin werken leraren, wetenschappers en bedrijven aan een verantwoorde ontwikkeling en evaluatie van geavanceerde digitale innovaties als AI in het funderend onderwijs.[7] Zo wordt er bijvoorbeeld met behulp van AI gewerkt aan een centraal lerarendashboard en wordt onderzocht hoe AI kan bijdragen aan persoonlijke ondersteuning van leerlingen in hun leerproces.
- Aanvullend ontwikkelt het Nationaal Groeifondsprogramma Npuls onder meer een **landelijk AI-punt en AI-visie voor het mbo, hbo en wo** om zo de sectoren voor te bereiden op de transformatie van het onderwijs en om samen met partners en instellingen deze veranderingen ook mede vorm te geven.
- Op het gebied van **veilige AI** zijn al grote investeringen gedaan. Via het Innovation Center for Artificial Intelligence (ICAI) wordt in een samenwerking tussen het bedrijfsleven, de overheid en de kennissector volop geëxperimenteerd en onderzoek gedaan (totaalbudget van 87 miljoen voor het ROBUST programma). Zo zijn er via ICAI de afgelopen jaren in verschillende Nederlandse steden AI-labs gestart, waarin samenwerking plaatsvindt tussen overheden, bedrijven en wetenschap.

ii Wet- en regelgeving in Nederland en de EU

Voor de ontwikkeling en toepassing van generatieve AI zijn verschillende juridische kaders van toepassing. Hieronder wordt ingegaan op hoe de ontwikkeling van generatieve AI zich verhoudt tot fundamentele rechten en wordt aandacht besteed aan de Europese AI-verordening.

Fundamentele rechten

Door de ontwikkeling en het gebruik van generatieve AI kan de verwezenlijking van fundamentele rechten onder druk komen te staan. De fundamentele rechten die geraakt worden door generatieve AI zijn het discriminatieverbod en het recht op privacy en gegevensbescherming.

Het discriminatieverbod is samen met het gelijkheidsbeginsel opgenomen in artikel 1 van de Grondwet. Op Europees niveau is het verbod op discriminatie onder andere verankerd in artikel 14 van het Europees Verdrag voor de Rechten van de Mens (EVRM) en op internationaal niveau in artikel 26 van het Internationaal Verdrag inzake burgerrechten en politieke rechten (IVBPR). Zoals al eerder in deze visie naar voren is gebracht (in hoofdstuk 3b), kan generatieve AI vooringenomenheid of discriminatie in de hand werken. Bias kan op verschillende manieren in de generatieve AI-systemen terecht komen, onder andere via ontwikkelaars en trainingsdata. Omdat een handvol AI-ontwikkelaars generatieve AI en toepassingen vormgeven, kan (onbewust) bias in de modellen terechtkomen. Daarnaast kunnen de trainingsdata bias bevatten en in het model terechtkomen, waardoor deze bias op grote schaal wordt verspreid en versterkt.[8]

Bij het gebruik en het ontwikkelen van generatieve AI kan inbreuk worden gemaakt op het recht op privacy en het recht op gegevensbescherming. In de Grondwet is het recht op eerbiediging van de persoonlijke levenssfeer neergelegd in artikel 10. Daarnaast wordt het recht op privacy beschermd door artikel 8 van het Europees Verdrag voor de Rechten van de Mens

6. Kamerstuk 2022-2023, 26 643 nr. 1056.
7. rijksoverheid.nl/documenten/kamerstukken/2023/07/06/visiebrief-digitalisering-in-het-funderend-onderwijs
8. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, Volume 3, 2023.

(*Right to respect private and family life*) en artikel 17 van het Internationaal Verdrag inzake burgerrechten en politieke rechten.

Privacy- en gegevensbeschermingsrecht

Krachtens artikel 8 lid 1 van het Handvest van de grondrechten van de Europese Unie en artikel 16 lid 1 van het Verdrag betreffende de werking van de Europese Unie heeft eenieder recht op bescherming van diens persoonsgegevens. De belangrijkste algemene regels voor het verwerken van persoonsgegevens staan in de Algemene verordening gegevensbescherming (AVG) en, aanvullend daarop voor Nederland, de Uitvoeringswet AVG (UAVG). Deze regels gelden zowel voor overheden en private partijen wanneer zij onder de reikwijdte van deze wetgeving vallen. Bevoegde autoriteiten met taken op het gebied van de rechtshandhaving vallen onder afwijkende regels voor de bescherming van persoonsgegevens. Dit is de Wet politiegegevens (Wpg) en de Wet justitiële en strafvorderlijke gegevens (Wjsg). Het gegevensbeschermingsrecht bevat normen voor de behoorlijke en rechtmatige verwerking van persoonsgegevens, respectievelijk politiegegevens zoals algemene beginselen en grondslagen, regels over transparantie, over de beveiliging van persoonsgegevens en rechten van betrokkenen zoals het recht op inzage, correctie en verwijdering. Voor het verwerken van bijzondere categorieën persoonsgegevens, zoals gegevens waaruit etnische afkomst blijkt, biometrische gegevens met het oog op de unieke identificatie van een persoon of gegevens over gezondheid geldt in principe een verbod, tenzij aan strikte voorwaarden uit de wet wordt voldaan. Om te voorkomen dat er een ernstig risico op omzeiling zou ontstaan, is het recht op gegevensbescherming onafhankelijk van gebruikte technologieën, zoals generatieve AI. De Telecommunicatiewet (Tw) is als lex specialis aanvullend op de AVG. Volgens de Tw is er als hoofdregel toestemming vereist van de internetgebruiker wanneer er een inmenging is in de vertrouwelijke communicatie of wanneer er cookies of vergelijkbare technieken worden gebruikt.

Het toezicht op de naleving van de wetgeving voor de bescherming van persoonsgegevens is in handen van een onafhankelijke toezichthouder. In Nederland is de Autoriteit Persoonsgegevens (AP) daartoe bevoegd. De AP heeft een mandaat en bevoegdheden om te onderzoeken of partijen voldoen aan hun verplichtingen op grond van de wetgeving voor de bescherming van persoonsgegevens en op grond daarvan de nodige corrigerende maatregelen te nemen. Het toezien op de rechtmatigheid van gegevensverwerkingen in de private sector in concrete zaken, is dan ook geen taak van het kabinet.

De European Data Protection Board (EDPB) is een onafhankelijk orgaan waarin alle nationale privacytoezichthouders uit de EU en de European Data Protection Supervisor (EDPS) samenwerken. Omdat generatieve AI een grensoverschrijdend fenomeen is dat vraagt om een geharmoniseerde aanpak, heeft de EDPB een taskforce ChatGPT ingesteld om de samenwerking en informatie-uitwisseling over mogelijke handhavingsmaatregelen te bevorderen.[9] De Autoriteit Consument en Markt (ACM) is de toezichthouder op de naleving van de Tw.

Auteursrecht

Het auteursrecht bevat regels voor de bescherming van werken van letterkunde, wetenschap of kunst tegen ongeautoriseerde openbaarmaking en verveelvoudiging. De Auteurswet bevat een beperking op het verveelvoudigingsrecht ten behoeve van tekst- en datamining[10] waarvan gebruik kan worden gemaakt om AI te trainen. Iedereen kan zich op die beperking beroepen maar alleen wanneer op rechtmatige wijze toegang tot de werken is verkregen. Deze beperking geldt niet wanneer rechthebbenden op passende wijze uitdrukkelijk het recht hebben voorbehouden dat er kopieën mogen worden gemaakt ten behoeve van tekst- en datamining. In het geval van online ter beschikking gestelde werken moet het voorbehoud worden gemaakt met machinaal leesbare middelen, zoals in de metadata waar de voorwaarden voor het gebruik van een website worden geduid. Als een correct voorbehoud is gemaakt, dan is weer voorafgaande toestemming van de rechthebbenden vereist om de voor tekst- en datamining benodigde kopieën te kunnen maken. Aan het verlenen van toestemming kunnen voorwaarden worden verbonden, zoals het betalen van een vergoeding en bronvermelding waaronder de naam van de maker. Voor het trainen van generatieve AI-modellen met kopieën van de nabuur- en databankrechtelijk beschermde prestaties gelden soortgelijke regels.

Een voortbrengsel (output) van generatieve AI als zodanig komt niet voor auteursrechtelijke bescherming in aanmerking. Er is namelijk geen sprake van een schepping van de menselijke geest. Als er sprake is van samenwerking tussen een mens en een AI-systeem kan het voortbrengsel onder omstandigheden wel voor auteursrechtelijke bescherming in aanmerking komen. Daarvoor moeten creatieve keuzes zijn gemaakt bij de uitwerking van het idee in de prompt. Waarschijnlijk is de menselijk invloed op het uiteindelijke resultaat nog niet voldoende om van een werk in de auteursrechtelijke betekenis van het woord te kunnen spreken. Dat kan wel veranderen door later creatieve keuzes toe te voegen tot het uiteindelijk werk dat openbaar wordt gemaakt. Het is niet uitgesloten dat de output inbreuk maakt op het auteursrecht, de naburige rechten, het databankenrecht of het portretrecht.

Het is aan de rechter, en uiteindelijk aan het Hof van Justitie van de EU, om een oordeel te geven of huidige generatieve AI-modellen een succesvol beroep kunnen doen op de tekst- en datamining-exceptie. Op dit moment is de rechter nog niet in de gelegenheid gesteld om hierover te oordelen. Buiten de EU zijn er wel al verschillende rechtszaken aangespannen waarbij de vraag centraal stond of het trainen van (generatieve) AI-modellen een inbreuk op het auteursrecht.[11] Er is een realistische kans dat bij de ontwikkeling van generatieve AI-tools auteursrechtinbreuken plaatsvinden, omdat er geen

9. edpb.europa.eu/system/files/2023-09/20230919-20plenagenda_public.pdf
10. Een geautomatiseerde analysetechniek die gericht is op de ontleding van tekst en gegevens in digitale vorm om informatie te genereren, zoals patronen, trends en onderlinge verbanden.
11. Rathenau Instituut (2023), Generatieve AI.

rechtmatige toegang tot de werken bestaat en/of de gemaakte voorbehouden niet worden gerespecteerd.[12] Daarom is het van belang om de komende jaren te blijven monitoren en, indien nodig, in Europees verband het beleid ter zake aan te scherpen.

Bescherming bedrijfsgeheimen

Innovatieve bedrijven worden steeds meer blootgesteld aan praktijken die zijn gericht op het onrechtmatige verkrijgen van bedrijfsgeheimen, zoals ontvreemding, kopiëren zonder toestemming, economische spionage of inbreuk op vertrouwelijkheidsvereisten, zowel van binnen als van buiten de EU. Zoals bepaald in de Wet bescherming bedrijfsgeheimen gaat het bij bedrijfsgeheimen om knowhow en bedrijfsinformatie die waardevol is omdat zij geheim is en die ook bedoeld is om vertrouwelijk te blijven. De houder heeft hiervoor ook maatregelen genomen om deze geheim te houden. Ontwikkelingen zoals generieke AI brengen echter een verhoogd risico met zich mee. Bijvoorbeeld, in aanvulling op hetgeen hiervoor ten aanzien van auteursrecht is aangegeven, kan in de trainingsfase van generatieve AI het model gevoed worden met gegevens die aangemerkt worden als een bedrijfsgeheim. Omdat investeringen in intellectueel kapitaal door bedrijven van invloed is op hun innovatief vermogen en concurrentiepositie dient scherp te worden toegezien dat toepassing van generieke AI dit niet aantast. Als dit wel het geval is, kan dit negatieve invloed hebben het rendement en de wil om verder te innoveren.

Mededinging en marktordening

Effectieve marktwerking is een randvoorwaarde om Nederlandse bedrijven voldoende keuze te bieden tegen een eerlijke prijs en is een stimulans voor innovatie. Ervaringen uit het verleden met machtsconcentratie en afhankelijkheid op andere technologiemarkten hebben geleerd dat de voordelen van sommige technologieën te veel in handen blijven van enkele grote technologiebedrijven en weinig worden doorgegeven aan ondernemers en consumenten. Dit kan de productiviteitsgroei in de bredere economie op den duur belemmeren. Eenzelfde dynamiek is zichtbaar op het gebied van generatieve AI, bijvoorbeeld in de positie van ontwikkelaars van modellen. Het mededingingsrecht kan door toezichthouders worden ingezet om concurrentieverstorend gedrag (zoals misbruik van economische machtspositie) tegen te gaan en te voorkomen dat toegang ondeugdelijk wordt belemmerd.

Daarnaast bevat de Europese Digital Markets Act (DMA) specifieke regels voor de grootste online platforms, zogeheten poortwachters. Dit zijn platforms waar ondernemers en consumenten nauwelijks meer omheen kunnen. Veel van deze poortwachters hebben ook een belangrijke positie verworven in AI-markten en kunnen hun marktmacht op andere technologiemarkten zoals cloud computing inzetten om die positie te verstevigen (overheveling). De regels in de DMA zijn erop gericht om de betwistbaarheid van de positie van poortwachters te vergroten en de afhankelijkheid van die poortwachters te verminderen. Zo bevat de DMA bijvoorbeeld interoperabiliteits- en dataverplichtingen en diverse verboden om overheveling van marktmacht tegen te gaan. De DMA biedt mogelijkheden voor toepassing van de regels op AI-markten. Zo vallen diverse AI-toepassingen mogelijk al binnen de reikwijdte van de DMA en biedt de DMA de flexibiliteit om op basis van marktonderzoek zo nodig de reikwijdte te verbreden en aanvullende verplichtingen op te nemen.

Europese AI-verordening

De AI-verordening vormt het belangrijkste wetgevende kader voor de ontwikkeling en het gebruik van AI in de EU. Ook voor generatieve AI worden er eisen gesteld en toezicht op die eisen ingericht. Op 8 december 2023 is in Brussel een voorlopig politiek akkoord bereikt over de AI-verordening. Na goedkeuring door alle EU-lidstaten en het hele Europees Parlement treedt de wet in werking. Het doel van deze Europese wet is dat er veilige AI-systemen op de interne markt komen met waarborgen voor de bescherming van gezondheid en fundamentele rechten. Om dit te bereiken gaan eisen gelden voor AI-systemen op basis van het risico dat deze met zich meebrengen. Sommige AI-praktijken worden verboden en andere AI-systemen worden aan hoge eisen onderworpen vanwege het risicovolle toepassingsgebied, zoals bij werving en selectie of voor rechtshandhaving. De verordening geldt direct als wet in Nederland. Een deel ervan, zoals het toezicht op verboden en hoog-risico AI-toepassingen, wordt via een Nederlandse wet verder ingericht. Generatieve AI en de krachtige AI-modellen die vaak hieraan ten grondslag liggen en voor een breed scala aan toepassingen kan worden ingezet, ook wel 'general purpose AI' (GPAI) modellen genoemd, vallen ook onder de AI-verordening. Zo worden aan alle GPAI-modellen transparantie-eisen gesteld, zodat bedrijven die met specifieke AI-applicaties voortbouwen op deze modellen toegang hebben tot o.a. de benodigde technische documentatie om aan de eisen van de AI-verordening te voldoen. Voor de krachtigste GPAI-modellen met systeemrisico's gaan aanvullende verplichtingen gelden op het gebied van risicomanagement, monitoring van ernstige incidenten, en het uitvoeren van modelevaluaties. Deze verplichtingen worden geoperationaliseerd via praktijkcodes die de Europese Commissie samen met de industrie, de wetenschap, maatschappelijke organisaties en andere belanghebbenden gaat ontwikkelen. Er wordt een Europese toezichthouder ingericht binnen de Europese Commissie, de AI Office, die de nieuwe regels voor GPAI-modellen gaat handhaven.

Voor generatieve AI-systemen, zoals chatbots en systemen die afbeeldingen en video's genereren, staan in de AI-verordening aanvullende transparantie eisen. Aanbieders van generatieve AI-systemen moeten ervoor zorgen dat het voor mensen duidelijk is dat ze met een AI interacteren of dat content door AI is gemaakt.

