Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Overheidsbrede handreiking voor de verantwoorde inzet van generatieve AI

Versie 1.4

Datum Februari 2025
Status Concept

Naam	**Overheidsbrede handreiking voor de verantwoorde inzet van generatieve AI**
Type	Handreiking
Status	Concept
Versie	1.4

Versie	**Omschrijving**
0.78	Eerste versie voor feedbackronde 1
0.79	Tweede versie voor feedbackronde 2 IBWG
0.80	Derde versie voor CDO raad
0.90	Vierde versie voor vooroverleg CIO-Beraad
0.91	Vijfde versie voor feedbackronde 5
0.95	Zesde versie voor CIO-beraad
0.96	Zevende versie na CIO-beraad
1.2	Achtste versie voor feedbackronde, DT CIO Rijk
1.3	Negende versie CDO raad
1.31	Tiende versie CIO-beraad
1.32	Elfde versie voor PO Stass
1.4	Interbestuurlijke versie handreiking

Inhoud

Inleiding

Generatieve AI is een krachtig verlengstuk van het analytisch en scheppend vermogen van mensen. Het heeft uitgebreide mogelijkheden om maatschappelijke vraagstukken aan te pakken. Ook kan het de arbeidsproductiviteit verhogen. Tegelijkertijd staan er mogelijk rechten onder druk en zijn er ook ingrijpende risico's. Zo kan het discriminatoire dynamieken en sociaaleconomische ongelijkheden versterken.[1]

De Nederlandse overheid heeft een voorbeeldfunctie als het gaat om het verantwoord en veilig inzetten van opkomende technologieën zoals generatieve AI. Dit vraagt om overheidsorganisaties die over de juiste gereedschapskist beschikken om deze technologie – in verschillende facetten van hun werk – verantwoord en rechtmatig in te zetten.

Een uitgangspunt is hierbij het voldoen aan de eisen van de Europese AI-verordening (hierna: AI-verordening) en andere geldende relevante regelgeving. De juridische aspecten moeten zorgvuldig worden bekeken bij het inzetten van generatieve AI om te waarborgen dat er wordt voldaan aan wet- en regelgeving. Als duidelijk is dat generatieve AI positief kan bijdragen aan het behalen van de organisatiedoelen, rijzen de vragen 'kan het?', 'mag het?' en 'is het verantwoord?'. Hierbij is digitale ethiek fundamenteel. Naast de juridische en ethische overwegingen dient nagedacht te worden over de veiligheid waarbij privacy en cyberveiligheid van belang zijn. Ten slotte bestaat verantwoorde inzet voor een groot deel uit het toewijzen van de juiste rollen en verantwoordelijkheden.

Leeswijzer
In hoofdstuk 1 wordt de afbakening van deze handreiking uitgelegd aan de hand van positionering en scope van het document. Daarna volgt een hoofdstuk over wat generatieve AI is en hoe deze technologie wordt ingezet bij de overheid. Hoofdstuk 3 vertelt over de kansen van generatieve AI en laagdrempelige toepassingsgebieden bij overheidsorganisaties. Hoofdstuk 4 legt een stappenplan voor, met aanbevelingen voor organisaties die aan de slag willen met generatieve AI. Hierin wordt verwezen naar de volgende hoofdstukken, die informatie geven over de juridische-, ethische-, en veiligheidsaspecten en governance van generatieve AI. Hoofdstuk 9 bevat een lijst met concrete aanbevelingen die betrokken overheidsmedewerkers kunnen communiceren richting eindgebruikers. De handreiking sluit af met een begrippenlijst en een bijlage waarin beknopt wordt uitgelegd hoe generatieve AI-systemen tot stand komen.

[1] Overheidsbrede visie op generatieve AI

1. Positionering en scope

1.1 Scope

De handreiking gaat nader in op de technologische, organisatorische, ethische en juridische (rand)voorwaarden die van meewaarde zijn om als overheidsorganisatie tot de verantwoorde inzet van generatieve AI te komen. Deze handreiking bestrijkt het hele palet van de inzet van generatieve AI: van experimenteren tot en met toepassen en monitoren.