Deze benadering sluit aan bij de inzet van Nederland voor de AI-verordening. Het kabinet vindt het proportioneel dat verplichtingen worden opgelegd aan bedrijven en overheden om mensen te beschermen tegen bepaalde risico's die AI-toepassingen met zich mee kunnen brengen. Dit is belangrijk voor veilige ontwikkeling en gebruik van (generatieve) AI, en daar-

12. Zie het advies van de Landsadvocaat (2023) inzake gebruik generatieve AI-tools: open.overheid.nl/documenten/16d72572-da6b-422c-8cf8-cdc95a523093/file

mee voor het vertrouwen vanuit de maatschappij en de markt om de kansen die deze technologie biedt te benutten.

Over de Europese AI-verordening wordt sinds 21 april 2021 onderhandeld tussen de EU-lidstaten en met het Europees Parlement. Nadat de EU-lidstaten en het Europees Parlement de wet hebben goedgekeurd, treedt de wet in werking. Op basis van het politieke akkoord van december 2023 hebben Nederlandse overheden en bedrijven dan tussen een half en twee jaar de tijd om te zorgen dat AI-systemen die worden ontwikkeld, gekocht en gebruikt aan de eisen van de AI-verordening voldoen. Deze termijn is afhankelijk van het risico ervan, zo zijn sommige AI-praktijken al na 6 maanden verboden. Voor de hoog-risico AI-toepassingen gelden termijnen van 24 en 36 maanden en in die periode wordt de uitvoeringswet opgesteld in consultatie met belanghebbenden en behandeld door het parlement. Voor GPAI-modellen, waaronder de meeste grote AI-modellen die content genereren vallen, zijn de eisen na 12 maanden van toepassing en binnen die termijn wordt het Europese toezicht ingericht.

iii Internationale ontwikkelingen

Generatieve AI is een grensoverschrijdend fenomeen. Gezien de impact van deze technologie op de wereldbevolking en op geopolitieke en internationale verhoudingen, is het van belang dat Nederland een actieve rol speelt op het internationale toneel. Hieronder worden verschillende belangrijke ontwikkelingen op het gebied van (generatieve) AI geschetst.

- **Het CAI (Comité voor AI) van de Raad van Europa** is bezig met de ontwikkeling van een AI-verdrag, gericht op het reguleren van AI-systemen in overeenstemming met de normen van de Raad op het gebied van mensenrechten, democratie en de rechtsstaat. De Europese Commissie onderhandelt namens de EU en werkt nauw samen met EU-lidstaten. De streefdatum van het verdrag is april 2024.
- **De Global Partnership on AI (GPAI)** is een initiatief van Frankrijk en Canada ter bevordering van grensoverschrijdende samenwerking tussen experts die werken aan verantwoorde AI.
- **De AI-expertgroep van de OESO (AIGO)** werkt aan de uitvoering van de AI-principes van de OESO, onderzoek op uiteenlopende terreinen en uitwisseling van best practices op het gebied van AI in de OESO en andere landen. Nederland participeert in deze groep om samen met andere OESO-lidstaten bij te dragen aan verantwoorde en ethische AI-technologieën.
- **Het G7 Hiroshima AI-proces** richt zich op het opstellen van internationale richtsnoeren voor organisaties die geavanceerde AI-systemen ontwikkelen, en hebben tot doel veilige, beveiligde en betrouwbare AI wereldwijd te bevorderen. De niet-uitputtende lijst van leidende beginselen wordt besproken en uitgewerkt als een levend document om voort te bouwen op de bestaande AI-beginselen van de OESO in reactie op recente ontwikkelingen in geavanceerde AI-systemen.
- De **EU-US Trade and Technology Council**, opgericht tijdens de EU-VS top op 15 juni 2021 in Brussel, dient als een forum voor de VS en de EU om kwesties ten aanzien van belangrijke mondiale handels-, economische en technologie kwesties te coördineren, en om trans-Atlantische handels- en economische betrekkingen te versterken. Binnen deze council werken de EU en de VS samen aan o.a. de ontwikkeling van betrouwbare AI.[13]
- **Het VN High Level Advisory Body on AI**, opgericht door de Secretaris Generaal van de Verenigde Naties, heeft als taak aanbevelingen te formuleren voor internationale AI-governance structuur. Deze adviesraad bestaat uit 39 experts, publiceerde een tussentijds rapport[14] in december 2023 en zal een eindadvies presenteren in september 2024 tijdens de 'Summit for the Future'.[15]
- Ook **UNESCO** is actief op het gebied van AI en is ook partner in het Global Partnership on AI (GPAI). In 2021 is, mede op initiatief van Nederland, een aanbeveling voor de ethiek van AI aangenomen, in 2023 gevolgd door beleidsstukken over ChatGPT en AI Foundational Models in relatie tot de aanbeveling. Nederland wil blijvend aandacht genereren voor de implementatie van deze aanbeveling en daarbij de relatie tussen AI, ethiek en mensenrechten bevorderen.
- Op initiatief van de regering Biden hebben vooraanstaande Amerikaanse AI-bedrijven, op 21 juli 2023, de **"Voluntary Commitments on AI"** ondertekend, een reeks vrijwillige principes die de nadruk legt op veiligheid, vertrouwen en transparantie in AI-ontwikkeling. Parallel hieraan richtte de VS de AI Contact Group op, een cross-regionale groep van 21 landen, inclusief Nederland. Daarnaast heeft president Biden (VS) op 30 oktober 2023 het eerste presidentieel decreet (*executive order*) over AI uitgevaardigd. Het decreet heeft veilige, beveiligde, betrouwbare AI en AI-innovatie als doel, is toegespitst op de AI-ontwikkelingen en anticipeert op toekomstige AI-ontwikkelingen.[16]
- Nederland heeft in februari 2023 het initiatief genomen voor een eerste internationale dialoog over verantwoord gebruik van AI in het militaire domein, door het organiseren van de **REAIM summit** in Den Haag. Een belangrijk resultaat is de ondertekening van een Call to Action door meer dan 50 landen. Het kabinet beschouwt de Call to Action als een belangrijke eerste stap om zo veel mogelijk landen en andere stakeholders te betrekken bij de internationale agendavorming op dit thema en als brede basis voor verdergaande gesprekken over internationale kaders

13. commission.europa.eu/strategy-and-policy/priorities-2019-2024/stronger-europe-world/eu-us-trade-and-technology-council_en
14. https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf
15. https://www.un.org/techenvoy/ai-advisory-body
16. whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

voor verantwoorde toepassing van AI in het militaire domein. Nederland is co-host van de volgende editie van REAIM. Deze vindt in 2024 plaats in Zuid-Korea.

- Tijdens REAIM is **de Global Commission on Responsible AI in the Military Domain** gelanceerd, deze commissie moet over twee jaar een zo concreet mogelijk en beleidsmatig relevant advies opleveren over de governance van militaire AI, en daarmee belangrijke input vormen voor gesprekken/ onderhandelingen tussen staten op dit dossier (tegen die tijd zeer waarschijnlijk in VN context).
- Nederland levert een actieve bijdrage aan de Data and Artificial Intelligence Review Board (**DARB**) van de NAVO. Deze board ontwikkelt een certificeringsnorm voor kunstmatige intelligentie in het militaire domein. De norm moet ervoor zorgen dat bedrijven en instellingen binnen het bondgenootschap handelen overeenkomstig het internationaal recht en de normen en waarden van de NAVO. Binnen de DARB is er speciale aandacht voor de kansen die Generatieve AI biedt voor militaire toepassingen, met inachtneming van verantwoorde inzet zoals in de norm wordt vastgelegd.

b Vier uitgangspunten voor generatieve AI

De vorige paragraaf maakt duidelijk dat het kabinet op nationaal, Europees en mondiaal niveau werkt en bijdraagt aan regulering en stimulering van AI, waaronder generatieve AI. De regels en initiatieven die hieruit voortkomen zullen helpen om als samenleving op verantwoorde wijze te laten profiteren van generatieve AI. Tegelijkertijd moet er rekening worden gehouden met de *mogelijkheid* dat bestaand beleid bepaalde risico's niet adequaat dekt en bepaalde randvoorwaarden voor het benutten van kansen niet afdoende borgt.[17] Gelet op de verwachte impact van generatieve AI kunnen beleidshiaten grote gevolgen hebben voor mens, economie en samenleving. Dat kan aanleiding geven tot het inventariseren van mogelijke aanvullende beleidsacties en het borgen van randvoorwaarden. Dit vergt van de overheid een **proactieve en lerende houding, visie en durf**. Het effectief adresseren van zowel de mogelijkheden als de uitdagingen vragen om wendbaarheid in beleidskeuzes. Hierbij is open samenwerking tussen overheid, bedrijfsleven en wetenschap van belang, om zo vroegtijdig signalen op te vangen en bij te sturen. Een AI-ecosysteembenadering past bij snelle technologische ontwikkelingen rondom generatieve AI, die impact heeft op maatschappij, markt en burger.

De overheidsbrede visie op generatieve AI is gebaseerd op vier waardengedreven uitgangspunten die onder meer aansluiten bij de Werkagenda Waardengedreven Digitaliseren[18], de Strategie Digitale Economie, de Agenda Coalities voor de Digitale Samenleving. De Nederlandse overheid zet zich in voor **veilige** en **rechtvaardige** ontwikkeling en gebruik van generatieve AI, en voor generatieve AI die **menselijk welzijn, duurzaamheid en autonomie** dient, en **duurzaamheid** en **onze welvaart** vergroot. Deze vier uitgangspunten verwoorden streefdoelen (zoals ook in lijn met de Duurzame Ontwikkelingsdoelen (SDG's[19])) van het kabinet bij de ontwikkeling, het gebruik en de inbedding van generatieve AI. Daarnaast geven ze inzicht in de wijze waarop generatieve AI ingrijpt op de verschillende waarden die in de uitgangspunten centraal staan.

De (voorziene) beleidsacties dragen bij aan de realisatie van meerdere uitgangspunten en worden ten behoeve van het overzicht separaat gepresenteerd in hoofdstuk 5. Deze beleidsacties zullen niet echter voldoende zijn. Door middel van een lerende en iteratieve aanpak dient de komende jaren op frequente basis de noodzaak van eventuele (nieuwe) acties of beleid te worden geëvalueerd. Hierbij zullen ook de medeoverheden en uitvoeringsorganisaties actief betrokken blijven, zeker ook met het oog op uitvoerbaarheid.

1 Uitgangspunt 1: Generatieve AI wordt op een veilige manier ontwikkeld en toegepast

De Nederlandse overheid zet zich in voor veilige ontwikkeling en gebruik van generatieve AI-systemen. Meer specifiek betekent dit dat we actief bijdragen aan de mitigatie van **misbruik, ongelukken** en **systemische veiligheidsrisico's** van en door generatieve AI-modellen. Dit doen we op nationaal, Europees en internationaal niveau omdat veel van de mogelijke veiligheidsrisico's zich niet aan landsgrenzen houden en niet door unilaterale inzet van Nederlandse overheid kunnen worden verholpen.

Mitigatie van misbruik van generatieve AI-systemen

Zoals beschreven in hoofdstuk 3, kunnen generatieve AI-modellen door kwaadwillende actoren misbruikt worden wanneer ze onvoldoende robuuste veiligheidsmechanismen bevatten. De nieuwste generatie AI-modellen bevat al meer

17. Ook het Rathenau Instituut waarschuwt voor de mogelijkheid dat bestaande en aanstaande beleidskaders niet zijn opgewassen tegen de mogelijkheden geboden door generatieve AI. Zie: Rathenau Instituut (2023). Generatieve AI: p. 33.
18. Iedereen kan meedoen in het digitale tijdperk, iedereen kan de digitale wereld vertrouwen en iedereen heeft regie op het digitale leven.
19. Zie ook: rijksoverheid.nl/onderwerpen/ontwikkelingssamenwerking/internationale-afspraken-ontwikkelingssamenwerking

'vangrails' om misbruik te voorkomen. Toch is het nog steeds mogelijk om modellen te 'kraken' ('jailbreaken' in jargon).[20] Open-source-modellen zijn hier mogelijk extra kwetsbaar voor.[21] Het gevolg is dat generatieve AI-modellen ook nu al kunnen worden misbruikt om desinformatie en phishing e-mails te genereren,[22] om manipulatieve content te maken en om deepfakes te produceren. Toegenomen vaardigheden van toekomstige generaties generatieve AI-modellen zouden nog ingrijpender misbruik kunnen faciliteren, zoals geautomatiseerde cyberaanvallen, of hulp bij de synthese van gevaarlijke virussen of chemische stoffen. Misbruik van generatieve AI-modellen houdt zich niet aan landsgrenzen, zoals ook al beschreven in hoofdstuk 2. Om die reden zet de Nederlandse overheid in Europees verband in op het **beperken van mogelijkheden op misbruik**, en op het stimuleren van technieken die het misbruikpotentieel verkleinen. Daarnaast investeren we in de nationale weerbaarheid tegen misbruik, bijvoorbeeld in het cyberdomein (zie bijvoorbeeld actielijn 'vergroten kennis en kunde').

Mitigatie van ongelukken door gebruik van generatieve AI-systemen

Het kabinet stimuleert de ontwikkeling van intrinsiek veilige generatieve AI-modellen– gebruik van generatieve AI-modellen mag immers niet leiden tot (grootschalige) ongelukken. Ontwikkelaars dienen daartoe te voorkomen dat systemen foutieve of risicovolle informatie genereren. Dit is geen eenvoudige opgave: ondanks flinke inspanningen van AI-ontwikkelaars om zogenaamde hallucinatie te beperken, genereren tekstmodellen nog regelmatig met grote stelligheid onjuiste of zelfs gevaarlijke output. Moderne generatieve AI-modellen hebben bovendien een significant black box-karakter. Dit betekent dat we niet kunnen voorspellen wanneer een model op onwenselijke of onbetrouwbare wijze te werk zal gaan, en ook niet kunnen verifiëren dat de doelen die we een model meegeven correct in het model hun inbedding hebben gekregen. Dit gebrek aan interpreteerbaarheid en uitlegbaarheid wordt problematischer wanneer toekomstige modellen worden gebruikt om zelfstandig acties te nemen of beslissingen te maken. Het kabinet pleit daarom in internationaal verband (bijvoorbeeld via de OESO en de VN) voor veiligheids- en interpretatieregels voor generatieve AI-modellen en stimuleert onderzoek naar verantwoorde en transparante AI-modellen. Het kabinet merkt ook op dat open-sourcemodellen op het gebied van transparantie en uitlegbaarheid uitkomst kunnen bieden en stimuleert open source o.a. door het principe 'open source, tenzij' bij aanbesteding en ontwikkeling te hanteren.[23] Daarbij moet wel worden opgemerkt dat transparantie niet ten koste mag gaan van de veiligheid van generatieve AI-modellen.

Mitigatie van systemische veiligheidsrisico's

Grootschalige omarming van generatieve AI mag niet bijdragen aan maatschappelijke ongelijkheid, instabiliteit en de verstoring van vitale processen. Dit type **systemische veiligheidsrisico's** zou kunnen voortkomen uit de versteviging van bestaande ongelijkheden door de inzet van generatieve AI, snelle en grootschalige veranderingen op de arbeidsmarkt of door verschuivingen in economische en militaire verhoudingen als gevolg van de inzet van geavanceerde generatieve AI. Systemische veiligheidsrisico's hebben een diffuus karakter en kunnen dus niet eenvoudig voorkomen worden door gerichte, specifieke acties. Het kabinet zet daarom in op **onderzoek en monitoring** van de wijze waarop generatieve AI tot bredere, ongewenste maatschappelijke verandering kan leiden. Een proactieve houding is hierbij cruciaal: ongewenste systemische effecten van generatieve AI mogen ons niet zomaar 'overkomen' zoals is gebeurd bij sociale media. Systemische veiligheidsrisico's hebben vaak een internationale component. Het kabinet zet daarom actief in op samenwerking met gelijkgestemde landen, als het gaat om de mitigatie van dergelijke risico's.

2 Uitgangspunt 2: Generatieve AI wordt op een rechtvaardige wijze ontwikkeld en toegepast

Het kabinet streeft ernaar dat de ontwikkeling, gebruik en impact van generatieve AI in overeenstemming zijn met rechtvaardigheidsprincipes.

Rechtmatige ontwikkeling en toepassing

Generatieve AI dient rechtmatig te worden ontwikkeld en gebruikt. Hierin onderscheiden we verschillende elementen die specifiek zijn voor generatieve AI. Ten eerste moet duidelijk zijn wie **verantwoordelijkheid** draagt voor het goed functioneren van AI-modellen en wie onder welke voorwaarden verantwoordelijk is voor eventuele schadelijke of ongewenste uitkomsten. Door het black box-karakter van AI-modellen én de complexiteit van de sociaal-technologische structuur van AI-systemen is er ruimte voor verantwoordelijkheidsverwarring.[24] Wie is er bijvoorbeeld verantwoordelijk voor schadelijke dan wel illegale inzet of output van een bepaald generatieve-AI-model? De modelontwikkelaar, de ontwikkelaar van de toepassing, de organisatie die de tool implementeert of de gebruiker?[25]

20. Binnen de context van LLM's, verwijst jailbreaking naar het ontwerpen van prompts met de intentie model biases uit te buiten om zo output te genereren die niet strookt met het doel van het model. Het model zal bijvoorbeeld antwoord geven op vragen die normaliter door het model niet zouden worden beantwoord.
21. Abdelnabi, S., Greshake, K., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (pp. 79-90). Zie ook bijlage 2 voor meer informatie m.b.t. open source generatieve AI modellen.
22. Hier wordt ook op gewezen door de Cyber Security Raad (2023). Zie ook: cybersecurityraad.nl/actueel/nieuws/2023/12/22/csr-brief-over-ai-en-cybersecurity
23. Voor een overzicht van de inspanningen van het kabinet op het gebied van open-source-software zie Kamerstukken II 2022/2023, 26 643, nr. 1057.
24. Novelli et al. (2023). 'Accountability in Artificial Intelligence: What It Is and How It Works.' AI & Society.
25. Over deze kwestie woedt een levendig juridisch-ethisch academisch debat. Zie bijvoorbeeld: Zech (2021). 'Liability for AI: Public Policy Considerations'. ERA Forum, 22: pp. 147-158 & Hacker (2023). 'The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future'. Computer Law & Security Review 51.

Ten tweede moet bij de ontwikkeling en toepassing van generatieve AI al **bestaande wet- en regelgeving** zoals de **Grondwet, het privacy- en gegevensbeschermingsrecht** en **het auteursrecht** worden gerespecteerd. Zoals eerder uiteengezet (in hoofdstuk 3), kunnen de manieren waarop generatieve AI-modellen data 'oogsten' en verwerken inbreuk maken op deze rechten. Conformiteit met bestaande regelgeving vergt vertaling van bestaande wettelijke kaders naar de context van generatieve AI. Daar waar onduidelijkheden of beleidsleemtes ontstaan, of waar regelgeving niet meer 'fit for purpose' wordt geacht, dient bezinning op aanscherping van bestaande kaders plaats te vinden. Daarnaast blijft van belang om te waarborgen dat toezichthouders de kennis, capaciteit en middelen tot hun beschikking hebben om hun taken nu en in de toekomst effectief uit te kunnen voeren. Het belang van analyse en (waar nodig) aanpassen van regelgevende kaders en capaciteiten van toezichthouders wordt ook herhaaldelijk benadrukt door het Rathenau Instituut.[26]

Kansengelijkheid

Het kabinet onderkent dat generatieve AI-ontwikkeling en -toepassing kansengelijkheid onder druk kan zetten. De mogelijke oorzaak hiervan is tweeledig. Een eerste potentiële factor is **ongelijke toegang** tot generatieve-AI-toepassingen voor inwoners (als gevolg van inkomensverschillen). Een tweede potentiële factor is een **digitale vaardighedenkloof**. Beide factoren kunnen een significante doorwerking hebben in **maatschappelijke en economische kansen**. Het kabinet streeft ernaar te borgen dat iedereen in onze samenleving over de middelen en vaardigheden beschikt te kunnen profiteren van de mogelijkheden die generatieve AI-toepassingen bieden. Dat vergt investeringen in laagdrempelige toegankelijkheid en in technologisch burgerschap.

Non-discriminatie

Zoals beschreven gaan veel AI-systemen gebukt onder **bias** (vooringenomenheid), **selectiviteit** en **stereotypisch** gedachtegoed, verankerd in onderliggende data en modelparameters. Daarmee kan generatieve AI als een vliegwiel fungeren voor discriminatoire dynamieken, bijvoorbeeld als het gaat om aannamebeleid bij organisaties en bedrijven.[27] Met het oog op de beginselen van non-discriminatie acht het kabinet dit onacceptabel. Het kabinet onderschrijft het belang van de ontwikkeling en inzet van methodes om bias en discriminatie te mitigeren. Er zijn verschillende methodes in ontwikkeling, zoals datacuratie, 'constitutional AI' (waarbij een AI-model automatisch getraind wordt om antwoorden te geven die passen bij constitutionele principes.)[28] of 'democratische AI', waarbij een representatieve selectie van mensen wordt betrokken bij de ontwikkeling van een AI-model of –toepassing. Ook stimuleert het kabinet het uitgebreid testen van generatieve AI-modellen om bias en discriminerende uitkomsten te verifiëren. Dit moet periodiek plaatsvinden omdat modellen over tijd kunnen verslechteren.

Transparantie en uitlegbaarheid

Het kabinet onderschrijft het belang van transparantie en uitlegbaarheid van (generatieve) AI-modellen. Het black box-karakter[29] van generatieve AI-modellen staat een basaal begrip van de werking van AI-modellen in de weg. Dit schaadt **procedurele rechtvaardigheid**: zonder inzicht en uitleg kan de eerlijkheid van processen en procedures die door generatieve AI worden gestuurd, niet worden gecheckt. Bovendien zit het ontbreken van transparantie de **correctie van vooroordelen** in de onderliggende data en modelparameters in de weg. Het kabinet zet zich in voor meer transparantie van en inzicht in de trainingsdata en de werking van AI-modellen, afgestemd op de behoeften van de context waarin een model wordt toegepast. De transparantieverplichtingen die onder de AI-verordening voor AI-systemen gaan gelden bieden hierbij uitkomst. Als overheid geven we daarnaast zelf het goede voorbeeld door bij inkoop van generatieve AI-systemen voorwaarden te stellen aan onder andere de herkomst van (trainings)data. Bovendien moet het voor wetenschappers mogelijk zijn om onder de motorkap van AI-modellen te kijken. Gebruikers kunnen gebaat zijn bij begrijpelijke en overzichtelijke model cards – een soort bijsluiters met technische details en mogelijke beperkingen van het betreffende AI-model.[30] Zo'n bijsluiter kan (eind)gebruikers ook helpen bij het bepalen of een model geschikt is, en of er eventuele gevaren bestaan bij de inzet ervan. Dit signaleert ook de AP in haar tweede Rapportage AI- & Algoritmerisico's Nederland.[31] Ook kunnen opensource-modellen in sommige gevallen uitkomst bieden als het gaat om transparantie.[32]

Generatieve AI voor meer gelijkheid

Generatieve AI moet eraan bijdragen dat gelijkheid wordt bevorderd en (sociaaleconomische) kloven worden overbrugd, zowel binnen als tussen landen. Dit is in lijn met verschillende van de Duurzame Ontwikkelingsdoelen (SDG's).[33] Dat vergt ten eerste dat de effecten van automatisering op loonontwikkeling en werkgelegenheid gemonitord worden, met name waar het gaat om **toenemende inkomens- en vermogensongelijkheid**. Door AI kunnen lonen voor verschillende banen meer uit elkaar gaan lopen, net als bij eerdere automatisering het geval was.[34] Ten tweede kunnen beleidsmaatregelen en initiatieven die economische gelijkheid bevorderen – zoals onderwijs- en omscholingsprogramma's, sociale vangnetten en inclusieve AI-ontwikkeling die zorgt voor een meer evenwichtige verdeling van kansen – helpen om economische ongelijkheid tegen

26. Rathenau Instituut (2023), Generatieve AI: hoofdstuk 4.
27. forbes.com/sites/forbeshumanresourcescouncil/2021/10/14/understanding-bias-in-ai-enabled-hiring/?sh=3f9734837b96
28. Voor een voorbeeld van constitutional AI, zie: anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input
29. Black box-modellen: Een (AI-)model waarvan er inzicht ontbreekt in hoe de voorspelling van het model tot stand is gekomen en wat de grondslag voor het gevormde model is.
30. Zie: Model Cards (huggingface.co)
31. Autoriteit Persoonsgegevens (2023). Rapportage AI- & algoritmerisico's Nederland (RAN) – najaar 2023.
32. Zie ook bijlage 2 voor meer toelichting over open source en generatieve AI en bijbehorende uitdagingen.
33. Zoals de bestrijding van alle vormen van armoede (SDG 1) en het verminderen van ongelijkheid (SDG 10).
34. Het omgekeerde kan ook gebeuren: er zijn scenario's denkbaar waarin de gemiddelde werknemer productiever wordt door generatieve AI die op een ondersteunende (i.p.v. verdringende) wijze wordt ingezet. Dit is beschreven in hoofdstuk 3.

te gaan en de digitale kloof te verkleinen. Daarbij is het van belang dat ons sociaal vangnet goed toegerust is op de sociaaleconomische transities die generatieve AI op de (midden) lange termijn naar verwachting in gang zet. Tot slot vereist rechtvaardigheid eerlijke voorwaarden voor de mensen die betrokken zijn bij de ontwikkeling en training van de modellen. Relevant in dit verband is dat de menselijke 'labellers' die helpen AI-modellen te verbeteren eerlijke beloning dan wel arbeidsvoorwaarden verdienen, iets wat nu niet altijd het geval is.[35]

3 Uitgangspunt 3: Generatieve AI die het menselijk welzijn dient en de menselijke autonomie borgt

De inzet en ontwikkeling van generatieve AI moet het menselijk welzijn dienen. Volgens de definitie van de Wereldgezondheidsorganisatie (WHO) is welzijn een positieve staat van lichamelijk, geestelijk en sociaal welbevinden.[36] Welzijn omvat volgens de WHO zowel kwaliteit van leven als het gevoel een betekenisvolle bijdrage te (kunnen) leveren aan de wereld.

Gezondheid

Generatieve AI-toepassingen die bijdragen aan **fysieke en mentale gezondheid** (bijvoorbeeld door het adequaat en efficiënt stellen van diagnoses, het verbeteren van de zorg c.q. andere publieke welzijnsdienstverlening en het bijdragen aan geneeskundig onderzoek) moeten worden aangemoedigd. Tegelijkertijd onderschrijft het kabinet dat het van belang is dat we oog houden voor de **gemeenschappelijkheid** en **menselijkheid** in onze leefwereld, ten behoeve van sociale cohesie en mentaal welzijn. Het kan onwenselijk zijn om bepaalde vormen van menselijk contact te vervangen door AI-gestuurde processen. Automatisering van contact kan mogelijk eenzaamheid en maatschappelijke vervreemding in de hand werken, of afbreuk doen aan onze sociale vermogens.[37] Dit alles behoeft een solide maatschappelijk debat over de gewenste rol van generatieve AI in de samenleving, met oog voor de kansen en de keerzijdes, de mogelijkheden en eventuele grenzen die moeten worden gesteld aan de inzet van generatieve AI. We organiseren dit in brede maatschappelijke coalities.[38]

Persoonlijke en professionele autonomie

Het kabinet acht het ook van belang dat de inzet van generatieve AI menselijke **zelfontplooiing** bevordert en niet ten koste gaat van **persoonlijke autonomie**, zowel in de privésfeer als op de werkvloer. Zoals uiteengezet in hoofdstuk 3 kunnen generatieve AI-toepassingen de autonomie van gebruikers inperken. Persoonlijke autonomie kan in het gedrang komen door zowel directe sturing van gedrag (via misleiding) en via indirecte sturing (via zogenoemde 'dark patterns' of het microtargeten van (des)informatie). Een combinatie van digitale weerbaarheid en een adequate handhaving van het Digital Services Act-verbod op dark patterns en bepaalde vormen van profilering kan uitkomst bieden. Ook op de werkvloer kan **professionele autonomie** in het gedrang komen door de inzet van generatieve AI, bijvoorbeeld door toenemende intransparantie van processen als gevolg van AI-automatisering. Voorkomen moet worden dat verregaande automatisering van werkprocessen de menselijkheid en zingeving die met werk gepaard gaan (te veel) aantast. Daarbij is het nodig dat werkenden grip blijven ervaren op hun werkinhoud en sociale werkrelaties onderhouden, zoals bepleit door de WRR.[39] Het is daarbij essentieel om adequaat te anticiperen op de arbeidsmarkt van de toekomst, en te zorgen dat mensen dan de vaardigheden hebben om binnen de omgang met nieuwe technologieën griphouden op hun werkinhoud. Ook kunnen sociale partners mogelijk van betekenis zijn bij de inbedding van generatieve AI bij bedrijven en organisaties, via medezeggenschapsraden en collectieve arbeidsovereenkomsten.[40]

4 Uitgangspunt 4: Generatieve AI draagt bij aan duurzaamheid en onze welvaart

De Nederlandse overheid zet zich in voor **duurzame** generatieve AI die op een bijdraagt aan onze **welvaart**. Dat betekent dat generatieve AI wordt ingezet om duurzame economische groei te bevorderen, personeelstekorten te verkleinen, en leidt tot innovatieve nieuwe producten en oplossingen, waaronder oplossingen voor maatschappelijke vraagstukken zoals klimaatverandering.

Eerlijk speelveld en productiviteitsgroei

Randvoorwaarde hiervoor is het borgen van **gezonde mededinging** tussen AI-ontwikkelaars, om zowel toegankelijkheid van markt én modellen als competitieve prijzen te bevorderen. Dit vergt effectieve handhaving van de Europese en nationale mededingingsregels, en ook van de Digital Markets Act (DMA). Mededingingsautoriteiten hebben de instrumenten, expertise en capaciteit nodig om snel in te grijpen als dat nodig is om concurrentieverstorend gedrag te voorkomen. Andere randvoorwaarden zijn de aanwezigheid van de juiste kennis van AI binnen organisaties, beschikbaarheid van werkenden met de juiste vaardigheden om AI technologie te kunnen gebruiken, voldoende draagvlak binnen organisaties, en de aanwezigheid van een adequate digitale infrastructuur.

35. ssir.org/articles/entry/ai-workers-mechanical-turk
36. Zie: WHO (2021). Health Promotion Glossary of Terms 2021. URL: Health Promotion Glossary of Terms 2021 (who.int)
37. Turkle, S. (2015). Reclaiming Conversation: The Power of Talk in a Digital Age. New York: Penguin Press. &
38. Een voorbeeld hiervan is de maatschappelijke coalitie 'Over Informatie Gesproken'.
39. Het betere werk. De nieuwe maatschappelijke opdracht | Rapport | WRR
40. Workers could be the ones to regulate AI | Financial Times (ft.com)

Wanneer generatieve AI op verantwoorde wijze wordt gebruikt, biedt het allerlei mogelijkheden om **productiviteit** op de werkvloer te vergroten. Nu al kunnen coding assistents software engineers helpen om code van hogere kwaliteit te schrijven in dezelfde tijd, en taalmodellen helpen bij het schrijven of redigeren van teksten. Wanneer generatieve AI-systemen in de toekomst nog vaardiger worden, zullen dit soort systemen op steeds meer manieren tijdrovende taken kunnen overnemen, waardoor meer tijd over blijft voor kerntaken. De Nederlandse overheid zet zich daarom in voor verantwoorde adoptie van (generatieve) AI-modellen die assisteren bij menselijke taken en die – waar mogelijk en wenselijk – ook taken volledig kunnen automatiseren. We hebben daarbij ook oog voor adoptie van generatieve AI-toepassingen in publieke sectoren zoals de zorg. Niet alleen draagt dit bij aan onze economische groei, de inzet van generatieve AI kan organisaties ook helpen hun personeelstekorten te verminderen. Productiviteitsverhogingen kunnen ook nieuwe werkgelegenheid creëren omdat de vraag naar goederen en diensten toeneemt wanneer inkomens stijgen.