Dit document beperkt zich tot het bieden van een handreiking. Het document is niet bindend. Het volgende valt buiten de scope van dit document:

- Het opstellen van of voorzien in nieuwe verplichtingen of voorwaarden bij de inzet van generatieve AI;
- Hoe de toezichthoudende instanties hun taak als toezichthouder op het gebruik van generatieve AI-toepassingen vervult.
- Het geven van richting over de inkoop, de ontwikkeling en het gebruik van AI in de brede zin, dit document gaat enkel in op generatieve AI;
- Normuitleg of het samenbrengen van bestaande kaders.

1.2 Gerelateerde documenten

- Overheidsbrede visie op generatieve AI: in deze overheidsbrede visie wordt het belang van generatieve AI die in dienst staat van het vergroten van het menselijk welzijn en autonomie, duurzaamheid, welvaart, rechtvaardigheid en veiligheid, benadrukt.
- PM: Verwijzing naar Overheidsbrede standpunt inzet generatieve AI.
- Het Algoritmekader: geeft een overzicht van bestaande wet- en regelgeving die op algoritmes van toepassing kunnen zijn.
- Het Algoritmeregister: dit is een nationale openbaar beschikbare publicatie van door overheid ingezette algoritmes met bijbehorende toelichting.

Deze handreiking sluit aan op de overheidsbrede visie op generatieve AI en het herijkte standpunt voor overheidsorganisaties voor de verantwoorde inzet van generatieve AI. Daarnaast kan deze handreiking aanvullend gebruikt worden, wanneer van toepassing, bestaande wet- en regelgeving die een rol kan spelen bij de inzet van generatieve AI en in het bijzonder ook de Europese AI-verordening. De reden hiervoor is tweeledig. Ten eerste wordt generatieve AI momenteel al veelvuldig ingezet binnen overheidsorganisaties[2], terwijl er nog een beperkt aantal Europese en nationale standaarden zijn vastgesteld en verschillende bepalingen uit de AI-verordening momenteel nog niet volledig van kracht zijn. Deze handreiking geldt als direct inzetbaar document, om overheidsorganisaties de juiste handvatten te bieden om generatieve AI op een verantwoorde manier in te zetten. Ten tweede biedt deze handreiking concrete aanbevelingen voor overheidsorganisaties en eindgebruikers aan. In deze context zijn eindgebruikers medewerkers van overheidsorganisaties.

[2] Zie de rapporten: Focus op AI bij de rijksoverheid | Rapport | Algemene Rekenkamer en 3 Snelle doorlichting impact generatieve AI op (IT-) overheidspersoneel | Rapport | Rijksoverheid.nl

1.3 Organisatiespecifieke versie van de handreiking

Organisaties binnen de overheid worden aangemoedigd om een eigen versie van de handreiking uit te brengen:

- De aanbevelingen voor overheidsmedewerkers die zij kunnen omzetten naar aanbevelingen zijn te vinden in het hoofdstuk Aanbevelingen.
- Overheidsorganisaties kunnen aanvullende (strengere) regels vaststellen, bijvoorbeeld indien een organisatie veel vertrouwelijke of gevoelige informatie verwerkt.
- Als een organisatie een eigen versie maakt, wordt in de titel de naam van de organisatie vermeld.

1.4 Status en beheer van het document

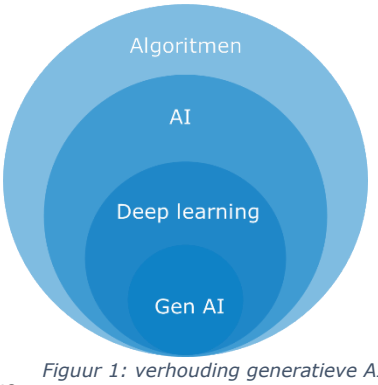
Dit document behoeft zodoende geen centrale toezichts- of controlemechanismen. Wel wordt aangeraden dit document decentraal te vertalen met bijbehorende controlemechanismen. Voor hulp bij specificeren van de handreiking kunnen organisaties bijvoorbeeld terecht bij hun eigen of lokale CIO en CDO offices.

Deze handreiking is een levend document, aangezien de ontwikkelingen in generatieve AI zich in een snel tempo voortzetten. De handreiking wordt daarom periodiek aangevuld en bijgewerkt met nieuw opgedane kennis uit bijvoorbeeld pilots en proeftuinen en nieuwe wet- en regelgeving.