Verantwoorde en innovatieve generatieve AI-toepassingen
De toepassing van generatieve AI-systemen kan niet alleen bestaande taken versnellen, het kan ook bijdragen aan het **innovatief** vermogen van Nederland. We zien een toekomst voor ons waarin Nederlandse bedrijven en organisaties voorop lopen in de toepassing van generatieve AI-modellen voor innovatieve producten en businessmodellen. Generatieve AI-modellen kunnen ook worden ontwikkeld en toegepast om wetenschappelijk onderzoek en R&D te versnellen. Dit stimuleren wij als overheid middels investeringen, publiek-private samenwerking en samenwerking met kennisinstellingen te sturen op verantwoorde generatieve AI. Op deze wijze kunnen we de kansen die generatieve AI biedt voor het oplossen van maatschappelijke vraagstukken verzilveren.

Niet alleen Nederlandse bedrijven, maar ook Nederlandse consumenten moeten de vruchten van generatieve AI kunnen plukken. Daartoe willen we dat consumenten toegang hebben tot een breed aanbod van verantwoord generatieve AI-gedreven producten en diensten. Nederland en de EU pleiten in internationaal verband voor verantwoorde ontwikkeling van generatieve AI-systemen waardoor nuttige generatieve AI-toepassingen die in het buitenland worden bedacht ook direct in Nederland beschikbaar zijn. Europese regelgeving, in het bijzonder de AI-verordening, bevat de randvoorwaarden om (generatieve) AI op een wijze in te zetten die bijdraagt aan veiligheid, gezondheid en fundamentele rechten. Het borgen van de Europese waarden en stimuleren van innovatie kunnen zo hand in hand gaan.

De weg van onderwijs en wetenschap
Generatieve AI kan ook bijdragen aan verdere verbetering van onderwijs, zowel voor kinderen als voor werkenden die willen bijleren of omscholen. Generatieve AI kan bijvoorbeeld worden gebruikt om nieuw lesmateriaal te genereren, of om onderwijsmethodes te personaliseren. Wanneer het Nederlandse onderwijs dit soort kansen benut, kan generatieve AI ook bijdragen aan hogere onderwijskwaliteit en ons toekomstig **verdienvermogen**.

In de wetenschap kan generatieve AI een wezenlijke bijdrage leveren aan het oplossen van complexe problemen, zeker met behulp van specifieke datasets bestaande uit bijvoorbeeld medische beelden en teksten, eiwitstructuren of wiskundige vraagstukken.[41] Hierdoor kan deze technologie een belangrijke rol spelen in het **aanjagen van innovatie**.

Duurzaamheid
Een zwaarwegend punt is dat de ontwikkeling en inzet van generatieve **AI geen onwenselijke impact** mag hebben op **ons klimaat** en ecosystemen. Dat betekent dat we de technologie als overheid niet inzetten als deze grote schade aanricht aan mens en planeet. In lijn met onder meer de Duurzame Ontwikkelingsdoelen streven we naar **duurzame innovatie** en gaan we klimaatverandering tegen.[42] Duurzaamheid moet worden bevorderd door in te zetten op energie-efficiënte trainingsprocessen en implementatie, waarbij de inzet van hernieuwbare energiebronnen als prioriteit wordt beschouwd. Tegelijkertijd kan de toepassing van generatieve AI ingezet worden om bij te dragen aan de mitigatie van klimaatverandering. Zo kunnen generatieve AI-modellen energiegebruik helpen optimaliseren, of wetenschappelijk onderzoek naar schone energiebronnen ondersteunen.

41. Rathenau Instituut (2023), Generatieve AI: p. 14.
42. Zie onder meer SDG 9 (duurzame industrie, innovatie en infrastructuur) en SDG 13 (klimaatverandering tegengaan).

5 Acties

Om tot verantwoorde inbedding van generatieve AI in de Nederlandse samenleving te komen, worden hieronder (concrete) acties, geclusterd in zes actielijnen benoemd.[1] Een aantal daarvan zijn reeds lopende acties, terwijl andere nieuw zijn.[2] Daarnaast worden er ook nog een aantal mogelijk nieuwe acties verkend of onderzocht. De acties zien erop toe dat de samenleving optimaal kan profiteren van de kansen van generatieve AI en de risico's worden beperkt.

Sommige acties grijpen in op de technologie zelf, andere op de bedrijven die de technologie maken en weer andere op de maatschappelijke context waarin deze wordt gebruikt. Sommige acties leveren een bijdrage aan de verwezenlijking van alle hierboven genoemde uitgangspunten, terwijl andere acties bij kunnen dragen aan meerdere, of één enkel, uitgangspunt. Er worden daarbij verschillende mogelijk nieuwe beleidsacties verkend of onderzocht waarvoor nog geen financiële dekking is. De uitkomsten hiervan worden ook voorgelegd aan het volgende kabinet.[3]

Generatieve AI raakt de hele samenleving. Omgaan met de kansen en uitdagingen van generatieve AI is daarmee bij uitstek een opgave die **in gezamenlijkheid** moet plaatsvinden op basis van een **lerende en experimenterende aanpak**. Dat betekent het continu voeren van een **brede maatschappelijke dialoog** in Nederland en mede op basis daarvan zoeken van **internationale samenwerking**, binnen de EU en mondiaal. Dit vereist dat we goed **op de hoogte zijn van de actuele ontwikkelingen** van generatieve AI en ons bewust zijn van de consequenties op sociaaleconomisch en duurzaam vlak. Zo kunnen we sociaaleconomische en verdere digitale transities tijdig zien aankomen en ons bewust zijn van de duurzaamheid van de inzet van generatieve AI. We leggen daarbij de nadruk op verantwoorde toepassing van generatieve AI, zodat de hele samenleving profiteert van het potentieel dat deze technologie te bieden heeft. Als Nederland willen we vanuit deze sterk waardengedreven benadering een mondiale koploper zijn.

Op de hoogte blijven is niet voldoende, we gaan actief inzetten op het **vergroten van kennis en kunde bij overheden, organisaties en burgers**. Dat doen we allereerst door als overheid het goede voorbeeld te geven, bijvoorbeeld door **innovatie** te blijven stimuleren en te investeren in talent binnen een **Nederlands en Europees AI-ecosysteem**. Tegelijkertijd is het van belang dat de **kennis en vaardigheden** over generatieve AI vergroot wordt. Het onderwijs speelt hierin een belangrijke rol, en daarom ondersteunt het kabinet onderwijsinstellingen zodat deze goed in staat zijn om in te spelen op technologische ontwikkelingen. Hierdoor kunnen leerlingen, studenten en docenten over actuele kennis en vaardigheden beschikken. Daarmee hebben we een goede basis om generatieve AI gericht en onder de juiste ethische randvoorwaarden in te zetten en te kunnen beoordelen. Gelet op de maatschappelijke impact daarvan zijn **toekomstbestendige wet- en regelgeving, rechtsbescherming** en, als sluitstuk, sterk en helder **toezicht en handhaving** daarbij essentieel.

1. 1. Samenwerken; 2. Nauwgezet volgen van alle ontwikkelingen; 3. Vormgeven en toepassen van wet- en regelgeving; 4. Vergroten kennis en kunde; 5. Innoveren met generatieve AI; 6. Sterk en helder toezicht houden en handhaven.
2. Deze acties zijn voorzien van financiële dekking.
3. Bij deze (mogelijke) acties staat duidelijk aangegeven dat het om verkennen of onderzoeken gaat.

a Samenwerken

In gesprek en debat
Publiek vertrouwen in nieuwe technologie, en de rol die de overheid speelt om deze technologie op verantwoorde wijze te laten plaatsvinden in de samenleving, is van cruciaal belang voor een goed functionerende digitale samenleving en economie. Dit vraagt niet alleen een duidelijke rol van de overheid, maar ook om het creëren van bewustwording en het voeren van debat bij en met inwoners over de mogelijke impact van generatieve AI én van digitaal burgerschap. Maatschappelijke dialogen en debat (bijvoorbeeld via de Nederlandse AI Parade en de aanpak begeleidingsethiek van ECP[4]) dragen bij aan het nadenken over de rol van generatieve AI in de samenleving en hoe gemeenschappelijkheid en menselijkheid van onze leefwereld behouden kan worden. Dit gesprek is ook van belang bij het vinden van een evenwicht tussen het benutten van het maatschappelijk en economisch potentieel van AI en het omgaan met uitdagingen waarvoor generatieve AI ons stelt.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het uitbreiden van een continue maatschappelijke dialoog met inwoners, werknemers, vakbonden en ondernemers over de impact en rol van generatieve AI op hun levens en de samenleving als geheel. De Nederlandse AI Parade van de NL-AI Coalitie heeft hierin een belangrijke rol, mogelijk ook via (verdere) uitbreiding naar o.a. het onderwijs[5] en de Legal Parade[6].	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	Doorlopend	BZK, OCW, EZK

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het onderzoeken van de mogelijkheid om een specifieke kwartiermaker of organisatie aan te stellen die zich richt op het actief stimuleren en coördineren van diverse (zowel bestaande als nieuwe) initiatieven specifiek gericht op de maatschappelijke dialoog over verantwoorde inzet van generatieve AI.	Alle uitgangspunten	2024	BZK
Dialoogsessies georganiseerd door het Rathenau Instituut in 2024 lenen zich ook bij uitstek voor maatschappelijk debat over de impact van generatieve AI en de rol die het kan, dan wel zou moeten, hebben in onze samenleving en economie.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024	BZK

4. ecp.nl/project/aanpak-begeleidingsethiek/
5. De NL-AIC biedt al een online leerprogramma over de toepassing van AI in het po en vo, zie: onderwijs.ai-cursus.nl/home
6. nlaic.com/bouwstenen/mensgerichte-ai/juridisch/legal-parade/

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het bevorderen van bewustzijn en vaardigheden voor burgers om hun privacy online te beschermen, in het bijzonder de gegevens die van burgers kunnen worden gebruikt bij het trainen van generatieve AI-modellen.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	N.t.b.	JenV
Het bevorderen van participatie in de totstandkoming van generatieve AI-modellen bij en door de overheid, bijvoorbeeld door het stimuleren van initiatieven die modellen trainen op basis van democratische input. Bijvoorbeeld via het (Rijks-)AI-validatieteam (zie ook actielijn 'Innoveren met generatieve AI').	Uitgangspunt 2. Rechtvaardigheid	2025-2028	BZK

Interbestuurlijk samenwerken

Omdat generatieve AI diepgaande implicaties heeft voor de gehele samenleving, is interbestuurlijke en interdepartementale samenwerking cruciaal. De afstand tot burger of ondernemer per bestuurslagen verschilt, het aanpakken van deze kansen en uitdagingen is dan ook een gezamenlijke verantwoordelijkheid voor alle bestuurslagen. Decentrale overheden kunnen generatieve AI-systemen gebruiken om oplossingen te ontwikkelen die aansluiten bij lokale of regionale behoeften. Om consistentie in de te hanteren richtlijnen en normen te kunnen borgen, is interbestuurlijke samenwerking vereist en kunnen decentrale overheden nationaal beleid versterken met inzichten vanuit de lokale praktijk.

Een gecoördineerde aanpak is vereist om als land succesvol voorruit te lopen op deze ontwikkeling. Lokale, regionale en nationale overheden en uitvoeringsorganisaties, maar ook andere stakeholders zoals het bedrijfsleven en maatschappelijke organisaties, moeten gezamenlijk optrekken. Deze samenwerking is niet alleen noodzakelijk voor het ontwikkelen van een coherent en effectief beleid, maar ook voor het voeren van een brede maatschappelijke dialoog. Daarbij is het ook van meerwaarde om te verkennen hoe en in welke mate generatieve AI een rol kan spelen bij het verbeteren van de informatiehuishouding van de overheid. Hetzelfde geldt voor het wegwerken van legacyvraagstukken, waarmee verschillende overheidspartijen te kampen hebben.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Inrichting van interbestuurlijk triageloket voor gegevensdeling, mede ook in het licht van generatieve AI. Het doel van triage is om met een advies gegevensdeling mogelijk te maken en kennis te vergaren over (on)duidelijkheden over de knelpunten en om deze kennis vervolgens ook te kunnen delen.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2023-2024	BZK i.s.m. onder meer SZW, JenV, VWS en FIN en medeoverheden

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het via pilots beproeven van verantwoorde generatieve AI-toepassingen in concrete (proactieve) dienstverlening bij de overheid.	Uitgangspunt 3. Welzijn en autonomie en 4. Duurzaamheid en welvaart	2024-2026	BZK i.s.m. medeoverheden en uitvoering
Het verkennen van de wijze waarop generatieve AI ingezet kan worden bij juridische en administratieve processen ('Legal Tech').	Uitgangspunt 3. Welzijn en autonomie en 4. Duurzaamheid en welvaart	2024 en verder	BZK i.s.m. medeoverheden
Het gebruiken van generatieve AI voor het analyseren van grote datasets voor beleidsvorming en -evaluatie. Hiermee kan de overheid beter inspelen op maatschappelijke behoeften en de effectiviteit van huidige maatregelen toetsen.	Uitgangspunt 4. Duurzaamheid en welvaart	2024 en verder	BZK i.s.m. medeoverheden en uitvoering
We onderzoeken de meerwaarde van verantwoorde generatieve AI bij het bevorderen van transparantie en het verbeteren van onze informatiehuishouding als (Rijks)overheid. Net als de mogelijke rol van generatieve bij het oplossen van legacy vraagstukken.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024 en verder	BZK (i.s.m medeoverheden en uitvoering)

Europese en internationale samenwerking
Generatieve AI is een grensoverschrijdend fenomeen, een (geo)politiek vraagstuk met verstrekkende gevolgen voor de internationale orde. Dit vereist internationale samenwerking met gelijkgestemde landen op aspecten als mensenrechten en persoonlijke en internationale veiligheid. De komende maanden en jaren worden daarom de gesprekken over internationale governance van AI geïntensiveerd.

Gezien de actieve rol die Nederland speelt in de governance en normering van cyber, non-proliferatie en ontwapening, en dankzij een vooraanstaande Research & Development positie, beschikt Nederland over waardevolle inzichten en netwerken om een actieve bijdrage te leveren aan de internationale governance van AI. Nederland zal hier vanuit een sterk waardengedreven benadering in participeren.

De VS en China zijn wereldspelers als het gaat om AI-capaciteiten. Ook de EU scoort goed, behalve als het gaat om een (AI-)ecosysteem waarin bedrijven (generatieve) AI productief kunnen maken. Nederland heeft een sterke wetenschappelijke basis voor onderzoek op het gebied van AI en behoort binnen de EU tot een categorie van landen die door hun specialismen ook relevante spelers op het wereldtoneel kunnen zijn.7 Op het gebied van generatieve AI is de VS de belangrijkste speler, met afstand gevolgd door China. De veelzijdigheid van generatieve AI stelt ons voor strategische keuzes op het gebied van digitale open strategische autonomie, zoals hoe we het politiek-economisch fundament kunnen versterken, risicovolle strategische afhankelijkheden kunnen mitigeren, en het geopolitieke handelingsvermogen van de EU kunnen vergroten. Hiervoor heeft het kabinet in oktober 2023 de Agenda Digitale Open Strategische Autonomie (DOSA) gepubliceerd, waarbij AI één van de beleidsprioriteiten is.8

Schaal is essentieel bij generatieve AI: een half systeem levert veel minder dan de halve waarde. Dit maakt dat in een competitieve markt waarin de beste systemen een toename van gebruikers en investeringen zien, de schaal (data, rekenkracht en investeringen) bepalend is voor succes. Europa biedt een kans om deze schaalvergroting te bereiken. Ook in EU-verband lopen initiatieven om de kennis- en innovatiepositie op generatieve AI te versterken, zoals de Alliance for Languages Technologies EDIC (ALT-EDIC).[9] Dit is een door Frankrijk gecoördineerd initiatief bestaande uit verschillende lidstaten, momenteel verkent het kabinet een deelname aan dit traject. Naast samenwerking zijn de doelen van deze EDIC om taalkundige en culturele diversiteit in

7. Zie het WRR-rapport 'Opgave AI' (2021) 'De nieuwe systeemtechnologie', voor een uitgebreide toelichting.
8. open.overheid.nl/documenten/5cb9749c-7efa-40db-9328-5da7fa5fcb7c/file
9. Een EDIC is een juridisch kader in de EU dat de lidstaten helpt bij het opzetten en uitvoeren van meerlandenprojecten. ALT-EDIC is een van de EDIC's die in voorbereiding zijn. Zie ook: ALT-EDIC (europa.eu)

Europa te behouden, technologisch leiderschap en strategische autonomie te versterken, Europese normen en waarden te respecteren en bewustwording te creëren.

Ook in andere EU-netwerken staat de discussie rondom generatieve AI niet stil. Zo heeft de AI Data Robotics Association (ADRA) een taskforce generatieve AI in het leven geroepen om kansen voor Europa in generatieve AI te helpen identificeren en de Europese Commissie te adviseren in besluiten over investeringen vanuit Horizon Europe, Digital Europe en andersoortige i nstrumenten. Het AI-Alliance Forum, een AI-platform dat zich richt op ondernemers en beleidsmakers om gezamenlijk de weg voorwaarts te bepalen voor Europese AI-innovatie, heeft aan de Commissie een oproep gedaan om zich in te zetten op multimodale AI.[10]

De toenemende behoefte aan rekenkracht en AI-chips die de volgende generatie krachtige AI-modellen met zich mee zal brengen, vraagt om significante investeringen in AI-infrastructuur. Dit is een randvoorwaarde om te voldoen aan de ambitie van Nederland om voorloper te zijn op het gebied van (verantwoorde) generatieve AI en om op dit vlak concurrerend te kunnen zijn.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Nederland neemt deel aan het partnerschap EuroHPC onder Horizon Europe op gebied van high performance computing (HPC), Nederlandse bedrijven en kennisinstellingen kunnen zo deelnemen aan Europese projecten op gebied van HPC en quantum computing.	Uitgangspunt 4. Duurzaamheid en welvaart	2023	EZK
Nederland zet zich in voor het versterken van de internationale rechtsorde. Het kabinet zet zich ervoor in dat waardengedreven veilige AI inzet van AI overal in de wereld de norm wordt.Deel hiervan is het bijdrage aan de internationale organisaties die richtlijnen voor AI proberen op te stellen, zoals binnen de VN gebeurt. We pleiten voor pre-deployment audits van de meest geavanceerde modellen en het opstellen van regels voor het beschikbaar stellen van de model weights[11] van grote modellen. Dit biedt ook de mogelijkheid om samen problemen aan te pakken waarover Nederland niet unilateraal regels voor kan opstellen, zoals het ontwikkelen van onveilige AI buiten Nederland, of het aanpakken van de arbeidsomstandigheden waaronder het finetunen van modellen nu vaak plaatsvindt.	Uitgangspunt 1. Veiligheid en 3. Welzijn en autonomie	Doorlopend	BZ
Als het gaat om het militaire gebruik van (generatieve) AI, streeft Nederland naar internationaal breed gedragen of mondiale normstelling en bouwt daarbij voort op de gezette stappen tijdens de REAIM Summit in Den Haag.[12]	Uitgangspunt 1. Veiligheid en 3. Welzijn en autonomie	Doorlopend	BZ i.s.m. Defensie

10. Multimodale AI is een AI-gebied waarbij informatie uit meerdere sensorische modaliteiten, zoals afbeeldingen, tekst, audio en video, wordt verwerkt en geïnterpreteerd. Zie verder: The AI4Media Strategic Research Agenda on AI for the Media Industry | Futurium (europa.eu)
11. Model weights zijn parameters die een LLM heeft geleerd tijdens het trainingsproces. Deze parameters zijn interne aanpassingen die het model maakt om zich aan te passen aan de inputdata en een taak uit te voeren, zoals het maken van voorspellingen.
12. Zo komt er o.a. een 'global commission AI' om wereldwijd het wederzijds bewustzijn te bevorderen, te verduidelijken wat te verstaan onder AI in het militaire domein en te bepalen hoe te komen tot de verantwoorde ontwikkeling, productie en toepassing hiervan. Zie ook: rijksoverheid.nl/ministeries/ministerie-van-buitenlandse-zaken/nieuws/2023/02/16/call-to-action-verantwoord-gebruik-ai-in-het-militaire-domein.

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het kabinet implementeert de Agenda DOSA, waarin AI één van de geïdentificeerde thema's is waar een strategische afhankelijkheid bestaat.[13]	Alle uitgangspunten	2023-2024	EZK (i.s.m. andere ministeries)
Op Europees niveau is in het kader van de Europese Economische Veiligheidsstrategie een start gemaakt met een nieuwe tranche risicoanalyses naar kritieke technologieën, waaronder AI. Deze zien toe op technologielekkage (weglekken van kennis) en technologieveiligheid (risico's voor de innovatie-capaciteit). Vanuit EZK is hierover op nationaal niveau een vragenlijst ingevuld. De Commissie verwacht in Q1 de eerste uitkomsten gereed te zullen te hebben.[14]	Uitgangspunt 1. Veiligheid	Doorlopend	EZK

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
EZK verkent op dit moment de mogelijkheid voor deelname aan de Alliance for Languages Technologies European Digital Infrastructure Consortium (ALT-EDIC). De ALT-EDIC brengt bestaande open source LLM's ten behoeve van gebruik door industriële spelers en mkb, met aandacht voor het mitigeren van bias.[15] Daarnaast fungeert de ALT-EDIC als een fonds voor het stimuleren van nieuwe LLM's en foundation models.	Uitgangspunt 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	2023-2024	EZK

13. open.overheid.nl/documenten/5cb9749c-7efa-40db-9328-5da7fa5fcb7c/file (zie ook p. 26/29).
14. Kamerstuk 22112-3826, Nr. 3826.
15. Open source-LLM's zijn transparant en komen daarmee ten goede aan de inzichtelijkheid van de betreffende modellen. Ook kan deze ontwikkelvorm bijdragen aan ethischere datacuratie ter voorkoming van bias en privacy- en auteursrechtschendingen.

b Nauwgezet volgen van alle ontwikkelingen

Het is van groot belang om de ontwikkelingen in generatieve AI in het algemeen nauwgezet te volgen. Daarnaast zal de aandacht worden gericht op specifieke onderwerpen: de gevolgen voor de werkgelegenheid, democratie, duurzaamheid en klimaat. Het kabinet zal daarbij de open aanpak die bij het opstellen van deze visie gehanteerd is voortzetten.[16] Dit betekent dat beleid in nauw contact blijft met medeoverheden, uitvoeringsorganisaties, kennisinstituten, commerciële partijen, publieke belangenorganisaties, werkgevers, werknemers en burgers.

Zicht houden op ontwikkelingen rondom (generatieve) AI

De maatschappelijke impact van generatieve AI zal de komende jaren steeds zichtbaarder worden en het is daarom van belang om een systematiek neer te zetten waarmee de (lange termijn) implicaties goed kunnen worden gemonitord. Dit geldt zowel voor technologische ontwikkelingen als voor de effecten van de ontwikkeling en het gebruik van generatieve AI in uiteenlopende sectoren en terreinen van onze samenleving en economie.[17] De WRR wijst in dit kader ook op het belang van het betrekken van het maatschappelijk middenveld bij het monitoren en het zicht houden op gevolgen van de verdere inbedding van (generatieve) AI in de Nederlandse samenleving.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het stimuleren van de uitnodiging van kennis uit het buitenland (ook buiten de EU) om bij te dragen aan onderzoeksprogramma's die betrekking hebben op (generatieve) AI.	Uitgangspunt 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	Doorlopend	OCW
De AP heeft in januari 2023 de nieuwe Directie Coördinatie Algoritmes (DCA) opgericht om coördinerend toezicht op algoritmes te versterken. Onderdeel hiervan is het proactief signaleren en analyseren van sector-overstijgende en overkoepelende risico's en effecten van de ontwikkeling en inzet van algoritmes, waaronder ook generatieve AI, en het verzamelen en delen van kennis daarover.	Uitgangspunt 1. Veiligheid, 2. Rechtvaardigheid en 3. Welzijn en autonomie	2023-2027	AP
In overleg met het bedrijfsleven signaleren waar innovatiebelemmeringen liggen, waarbij de overheid mogelijk een bijdrage kan leveren. o.a. via de Topsector ICT binnen het missiegedreven innovatiebeleid 2024-2027.	Uitgangspunt 4. Duurzaamheid en welvaart	Doorlopend (in 2024 eerste update voortgang)	EZK

16. Zie bijlage 1 voor meer informatie over de open aanpak die is gehanteerd om tot deze visie te komen.
17. Hier wordt ook op gewezen door het Rathenau Instituut in haar scan over generatieve AI (2023). https://www.rathenau.nl/sites/default/files/2023-12/Scan_Generatieve_AI_Rathenau_Instituut.pdf

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het kabinet verkent de mogelijkheden van een AI-adviesorgaan op het hoogste niveau (of Rapid Response Team AI (RRT-AI)).[18] Deze werkvorm/expertgroep kan het kabinet adviseren over belangrijke ontwikkelingen op het gebied van (generatieve) AI.	Alle uitgangspunten	2023-2024	BZK, EZK en OCW
Er komt een overheidsbrede inventarisatie en monitor van de initiatieven, ontwikkelingen en gebruik op het gebied van generatieve AI door overheden en (semi-)publieke organisaties, waarvan we de indicatoren interbestuurlijk opstellen. Hierbij is aandacht voor zowel kansen als risico's. De resultaten zullen ook periodiek worden gedeeld met uw Kamer.[19]	Alle uitgangspunten	2023-2025	BZK (i.s.m. medeoverheden en uitvoering)
Het is belangrijk om als (Rijks)overheid niet alleen bewust te zijn van de huidige capaciteiten van AI, maar ook om klaar te zijn voor toekomstige ontwikkelingen en capaciteiten van AI-systemen, daarom houden we vinger aan de pols met betrekking tot de toekomstige capaciteiten van (generatieve) AI. Hierbij kijken we goed waar we gebruik kunnen maken van bestaand (internationaal) onderzoek en waar kunnen aanvullen op de huidige state-of-the art, zoals de onderzoeken van Epoch of Metaculus,[20] organisaties en methodes met een goed trackrecord van het voorspellen van technologische doorbraken en ontwikkelingen.	Alle uitgangspunten	Doorlopend	BZK en EZK

18. Zie in dat kader ook de in oktober 2023 aangenomen motie van de leden Dekker-Abdulaziz en Rajkowski: tweedekamer.nl/kamerstukken/moties/detail?id=2023Z17682&did=2023D42892
19. Om hier toe te komen worden sessies georganiseerd met o.a. medeoverheden.
20. Zie ook: Exploring Metaculus's AI Track Record

Sturen op sociaaleconomische transities op het gebied van werk en inkomen
Zoals al aangeduid in hoofdstuk 3, kan wijdverbreid gebruik van generatieve AI-toepassingen in professionele en industriële context leiden tot arbeidsmarkt- en inkomensverschuivingen en gevolgen hebben voor de kwaliteit van werk. Het afstemmen van de vaardigheden van de beroepsbevolking op de arbeidsmarkt van de toekomst is essentieel. Ook nopen de transities tot het nadenken over de financiële houdbaarheid én adequaatheid van ons huidige socialezekerheidsstelsel.

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het kabinet heeft de Sociaal-Economische Raad (SER) gevraagd om de impact van AI (waaronder generatieve AI) op de arbeidsproductiviteit, kwantiteit en kwaliteit van werk in kaart te brengen.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024-2026	BZK, EZK, OCW en SZW

Duurzame ontwikkeling en gebruik (generatieve) AI
De uitdagingen op het gebied van klimaat zijn omvangrijk. Generatieve AI heeft de potentie om hier een positieve bijdrage aan te leveren. Momenteel bestaan er nog geen gestandaardiseerde methoden om de klimaatimpact van (generatieve) AI te meten. Vooralsnog zorgt de ontwikkeling en inzet van generatieve AI voor een grotere ecologische voetafdruk (zie ook hoofdstuk 3). Als overheid is duurzaamheid een belangrijk uitgangspunt bij de inzet van generatieve AI, dat betekent dat we de technologie niet toepassen als deze te schadelijk blijkt voor mens en planeet. Het is daarom ook noodzakelijk om in te zetten op ontwikkelingen die bijdragen aan het verduurzamen van generatieve AI. Hierbij kan worden gedacht aan energie-efficiëntere trainingsprocessen en het inzetten op publiek beschikbare en verantwoorde LLM's om zo het (pre)trainingsproces minder vaak te laten plaatsvinden. Bij duurzame generatieve AI hoort ook het verder verkennen van concrete mogelijkheden hoe generatieve AI kan bijdragen aan klimaatopgaven, via bijvoorbeeld innovatie en onderzoek.

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het duurzaamheidsaspect bij de ontwikkeling en het gebruik van generatieve AI (door de overheid) nader onderzoeken en waar mogelijk maatregelen nemen om de negatieve gevolgen te reduceren.	Uitgangspunt 3. Welzijn en autonomie en 4. Duurzaamheid en welvaart	2024	BZK
Verder onderzoek stimuleren naar de wijze waarop generatieve AI positief kan bijdragen aan verschillende klimaatopgaven bij de overheid.	Uitgangspunt 3. Welzijn en autonomie en 4. Duurzaamheid en welvaart	2024-2025	BZK

Democratie en (generatieve) AI
Generatieve AI brengt verschillende risico's met zich mee voor de democratie en onze democratische rechtsstaat. Met name de versnelde productie en verspreiding van desinformatie of strafbare content, zoals bedreigingen en hate speech, leiden tot zorgen. Het is daarom essentieel blijvend te monitoren hoe generatieve AI de dynamiek in de democratie verandert en welke aanpassingen dat vraagt in bestaand beleid om de democratie te vernieuwen en te beschermen. Het kabinet heeft daarom eerder al aangegeven dat onderzocht moet worden of de beleidsinzet op desinformatie met het oog op deze nieuwe technologieën moet worden herzien.[21] Tegelijkertijd moeten we ook de kansen van generatieve AI benutten om de democratie te versterken. Een voorbeeld daarvan is het stimuleren van onderzoek naar taalmodellen voor Fries en Papiaments en taalmodellen om communicatie met en voor burgers begrijpelijker en inclusiever te maken. Dit draagt bij aan een efficiëntere dienstverlening en een meer toegankelijke en inclusievere overheid.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
In de voortgangsbrief Rijksbrede strategie voor de effectieve aanpak van desinformatie wordt ingegaan op de vraag of de beleidsinzet op desinformatie moet worden aangepast met het oog op nieuwe technologieën.	Uitgangspunt 1. Veiligheid, 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024	BZK

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Verder onderzoek stimuleren naar de wijze waarop generatieve AI positief kan bijdragen aan de democratie, bijvoorbeeld op het gebied van burgerparticipatie.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024-2025	BZK
Het stimuleren van het ontwikkelen en verbeteren van (open en publieke) taalmodellen die getraind zijn op talen zoals bijvoorbeeld Fries, Papiaments en gebarentaal.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024-2025	BZK
Het verkennen van de wijze waarop generatieve AI kan ondersteunen om de communicatie met burgers te verhelderen en inclusiever te maken.	Uitgangspunt 3. Welzijn en autonomie en 4. Duurzaamheid en welvaart	Doorlopend	BZK i.s.m. medeoverheden

21. Kamerbrief II, 2023-2024, 35 165, nr. 64.

c Vormgeven en toepassen wet- en regelgeving

Wet- en regelgeving is van cruciaal belang om het vertrouwen in generatieve AI te bevorderen. Gelukkig valt generatieve AI niet in een juridisch vacuüm. Zoals in hoofdstuk 4 al aangegeven moet generatieve AI voldoen aan juridische kaders. Om normen toekomstbestendig(er) te maken, worden ze vaak breed en open geformuleerd. Daardoor kan onzekerheid ontstaan over hoe wet- en regelgeving moeten worden uitgelegd. Hierin is niet alleen een rol voor toezichthouders en rechters weggelegd, maar ook voor de Rijksoverheid, in samenwerking met medeoverheden. Bij een zich snel ontwikkelende technologie, zoals generatieve AI, is het daarbij belangrijk dat regulering zoveel als mogelijk technologieneutraal en adaptief is.

Er worden duidelijke regels opgesteld voor ontwikkelaars en aanbieders van generatieve AI, om zo (maatschappelijke) risico's en uitdagingen die gepaard gaan met de verdere groei van generatieve AI te mitigeren of te adresseren. In de (digitale) samenleving verdienen burgers en bedrijven rechtszekerheid en moeten zij erop kunnen vertrouwen dat de overheid passende kaders en regels opstelt en hier helder toezicht op inricht. Hierbij dient zowel aandacht te zijn voor de wijze waarop bestaande wettelijke instrumenten, alsook aankomende wet- en regelgeving, in staat zijn om een speelveld te creëren waarin generatieve AI op een veilige, rechtvaardige, rechtmatige én transparante wijze kan worden ontwikkeld en gebruikt. Duidelijke nationale, Europese en of internationale kaders dragen bij aan het versnellen van innovatie en nieuwe oplossingen. De Europese AI-verordening is hiervoor de basis. Daarom werken we in 2024 hard aan de overheidsbrede implementatie van de AI-verordening, partijen moeten hierop goed voorbereid zijn door onder andere goede voorlichting en guidance richting betrokkenen.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Implementatie van de Europese AI-verordening, waaronder • Aanwijzen toezichthouders via uitvoeringswet • Consultatie en parlementaire behandeling uitvoeringswet • Inrichten Europees toezicht • Voorlichting aan (mede-)overheden, bedrijfsleven en andere belanghebbenden • Stimuleren van guidance en normuitleg o.a. via toezichthouders • Inrichten Nederlandse regulatory sandbox • Ontwikkelen Europese standaarden • Vaststellen Europese uitvoeringshandelingen	Alle uitgangspunten	2024-2027	EZK en BZK i.s.m. JenV, OCW, SZW, I&W, VWS, LNV, BZ, Financiën, en Defensie
Het kabinet participeert actief in de onderhandelingen over het AI-verdrag in CAI (Comité voor AI) van de Raad van Europa.	Uitgangspunt 1. Veiligheid en 2. Rechtvaardigheid	2023-2024	BZK (i.s.m. JenV)

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Blijvend monitoren (in nationaal en Europees verband) of het huidig wettelijk kader (zoals auteursrecht en AVG) volstaat. Door middel van nader onderzoek en toezichtssignalen.	Alle uitgangspunten	Doorlopend(via de werkagenda)	Kabinetsbreed

d Vergroten kennis en kunde

Het kabinet zet proactief in op een toename van kennis en vaardigheden. Dit stelt ons in staat de kansen die generatieve AI biedt ten volle te benutten. Ook komt adequate inzet voor kennis en kunde ten goede aan grip op de technologie, aan maatschappelijke kansengelijkheid en aan participatie. Allereerst zullen we als overheid het juiste voorbeeld geven. Tegelijkertijd ziet het kabinet het belang van het versterken van kennis en vaardigheden op het gebied van generatieve AI in de hele maatschappij. Dit doen we door het onderwijs te ondersteunen zodat zij adequaat kunnen inspelen op technologische ontwikkelingen. Daarnaast zetten we binnen de overheid volop in op de ontwikkeling van digitale kennis en vaardigheden die nodig zijn om verantwoord om te kunnen gaan met generatieve AI.[22]

De overheid geeft het goede voorbeeld

De Nederlandse overheid heeft een voorbeeldfunctie als het gaat om het verantwoord en veilig ontwikkelen, inkopen en inzetten van generatieve AI. Dit vraagt ambtenaren die over de juiste gereedschapskist beschikken om deze technologie – in verschillende facetten van hun werk – verantwoord in te kopen, te ontwikkelen of in te zetten. Daarom is het voor de overheid van groot belang dat medewerkers over de juiste kennis en kunde beschikken (opleiden) en over de juiste kaders beschikken. Hiermee krijgen medewerkers handvatten om de technologie binnen het eigen werk verantwoord en rechtmatig in te zetten, de juiste voorwaarden, kaders en beleid om de technologie in te kopen en bewustwording van de uitdagingen en kansen van generatieve AI.

De vormgeving van de publieke dienstverlening kan het vertrouwen van de inwoner in de overheid in de hand werken óf juist schaden. De voorbeeldfunctie van de Nederlandse overheid moet zich dus op een integrale manier manifesteren.

Als overheid stimuleren wij innovatie en zien wij het belang in van experimenten om generatieve AI in te zetten voor publieke waarden. Hierbij dienen alle generatieve AI-toepassingen te voldoen aan geldende wet- en regelgeving.[23] Om vast te stellen welke specifieke vorm van inzet van generatieve AI wel of niet mogelijk is, dient voorafgaand aan het gebruik ervan per unieke casus een risicoanalyse te worden uitgevoerd. Dit zijn een Data Protection Impact Assessment (DPIA) en een algoritme impact assessment (zoals een Impact Assessment Mensenrechten en Algoritmes (IAMA)), waarin de risico's en risicobeperkende maatregelen worden vastgesteld. De uitkomsten hiervan dienen voorafgaand aan de inzet van de toepassing ter advies aan de (departementale) Chief Information Officer en de Functionaris Gegevensbescherming te worden voorgelegd. De bovengenoemde punten zijn van toepassing bij het gebruiken of (door)ontwikkelen van een open source generatieve AI-toepassing. In het kader van de Wet open overheid (Woo) en het stimuleren van transparantie, geldt de beleidslijn 'open (source), tenzij'.[24] Niet-gecontracteerde generatieve AI-toepassingen[25] voldoen over het algemeen niet aantoonbaar aan de geldende privacy- en auteursrechtelijke wetgeving. Zodoende is het gebruik hiervan door Rijksorganisaties (of in opdracht daarvan) niet toegestaan, in die gevallen waarin het risico bestaat dat wetgeving wordt overtreden, tenzij de aanbieder en de gebruiker aantoonbaar voldoen aan de geldende wet- en regelgeving. Gecontracteerde generatieve AI-toepassingen dienen bovendien te voldoen aan de Algemene Rijksvoorwaarden bij IT-overeenkomsten 2022 en aan departementale inkoopvoorwaarden (indien deze prevaleren). Bij het gebruik van een generatieve AI-toepassing is het van belang dat medewerkers voldoende worden geïnformeerd over hoe zij deze technologie op een verantwoorde wijze kunnen inzetten. Dit kan door training of richtlijnen voor verantwoord gebruik.

Met de opkomst van generatieve AI in de samenleving zijn de vooruitzichten voor individuen en organisaties veelbelovend. Aan de hand van ethisch gesprek, nadere risicoanalyses en risicocategorisering in lijn met de toekomstige AI-verordening worden de (on)mogelijkheden van de inzet van generatieve AI door Rijksorganisaties in komende jaren bepaald. Dit standpunt is geen categorisch verbod van de technologie, maar herhaalt geldende wet- en regelgeving. Het gebruik wordt niet ontzegd, maar gekaderd. Zo blijft het mogelijk te experimenteren met de technologie en wordt dit gestimuleerd.

22. Zie in dat kader ook de I-strategie Rijk (2021-2025) en de aandacht voor I-vakmanschap: I-strategie Rijk I-strategie Rijk 2021-2025 - Digitale Overheid.
23. Zie: Kamerbrief over voorlopig standpunt voor Rijksorganisaties bij het gebruik van generatieve AI | Kamerstuk | Rijksoverheid.nl
24. Het kabinet merkt ook op dat open-sourcemodellen op het gebied van transparantie en uitlegbaarheid uitkomst kunnen bieden en stimuleert open source o.a. door het principe 'open source, tenzij' bij aanbesteding en ontwikkeling te hanteren. Daarbij moet wel worden opgemerkt dat transparantie niet ten koste mag gaan van de veiligheid van generatieve AI-modellen (zie ook uitgangspunt 1).
25. Bijvoorbeeld openbaar toegankelijke, door grote techbedrijven ontwikkelde (en vaak online op het internet) aangeboden vormen van generatieve AI.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Er is een handreiking in ontwikkeling om (Rijks)overheidsorganisaties richting te bieden bij de inzet van generatieve AI. Deze handreiking voorziet in de (on)mogelijkheden voor gebruik van generatieve AI voor overheidsmedewerkers. We vertalen deze handreiking gezamenlijk naar de context van decentrale overheden.[26]	Uitgangspunt 1. Veiligheid en 2. Rechtvaardigheid	2023-2024	BZK (i.s.m. medeoverheden)
De AIVD analyseert en stelt publicaties op voor overheidsfunctionarissen alsmede voor de maatschappij over risico's en mogelijkheden voor vergroten van de weerbaarheid (m.b.t. onder meer generatieve AI.	Uitgangspunt 1. Veiligheid	Doorlopend	AIVD
Het NCSC en AIVD zetten in op het vergroten van de technische kennis op het onderwerp AI en weerbaarheid tegen cyberdreigingen in dit nieuwe domein en weerbaarheid tegen ongewenst gebruik van LLM's.[27] Daarbij worden organisaties geïnformeerd over urgente ontwikkelingen en krijgen zij concrete handvatten aangeboden voor de veilige ontwikkeling van AI.[28]	Uitgangspunt 1. Veiligheid en 4. Duurzaamheid en welvaart	2024	NCSC en AIVD

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Ambtenaren die bezig zijn met (generatieve) AI leren op dit thema moreel ethische kwesties te herkennen en te onderzoeken. Daartoe kan actief de samenwerking worden gezocht met het programma Dialoog & Ethiek.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024-2025	BZK
Via RADIO (Rijksacademie voor Digitalisering en Informatisering Overheid) aan kennisdeling doen over de (on)mogelijkheden van veilig gebruik van generatieve AI.	Uitgangspunt 1. Veiligheid en 2. Rechtvaardigheid	2024	BZK
Investeren in de kennis en vaardigheden van ambtelijke professionals en volksvertegenwoordigers in alle bestuurslagen door middel van cursussen en workshops. Interbestuurlijke kennisdeling faciliteren over de mogelijkheden voor veilig gebruik van generatieve AI door het delen van kennis en praktijkervaring.	Alle uitgangspunten	Doorlopend	BZK (i.s.m. medeoverheden)

26. Voorafgaand aan de Handreiking wordt indien vereist een risicoanalyse uitgevoerd op specifieke use case(s), via een DPIA en een IAMA.
27. AI: Cruciaal moment in de geschiedenis of een hype? | Expertblogs | Nationaal Cyber Security Centrum (ncsc.nl)
28. AI-systemen: ontwikkel ze veilig | Publicatie | AIVD

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het opstellen en ontwikkelen of aanscherpen van (interbestuurlijke) inkoopvoorwaarden met het oog op generatieve AI.	Alle uitgangspunten	2024-2025	BZK (i.s.m. medeoverheden)

Investeren in talent en rekenkracht
Zoals al eerder aangegeven, is generatieve AI een grensoverstijgend fenomeen. Vanuit de Nederlandse overheid is het belangrijk om, vooral in Europees verband, een ecosysteem voor (generatieve) AI te stimuleren door publiek-privaat samen te werken en investeringen daarin, evenals het investeren in (open) publieke alternatieve generatieve AI (zie ook actielijn 'Innoveren met generatieve AI'). We verkennen daarom de (noodzakelijke) investeringen in grootschalige wetenschappelijke en technologische infrastructuur (o.a. supercomputers en rekenkracht) op nationaal en EU-niveau om competitief te zijn op het terrein van LLM's en andere vormen van generatieve AI. Een expliciet aandachtpunt daarbij is ook het ontwikkelen en behouden van AI-talent zodat we generatieve AI kunnen ontwikkelen die voldoet aan Europese normen en waarden. Dit heeft ook meerwaarde voor de digitale open strategische autonomie van Europa.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Investeren in voldoende computerinfrastructuur om aantrekkelijke (wetenschaps)projecten te kunnen uitvoeren in Nederland en de EU.	Uitgangspunt 4. Duurzaamheid en welvaart	Doorlopend	Kabinetsbreed
Nederland blijft investeren in innovatieve projecten en onderzoek op het gebied van veilige en verantwoorde AI, bijvoorbeeld om interpreteerbaarheid en transparantie van AI-modellen te vergroten.	Uitgangspunt 1. Veiligheid en 4. Duurzaamheid en welvaart	Doorlopend	Kabinetsbreed

Kennis en vaardigheden vergroten (in het onderwijs)
De inzet van generatieve AI vereist verschillende vaardigheden. Zowel in het gebruik van generatieve AI-tools, alsook bij het beoordelen van de content die wordt genereerd door deze technologie. Dit vraagt om verdere inzet op mediawijsheid voor verschillende doelgroepen, waarbij specifiek aandacht dient te worden besteed aan bewustwording en het beoordelen van de betrouwbaarheid van (gegenereerde) content. Met de opkomst van generatieve AI worden in het gehele onderwijs vaardigheden als kritisch en gestructureerd denken belangrijker. Generatieve AI kan snel grote hoeveelheden tekst, beeldmateriaal, audio en (computer) code genereren, waardoor het kritisch kunnen beoordelen en waarderen van deze content des te belangrijker wordt.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het Nationaal Onderwijslab AI (NOLAI) doet onderzoek naar de pedagogische, maatschappelijke en sociale consequenties van generatieve AI.[29]	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2023-2024	OCW en EZK
Het stimuleren van de verantwoorde toepassing van (generatieve) AI-toepassingen voor maatschappelijke uitdagingen bij hbo-kennisinstellingen via 'AI in Actie'.[30]	Uitgangspunt 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	N.t.b.	BZK en OCW i.s.m. hbo's
Het kabinet zet zich in voor een vaste plek voor digitale geletterdheid in het landelijk curriculum voor primair en voortgezet onderwijs.	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	2023 en verder	OCW en BZK

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het versterken van digitale vaardigheden[31] en digitaal bewustzijn van mensen in Nederland zodat zij bewust, kritisch en actief kunnen omgaan met AI-toepassingen.[32]	Uitgangspunt 2. Rechtvaardigheid en 3. Welzijn en autonomie	Doorlopend	BZK en OCW

29. Zie ook: rijksoverheid.nl/documenten/kamerstukken/2023/07/06/visiebrief-digitalisering-in-het-funderend-onderwijs
30. Zie ook: Agenda 'Coalities voor de digitale samenleving': open.overheid.nl/documenten/10c88500-cdb5-4815-bd00-c915a5242ea3/file
31. rijksoverheid.nl/documenten/rapporten/2023/06/19/digitale-vaardigheden-van-nederlanders
32. Zie ook lijn 1.1 van de Werkagenda Waardengedreven Digitaliseren: Vergroten van digivaardigheden en kennis.

e Innoveren met generatieve AI

Bij een overheid die aan het stuur wil staan, hoort een overheid die experimenteert met veilige en verantwoorde generatieve AI. Op deze manier kunnen afhankelijkheden worden verminderd en ontdekken worden waar in concrete toepassingen risico's en kansen liggen. Dit heeft als doel de economische, wetenschappelijke en andere maatschappelijke kansen die generatieve AI ons biedt op een verantwoorde manier ten volle te kunnen benutten. Daarvoor is het ook belangrijk dat publieke organisaties samen met het Nederlandse bedrijfsleven de kennis en innovatiebasis op de ontwikkeling en toepassing van generatieve AI versterkt.

Gelet op de impact die de geconcentreerde ontwikkeling van (krachtige) generatieve AI met zich meebrengt, is het belangrijk dat er in Nederland een klimaat wordt gestimuleerd waarin er breed de ruimte is voor het experimenteren, testen en opschalen van betrouwbare en transparante (generatieve) AI-modellen en tools (rondom bijvoorbeeld validatie of bias-detectie). Daarbij neemt ook het belang van hoogkwalitatieve (Nederlandstalige) datasets als belangrijke bouwsteen voor generatieve AI-modellen toe. Om verdere kennis en ervaring op te doen met de validatie van AI[33], heeft het kabinet een (Rijks)-AI-validatieteam opgericht. Dit team buigt zich onder meer over het meetbaar maken van risico's en kansen van generatieve AI. Het team bestaat uit software engineers die samen met beleidsmakers gaan werken aan concrete hulpmiddelen om (generatieve) AI te valideren.

Een mooi voorbeeld van verantwoorde innovatie met generatieve AI in Nederland betreft de realisatie van GPT-NL.[34] Non-profitorganisaties TNO, NFI en SURF gaan samen een taalmodel ontwikkelen om zo een belangrijke stap te zetten richting een transparant, eerlijk en toetsbaar gebruik van AI naar Nederlandse en Europese waarden, met respect voor eigenaarschap van data. Er komt ook een connectie met de nationale supercomputer bij SURF. Het doel van GPT-NL is om minder afhankelijk te zijn van commerciële partijen en daarvoor een verantwoord en transparant alternatief te bieden. GPT-NL wordt een virtuele faciliteit die open staat voor partners die met data en kennis willen bijdragen aan een taalmodel en toepassingen willen ontwikkelen, bijvoorbeeld op het gebied van veiligheid, gezondheid, onderwijs en overheidsdienstverlening.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het stimuleren van de ontwikkeling van (open) Nederlandse en Europese LLM's in lijn met publieke waarden. GPT-NL vormt hier een startschot voor.[35]	Alle uitgangspunten	2023-2026	Kabinetsbreed
Open State Foundation (OSF) ontwikkelt als project (via een subsidie van BZK) een LLM die is getraind op Nederlandse open overheidsinformatie (waaronder openbaar beschikbare Kamerstukken en speeches). Doel hiervan is o.a. om de kansen en risico's van huidige taalmodellen voor de democratie in kaart te brengen.	Uitgangspunt 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	2023-2024	OSF (i.s.m. BZK)

33. Kamerstukken II 2022/23, 26643, nr. 1056
34. In november 2023 is voor de realisatie van GPT-NL 13,5 miljoen euro toegezegd via de regeling Faciliteiten Toegepast Onderzoek van RVO/EZK.
35. tno.nl/nl/newsroom/2023/11/nederland-start-bouw-gpt-nl-eigen-ai/

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het verkennen van de inrichting van een veilige en bruikbare publieke nationale AI-(test) faciliteit voor verantwoorde (generatieve) AI.	Alle uitgangspunten	2024-2025	Kabinetsbreed
Het realiseren dat generatieve AI op een verantwoorde manier kan worden ingezet in een veilige omgeving binnen de overheid. Dit gebeurt onder andere door verschillende pilots.	Uitgangspunt 1. Veiligheid, 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	Doorlopend	BZK (i.s.m. medeoverheden)
In 2024 zullen vanuit AiNed InnovatieLabs gestart worden. InnovatieLabs zijn publiek-private samenwerkingen gericht op de ontwikkeling van AI-innovaties, met een focus op mkb en start- en scale-ups. In de AiNed InnovatieLabs wordt kennis op het gebied van (generatieve) AI vanuit kennisinstellingen en (deep) techbedrijven samengebracht met als doel AI-innovaties sneller naar de markt te brengen en kennis te delen.[36]	Uitgangspunt 4. Duurzaamheid en welvaart	2024	EZK
Het inzetten op een AiNed call die zich o.a. richt op ELSA (Ethical, Legal and Societal) aspecten van de AI-verordening voor bestaande en sterk in ontwikkeling zijnde AI technologie zoals generatieve AI.[37]	Alle uitgangspunten	2024	EZK
Een Rijks-AI-validatieteam faciliteert publiek beschikbare benchmarking en tooling (zoals bias-detectie, op basis van bijvoorbeeld democratische input) om vangrails aan te brengen voor verantwoorde generatieve AI in Nederland.	Uitgangspunt 1. Veiligheid, 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	Doorlopend (vanaf 2024)	BZK (i.s.m. medeoverheden)
Het opnemen van ethische kaders, en eventueel ook tools, rondom het verantwoord gebruik van generatieve AI in de doorontwikkeling van het implementatiekader voor algoritmen (IKA)[38]. Hierin is ook aandacht voor de ondersteuning van ontwikkelaars en gebruikers bij de implementatie van de AI-verordening bij bepalingen die raken aan generatieve AI.	Uitgangspunt 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	2024-2026	BZK (i.s.m. medeoverheden)

36. Vooraankondiging AiNed InnovatieLabs (2024 Stichting AiNed call) - AiNed
37. Vooraankondiging AiNed ELSA Labs (2024 NWO call) - AiNed
38. rijksoverheid.nl/documenten/rapporten/2023/06/30/implementatiekader-verantwoorde-inzet-van-algoritmen

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het stimuleren van en/of onderzoek doen naar ontsluitingsmethoden (‘disclosure methods’) ten behoeve van transparantie over de herkomst en waarachtigheid van AI gegenereerde content. Zoals bijvoorbeeld ‘watermerken’ en het oormerken van AI-content met behulp van cryptografie.	Uitgangspunt 1. Veiligheid, 2. Rechtvaardigheid en 3. Welzijn en autonomie	2024-2026	BZK

f Sterk en helder toezicht houden en handhaven

Ontwikkelaars, beleidsmakers en toezichthouders op Europees en nationaal niveau moeten alert zijn op eventuele ongewenste effecten die de komende jaren kunnen ontstaan rondom generatieve AI. Een proactieve benadering is hierbij essentieel, waarbij toezichthoudende instanties en overheden vanaf het begin heldere kaders bieden om de ontwikkeling van generatieve AI in goede banen te leiden en ongewenste (generatieve) AI te weren.

Sectorale toezichthouders hebben een belangrijke rol bij het zorgen voor effectieve controle op het gebied van (generatieve) AI, controle die in lijn is met geldende wet- en regelgeving en publieke waarden. De in 2023 opgerichte Directie Coördinatie Algoritmes (DCA) bij de AP heeft onder meer coördinerende taken op deze gebieden.[39]

Om sterk en helder toezicht te kunnen houden, moeten toezichthouders tijdens de ontwikkeling en het gebruik kennis en informatie opdoen om de ontwikkelingen te kunnen blijven volgen en waar nodig bij te sturen. Goede samenwerking is daarom van groot belang, bijvoorbeeld via het Samenwerkingsplatform Digitale Toezichthouders (SDT) of de Inspectieraad. Het tijdig en effectief kunnen ingrijpen bij overtredingen en ongewenste effecten is een samenspel tussen toezichthouders, (rechterlijke) instanties, politiek en samenleving. Hierbij is openheid van belang zodat onder meer wetenschap, journalistiek, burgers en politiek een kans hebben om relevante ontwikkelingen rondom generatieve AI te analyseren en controleren.

Gezien het toenemende aantal generatieve AI-toepassingen in de komende jaren, is het van belang om te blijven evalueren of toezichthouders beschikken over de kennis en kunde, capaciteit en middelen om hun taken nu en in de toekomst effectief te kunnen uitvoeren.[40] Daarbij kan ook worden overwogen om de praktijkkennis van toezichthouders op het gebied van (generatieve) AI in te zetten voor wetgevingsadvisering. Dit sluit goed aan op een lerende aanpak, waarbij de komende jaren gemonitord dient te worden of geldende wet- en regelgeving nog voldoende effectief is en beschermt gelet op de ontwikkelingen rondom generatieve AI.

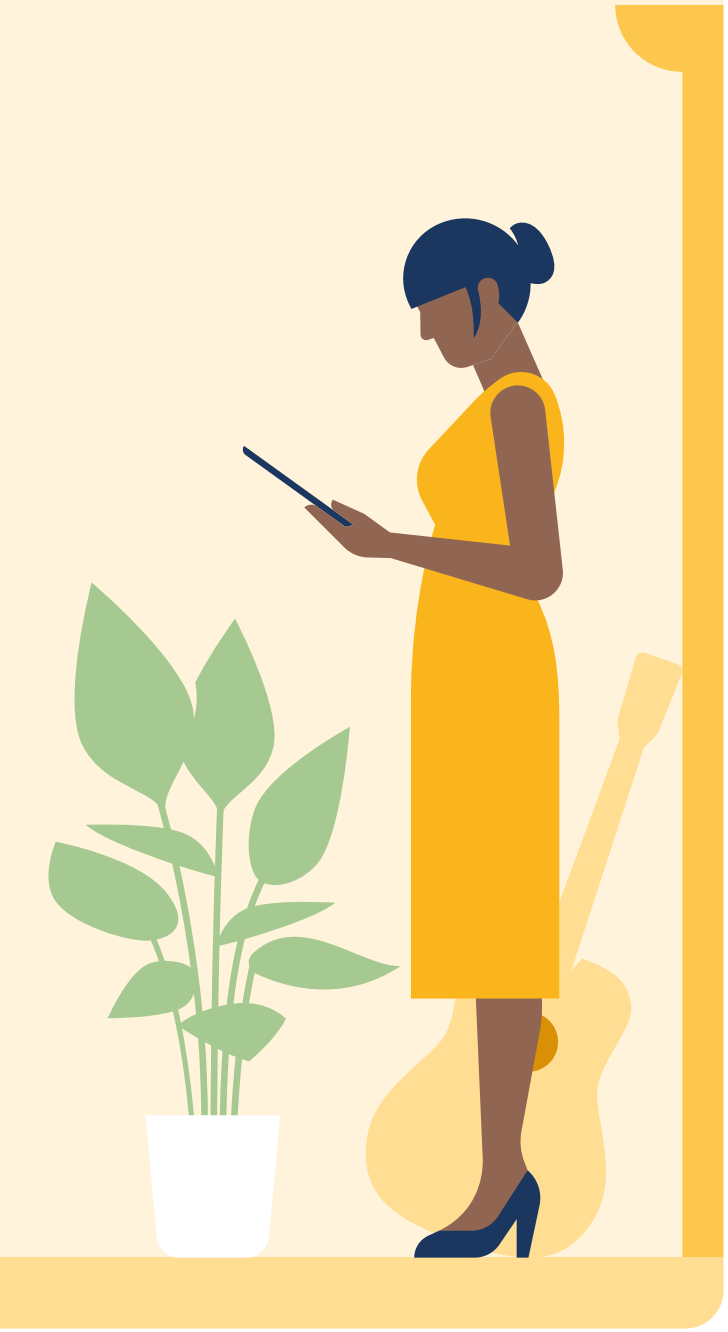
39. Zie ook: rijksoverheid.nl/documenten/kamerstukken/2022/12/22/kamerbrief-over-inrichtingsnota-algoritmetoezichthouder
40. Rathenau Instituut (2023), Generatieve AI: p. 38.

Lopende acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Implementatie van het toezicht op de AI-verordening	Alle uitgangspunten	2024-2027	EZK en BZK i.s.m. JenV, OCW, SZW, I&W, VWS, LNV, BZ, Financiën, Defensie en toezichthouders waaronder AP en RDI.
De implementatie van regulatory sandboxes uit de AI-verordening in Nederland in gezamenlijkheid met de toezichthouders.	Uitgangspunt 1. Veiligheid, 2. Rechtvaardigheid en 4. Duurzaamheid en welvaart	2023-2024	EZK (i.s.m. andere departementen, toezichthouders)
Het stimuleren van gezamenlijke guidance en uitleg (van toezichthouders) en het scheppen van overzicht in bestaande en nieuwe wettelijke kaders (zoals de AI-verordening) op het vlak van algoritmes en (generatieve) AI.	Uitgangspunt 1. Veiligheid en 2. Rechtvaardigheid	Doorlopend (vanaf 2023)	AP
Het kabinet zet Europees in op inclusie van AI-toepassingen binnen de scope van de Digital Markets Act en draagt bij aan effectieve handhaving.	Uitgangspunt 4. Duurzaamheid en welvaart	Doorlopend	EZK

Nieuwe acties

Actie samenvatting	Geeft invulling aan uitgangspunt(en)	Tijdlijn	Eigenaar
Het blijven inzetten op wetgevingsadvisering door toezichthouders over wettelijke kaders voor generatieve AI.	Alle uitgangspunten	Voorafgaand aan de inwerkingtreding van wet- en regelgeving en daarna doorlopend	Kabinetsbreed



6 Vervolg en slotwoord

Generatieve AI wordt een krachtig verlengstuk van het analytisch en scheppend vermogen van mensen. Generatieve AI maakt deel uit van bredere ontwikkelingen op het gebied van digitalisering en (traditionele) AI. Generatieve AI kenmerkt zich ten opzichte van traditionele vormen van AI door de schaal, ontwikkelingssnelheid en wijdverbreide beschikbaarheid van de technologie.

De impact van de technologie zal zich naar verwachting de komende jaren steeds sterker gaan manifesteren. Op basis van wetenschappelijke bevindingen en voorspellingen van experts tekenen de contouren van de impact van generatieve AI zich nu al af. Het is daarom belangrijk de ontwikkelingen en gevolgen van generatieve AI te blijven monitoren en analyseren.

Generatieve AI heeft zowel positieve als negatieve effecten. Veel hangt af van de specifieke ontwikkeling, toepassing en inbedding van de technologie. Adequate regie op generatieve AI is daarom belangrijk. In lijn met breder digitaliseringsbeleid, zoals de Werkagenda Waardengedreven Digitaliseren, hanteert deze visie een waardengedreven benadering. Het kabinet wil dat generatieve AI-toepassingen en de daaraan verbonden technologieën in dienst staan van het vergroten van het menselijk welzijn en autonomie, de duurzaamheid, welvaart, rechtvaardigheid en veiligheid. Het heeft de ambitie om daarin voorop te lopen in Europa en met Europa in de wereld. Het kabinet wil randvoorwaarden creëren voor de ontwikkeling en het gebruik van verantwoorde generatieve AI, onafhankelijk van commerciële of geopolitieke machtsblokken.

Om deze visie te realiseren, zijn concrete beleidsacties beschreven. Eind 2024 zal uw Kamer worden geïnformeerd over de voortgang. Hierbij zal aandacht zijn voor de noodzaak van eventuele nieuwe acties of beleid, ook gezien de komst van een nieuw kabinet. De razendsnelle ontwikkelingen rechtvaardigen een iteratieve en lerende aanpak.

Om de in deze visie geformuleerde uitgangspunten te verwezenlijken, zullen de hierbij geformuleerde concrete acties de komende jaren gemonitord worden. Deze acties zijn gericht op samenwerken, het nauwgezet volgen van alle ontwikkelingen, vormgeven en toepassen van wet- en regelgeving, vergroten van kennis en kunde, innoveren met generatieve AI en sterk en helder toezicht houden (en handhaven). Het succes van deze acties valt of staat bij het verder opbouwen van een werkend (generatief) AI-ecosysteem in Nederland en Europa. Zo kan de rol van ons land als een van de Europese koplopers op het gebied van veilige en rechtvaardige (generatieve) AI tot wasdom komen en kunnen mensen in Nederland daadwerkelijk de vruchten plukken van deze technologie.

Bijlage 1: Aanpak visietraject

Hieronder wordt nader ingegaan op de aanpak die is gehanteerd om tot de overheidsbrede visie op generatieve AI te komen.

a Open aanpak

Een breed scala aan stakeholders heeft bijgedragen om te komen tot een overheidsbrede visie op generatieve AI. Daarbij dankt het kabinet alle mensen en organisaties die de afgelopen maanden hebben meegedacht om deze visie te realiseren.

Vanaf mei 2023 zijn er diverse acties ondernomen om tot een open en breed gedragen overheidsbrede visie op generatieve AI te komen, getoetst in meerdere sectoren en domeinen binnen de Nederlandse samenleving. De aanpak kenmerkte zich door het ophalen van input bij een veelheid aan experts en door middel van verschillende soorten sessies, het tussendoor delen van resultaten met de buitenwereld en het delen en beproeven van uitkomsten via een online Pleio-community en andere kanalen.

- Tussen juni en november 2023 hebben verschillende (sector)sessies plaatsgevonden. Deze sessies zijn gehouden in sectoren als het openbaar bestuur, de zorg, werkgelegenheid en de economie. In deze sessies is input opgehaald voor de visie en is tijdens de verdere totstandkoming van de visie steeds ook getoetst of de juiste aspecten in de visie zijn meegenomen. In 2024 zullen deze gesprekken worden voortgezet.
- In samenwerking met ECP zijn er in het najaar van 2023 ontmoetingen georganiseerd met de media, het hoger onderwijs, de zorg en bij de politie.[1] In deze sessies stonden ethische dilemma's rondom de toepassing van generatieve AI centraal.
- In samenwerking met de Nederlandse AI Coalitie (NL-AIC) zijn er een aantal sessies met inwoners van Nederland georganiseerd. Deze sessies hadden enerzijds als doel om bewustwording te creëren over de technologie en anderzijds om te achterhalen hoe mensen aankijken tegen de impact van generatieve AI op hun levens en de rol van de overheid voor de inbedding van deze technologie in de samenleving. Gedurende het visietraject zijn er tussendoor verschillende keren uitkomsten en inzichten gedeeld, bijvoorbeeld via de Pleio-omgeving.[2]
- Om te borgen dat de visie een zo breed mogelijk perspectief op generatieve AI weerspiegelt is er een werkgroep generatieve AI gestart, met daarin deelnemers van een groot aantal ministeries, provincies (IPO) en gemeenten (VNG).
- Ook is er een klankbordgroep ingericht bestaande uit leden die verschillende maatschappelijke perspectieven vertegenwoordigden.[3]

b Catshuissessie

Op 6 september 2023 is er voor leden van het kabinet een Catshuissessie generatieve AI gehouden. Hierin werd stilgestaan bij de vraag hoe Nederland zich als land en verantwoorde proeftuin kan positioneren op het gebied van generatieve AI, waarbij veel aandacht was voor zowel de ethische aspecten als de kansen die deze technologie biedt. Ook werd in de Catshuissessie benadrukt dat Nederland na het initiatief van de REAIM Summit, een verschil kan maken op het thema verantwoorde AI in het militaire domein.

c Techscan Rathenau

Vanwege zijn onafhankelijkheid en kennis en expertise op het snijvlak van beleid en digitalisering, is het Rathenau Instituut gevraagd om met de zogenoemde 'techscan-methode' een analyse te maken van generatieve AI.[4] Dit rapport is in december 2023 gepubliceerd. Het drieledige doel hiervan was om: (1) de maatschappelijke impact van generatieve AI in een vroeg stadium te duiden door de kansen en de risico's te identificeren vanuit een publieke-waardenperspectief, (2) bestaand beleid om de kansen te verzilveren en de risico's te adresseren te evalueren; en (3) mogelijke handelingsopties in kaart te brengen.

Door middel van deze techscan wordt ook invulling gegeven aan de toezegging die op 22 maart 2023 door de staatssecretaris van BZK aan het lid Van Weerdenburg (PVV) is gedaan tijdens het Commissiedebat 'Digitale infrastructuur en economie' om na de zomer met een onderzoek naar de impact van AI op de samenleving te komen.

1. https://begeleidingsethiek.nl/cases/
2. https://generatieveai.pleio.nl/
3. In de klankbordgroep zaten vertegenwoordigers van: Bits of Freedom, CIO-Platform, IPO, FNV, NL-AIC, Politie, SER, VNG en VNO-NCW.
4. Rapport van het Rathenau Instituut over generatieve AI (2023): https://www.rathenau.nl/sites/default/files/2023-12/Scan_Generatieve_AI_Rathenau_Instituut.pdf

Bijlage 2: Hoe komt generatieve AI tot stand?

De totstandkoming van generatieve AI-modellen is onder te verdelen in drie fases: pre-training, finetuning, en toepassing (deployment). Hieronder worden deze fases beschreven.

De **pre-trainingsfase** is een cruciale stap in het trainen van generatieve AI-modellen. Tijdens deze fase wordt het model gevoed met trainingsdata, afkomstig van publieke en gesloten bronnen. De overvloed aan data is cruciaal omdat het de modellen in staat stelt om een breed scala aan concepten, (taal)structuren, contextuele nuances en representaties van de wereld te analyseren en categoriseren. Niet alleen tekst, maar ook audio, video en afbeeldingen kunnen als gegevensbronnen dienen, en deze gegevens kunnen gecombineerd worden in één model.

Tijdens de pre-trainingsfase leert het model om de parameters te optimaliseren zodat het model steeds nauwkeuriger correlaties en patronen in de trainingsdata kan ontdekken. Generatieve AI-modellen hebben richting de biljoen parameters die biljoenen iteraties nodig hebben om een gewenste waarde te bereiken. Deze immense omvang maakt het proces van input naar output minder inzichtelijk en geeft de modellen hun black box karakter. Het trainingsproces vereist aanzienlijke computerkracht en vormt vaak de beperkende factor bij het trainen van AI-modellen. Daarom worden generatieve AI-modellen getraind op gespecialiseerde hardware. Dankzij snelle ontwikkelingen in hardware in de afgelopen jaren is het voor steeds meer partijen mogelijk om grotere, complexere AI-modellen te trainen. Het is ook opmerkelijk dat de pre-trainingsfase bij dit soort modellen relatief weinig menselijke handelingen vereist. Dit maakte snelle schaalvergroting mogelijk.

Finetuningsfase: De pre-training fase resulteert in een rudimentair basismodel dat nog niet geschikt is voor breed gebruik. Het basismodel wordt door middel van finetuning verfijnd, waarbij het model leert om instructies van gebruikers op te volgen. Finetuning wordt ook gebruikt om het model specialistische kennis mee te geven of specifieke waarden en normen toe te voegen. Een ontwikkeling ten aanzien van finetuning is het gebruik van Reinforcement Learning from Human Feedback. Dit is een vrij complexe methode, waarbij (generatieve) AI-modellen niet worden beoordeeld op voorspellende vaardigheden, maar op de behulpzaamheid, eerlijkheid en veiligheid van de modeluitkomsten. Hierbij wordt de uitkomst door mensen beoordeeld en gelabeld. Ontwikkelaars hebben inmiddels methoden gepresenteerd waarbij AI-modellen zelf output integraal beoordelen op ethische aspecten, zoals Constitutional AI.[1]

In de zoektocht naar meer en beter trainingsmateriaal is het gebruik van synthetische data een logische volgende stap, nadat beschikbare open en besloten databronnen al zijn benut. Synthetische data wordt gegenereerd door een combinatie van verschillende databronnen en met behulp van generatieve AI. Dit heeft een impliciet versnellend effect. Dankzij snellere en betere AI kunnen zo ook hoogwaardige training- en finetuning data worden gegenereerd. Hoogwaardige AI resulteert in hogere kwaliteit van trainingsmateriaal, wat op zijn beurt weer leidt tot verbetering van het AI-systeem. De finetuningsfase resulteert in een model dat geschikt is voor gebruik.

Toepassingsfase: Tijdens de toepassingsfase wordt het model beschikbaar gesteld aan gebruikers. Het model kan worden hergebruikt en gedupliceerd, zodat bedrijven met toegang tot veel rekenkracht tienduizenden tot miljoenen gebruikers tegelijkertijd kunnen faciliteren. Waar in de trainingsfase maanden nodig zijn voor het trainen van een model, kan een model na toepassing binnen een paar seconden antwoord geven.[2]

Een ontwikkelaar van het model kan na de toepassingsfase besluiten het model open source te maken. Dit houdt in dat de broncode van het model (en soms ook andere componenten) wordt gepubliceerd om in te zien, te analyseren, te hergebruiken en op voort te bouwen met eigen aanpassingen (finetuning). Modellen kunnen op deze manier worden verbeterd, maar ook worden verslechterd, omdat er bijvoorbeeld geen controle meer is op verdere finetuning van het model. Het beschikbaar maken van de code en de parameters van generatieve AI-modellen neemt daarbij niet weg dat het black box-modellen blijven, waarvan de capaciteiten niet direct herleidbaar zijn. Daarnaast kan het per aanbieder verschillen hoeveel informatie over het model daadwerkelijk openbaar wordt gemaakt. Open source staat daarmee niet per definitie gelijk aan meer transparantie, veiligheid of verhoogde morele gepastheid.

1. Constitutional AI streeft ernaar (generatieve) AI te ontwikkelen die in lijn is met menselijke waarden door de uitkomsten van een AI-model automatisch getoetst te toetsen aan principes die ook democratisch opgesteld kunnen worden.
2. Een voorbeeld hiervan is Microsoft 365 Copilot.

Bijlage 3: Begrippenlijst overheidsbrede visie op generatieve AI

1. *Algoritme*
   Een set van regels en instructies die een computer uitvoert.

2. *Artificial general intelligence (AGI)*
   Een technologie die intelligentie heeft over een breed scala aan domeinen waaronder redeneren, plannen en leren, en met deze vaardigheden op of boven het menselijke niveau presteert.

3. *Artificial intelligence (AI)-chatbot*
   Digitale chatbots die via tekst en afbeeldingen kunnen communiceren op een manier die sterk lijkt op menselijke interactie. Een van de bekendste AI-chatbots op dit moment is ChatGPT.

4. *Artificiële intelligentie (AI)-systeem*
   Een op machines gebaseerd systeem dat, voor expliciete of impliciete doelstellingen, afleidt, uit de input die het ontvangt, hoe het output zoals voorspellingen, inhoud, aanbevelingen of beslissingen moet genereren die fysieke of virtuele omgevingen kunnen beïnvloeden. Verschillende AI-systemen variëren in hun mate van autonomie en aanpassingsvermogen na de implementatie (deployment) ervan (OESO 2023).

5. *Bijzondere categorieën persoonsgegevens*
   Persoonsgegevens waaruit ras, etnische afkomst, politieke opvattingen, religieuze of levensbeschouwelijke overtuigingen, het lidmaatschap van een vakbond blijken, of de verwerking van genetische gegevens, biometrische gegevens met het oog op de unieke identificatie van een persoon, gegevens over gezondheid, gegevens met betrekking tot iemands seksueel gedrag of seksuele gerichtheid.

6. *Black box-model*
   Een (AI-)model waarvan er inzicht ontbreekt in hoe de voorspelling van het model tot stand is gekomen en wat de grondslag voor het gevormde model is.

7. *Dark patterns*
   Interfaces, vooral in online gebruikersinterfaces, die de autonomie, besluitvorming en keuze van consumenten kunnen schaden. Ze misleiden, dwingen of manipuleren consumenten vaak, wat waarschijnlijk directe of indirecte schade veroorzaakt. Het kan echter moeilijk zijn om deze schade te meten.

8. *Deepfake*
   Een met technologie gecreëerde foto, video of audio waarop te zien of te horen is dat een persoon dingen doet of zegt, die hij of zij niet daadwerkelijk heeft gedaan of gezegd.

9. *Desinformatie*
   Desinformatie is het doelbewust, veelal heimelijk, verspreiden van misleidende informatie, met het doel om schade toe te brengen aan het publieke debat, democratische processen, de kenniseconomie of volksgezondheid.

10. *Finetuning*
    Tijdens het 'finetunen' van een AI-model wordt een reeds getraind model aangepast aan een specifieke taak of dataset. In plaats van een model 'from scratch' te trainen, wordt gebruik gemaakt van een bestaand model.

11. *Foundation model*
    Een foundation model is een basis machine learning model dat als fundament dient voor verdere gespecialiseerde modellen. Een large language model (LLM) is een type foundation model. Een voorbeeld van een foundation model is GPT-4, het foundation model voor ChatGPT.

12. *Generatieve AI*
    Een vorm van AI waarbij complexe algoritmes worden ingezet om nieuwe content te genereren zoals tekst, afbeeldingen, computercode of video's. Chatbot ChatGPT vormt hiervan een bekende exponent.

13. *Hallucinatie*
    Door een 'large language model' (LLM) gegenereerde informatie die feitelijk onjuist is. Het model genereert antwoorden die niet gebaseerd zijn op de gegeven input of op feitelijke informatie uit de trainingsdata.
    Een hallucinatie kan worden veroorzaakt door verschillende redenen, zoals een gebrek aan specifieke informatie in de trainingsdata, een gebrek aan context of foutieve en inconsistente informatie in de trainingsdata.

14. *Jailbreaking*
    Het ontwerpen van prompts met de intentie model biases uit te buiten om zo output te genereren die niet strookt met het doel van het model. Het model zal bijvoorbeeld antwoord geven op vragen die normaliter door het model niet zouden worden beantwoord.

15. *Large language model (LLM)*
    Een gespecialiseerd type AI-model dat getraind is op grote hoeveelheden tekst om bestaande content te begrijpen en content te genereren.

16. *Machine learning*
    Een deelgebied van artificiële intelligentie dat computers in staat stelt om te leren van data. Een machine learning algoritme leert van voorbeelden en ervaringen om patronen en regels in data te ontdekken.

17. *Model*
Een AI-model is het resultaat van het trainen van een algoritme op data. Een algoritme is een set instructies, en het model is het specifieke resultaat van het volgen van de set instructies op basis van bepaalde data. GPT-4 is een voorbeeld van een AI-model, in dit geval een large language model.

18. *Model collapse*
Een fenomeen wat kan voorkomen wanneer LLM's worden getraind met 'vervuilde' data; een database die door AI-gegenereerde data bevat. Er wordt gesproken van model collapse wanneer een LLM foutieve en minder gevarieerde output genereert, veroorzaakt door vervuilde data.

19. *Modelparameters*
Aanpasbare instellingen in AI-modellen die beslissen hoe een LLM output genereert. Modelparameters hebben invloed op de kwaliteit, diversiteit en creativiteit van de output. Modelparameters komen op verschillende manieren tot stand, waaronder wiskundige berekeningen maar ook menselijke inmenging.

20. *Open source*
Het open source publiceren van een generatief AI-model houdt in dat de broncode van het model (en soms ook andere componenten) wordt gepubliceerd om in te zien, te analyseren, te hergebruiken en op voort te bouwen met eigen aanpassingen (*finetuning*).

21. *Persoonsgegeven(s)*
Alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon ('de betrokkene'). Als 'identificeerbaar' wordt beschouwd een natuurlijke persoon die direct of indirect kan worden geïdentificeerd, met name aan de hand van een identificator zoals een naam, een identificatienummer, locatiegegevens, een online identificator of van een of meer elementen die kenmerkend zijn voor de fysieke, fysiologische, genetische, psychische, economische, culturele of sociale identiteit van die natuurlijke persoon.

22. *Pre-training*
Tijdens pre-training wordt een AI-model gevoed met grote hoeveelheden trainingsdata (tekst, audio, afbeeldingen, video) van verschillende bronnen. Het model leert tijdens pre-training patronen te herkennen in de data. Dit vereist aanzienlijke computerkracht en wordt uitgevoerd op gespecialiseerde hardware.

23. *Reinforcement Learning from Human Feedback (RLHF)*
In het geval van RLHF wordt menselijke feedback opgenomen in het trainingsproces van AI-algoritmes om het leren van het AI-algoritme te sturen of te verbeteren. Deze feedback van mensen kan als effect hebben dat het algoritme sneller en effectiever kan leren. Het doel is vaak om menselijke expertise te benutten om AI-algoritmes een bepaalde gewenste richting op te sturen.

24. *Systeem*
Een AI-systeem omvat niet alleen het model, maar ook de gehele infrastructuur eromheen. Dit omvat de hardware, software, gegevensverwerking, input- en outputinterfaces, en alle componenten die nodig zijn om het model effectief te laten werken. Een voorbeeld van een AI-systeem is ChatGPT.

25. *Taakspecifieke AI ('narrow AI')*
AI die geprogrammeerd is voor één specifieke taak, in tegenstelling tot generatieve AI, die kan worden ingezet voor een breed scala aan taken.

26. *Training*
Het proces waarbij een algoritme wordt geleerd om patronen in gegevens te herkennen.

27. *Transparantie*
Een model is transparant wanneer bekend is met welke formules, handelingen en waarden een model output genereert. Een transparant algoritme is het tegenovergestelde van een black box-algoritme.

28. *Uitlegbaarheid*
Een model is uitlegbaar wanneer uit te leggen en te begrijpen is waarom het model bepaalde output genereert. Uitlegbaarheid zorgt ervoor dat een mens kan begrijpen waarom een model iets doet zonder dat een mens de formules, handelingen en waarden binnen het model hoeft te kennen. Een uitlegbaar model is daarom niet per definitie transparant, en een transparant model is niet per definitie uitlegbaar.

29. *Vangrails ('guardrails')*
Beperkingen, richtlijnen of veiligheidsmaatregelen die worden ingesteld om ervoor te zorgen dat het gebruik van LLM's binnen ethische en verantwoorde grenzen blijft.

30. *Webscraping*
Het gebruik van software om informatie van webpagina's te extraheren om deze vervolgens te analyseren.

Colofon

Dit is een uitgave van:
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties
Postbus 20011
2500 EA, Den Haag

Januari 2024