2. Wat is generatieve AI en hoe wordt het ingezet bij de overheid?

2.1 Wat is generatieve AI?

Generatieve AI is een vorm van kunstmatige intelligentie waarbij algoritmes worden ingezet om content te genereren op verzoek van een gebruiker. Met een eenvoudige opdracht in natuurlijke taal, een *prompt*, kunnen gebruikers in een handomdraai content zoals tekst, beeld, geluid, video of computercode genereren. Generatieve AI is ook een voorbeeld van 'general purpose AI' (GPAI). Dit isvolgens de AI-verordening krachtige en AI die voor meerdere toepassingen en domeinen aangepast en ingezet kunnen worden.



Figuur 1: verhouding generatieve AI en bredere speelveld

Generatieve AI-technologie wordt getraind op grote hoeveelheden data, waaruit ze patronen en structuren leren herkennen. Op basis van deze patronen kan het getrainde model voorspellen wat het meest logische volgende element is in bijvoorbeeld een zin, plaatje of muziekstuk. Een uitgebreidere uitleg over hoe een generatief AI-systeem tot stand komt, is te vinden in de bijlage van deze handreiking. Een van de meest herkenbare toepassingen van generatieve AI voor het brede publiek zijn AI-chatbots. Bekende voorbeelden zijn: *ChatGPT*, *CoPilot* en *Gemini*. Dit zijn chatbots op basis van een *large language model* (LLM). Deze modellen zijn gespecialiseerd in natuurlijke taalverwerking en richten zich op het verwerken en voorspellen van tekst.

Generatieve AI-modellen kunnen op verschillende manieren aangeboden worden. Aanbieders kunnen ervoor kiezen de broncode, en eventuele andere componenten zoals de trainingsdata, performance *metrics* en controlemechanismen, toegankelijk te maken (*open source*), zodat het model herbruikbaar is voor derden. De *Open Language Models* (OLM) van 'Allen Institute for AI' zijn hier een voorbeeld van. Wanneer aanbieders de broncode en trainingsdata afgeschermd houden, spreken we van *closed source* modellen (*proprietary models)*. Een voorbeeld van een *closed source* generatief AI-model is *GPT-4*[3].

Ook kan generatieve AI worden ingezet in 'agentic AI'. Agentic AI verwijst naar autonome AI-systemen die zelfstandig doelen kunnen nastreven en acties kunnen ondernemen zonder voortdurende menselijke tussenkomst. Wanneer deze systemen gebruikmaken van generatieve AI, kunnen ze niet alleen beslissingen nemen, maar ook nieuwe content creëren, zoals teksten of afbeeldingen, om hun doelen te bereiken. Zorg bij de inzet van dit soort (generatieve) AI-systemen. Zonder adequate monitoring bestaat het risico dat AI-systemen ongewenste beslissingen nemen die niet in lijn zijn met overheidsbeleid of -ethiek. Wanneer het aantal agents toeneemt zal dit probleem alleen maar toenemen met het aanvullende risico van agents die tegenstrijdig zijn.

2.2 Generatieve AI bij de overheid

Voor de overheid biedt generatieve AI kansen om processen te verbeteren, de dienstverlening aan burgers te optimaliseren en het algemeen functioneren van de overheid te bevorderen. Verschillende overheidsorganisaties maken gebruik van generatieve AI of voeren pilots uit. Om een beter inzicht te krijgen in bepaalde *use cases* van generatieve AI binnen de overheid, en om een beter beeld te krijgen van de mogelijkheden voor de verantwoorde inzet van generatieve AI, is er een AI-community voor overheidsmedewerkers waarin inzichten en ervaringen gedeeld worden. Deze community is te vinden op: https://aienalgoritmes.pleio.nl/.

[3] Zie voor meer voorbeelden van open en closed source modellen deze tabel van Radboud Universiteit: https://opening-up-chatgpt.github.io

3. Kansen van generatieve AI

Dit hoofdstuk beschrijft manieren waarop de Nederlandse overheid de kansen op het gebied van generatieve AI kan verzilveren.

3.1 Vernieuwde doelstellingen

Generatieve AI kan bijdragen aan het behalen van doelen van overheidsorganisaties, zoals ook beschreven in de overheidsbrede visie op generatieve AI[4] en het rapport snelle doorlichting impact generatieve AI op (IT-) overheidspersoneel[5]. Generatieve AI stelt de overheid op twee manieren in staat te innoveren:

- Ten eerste kan de overheid haar huidige doelen mogelijk efficiënter en effectiever bereiken door middel van generatieve AI. Bijvoorbeeld doordat bepaalde processen sneller en gemakkelijker kunnen worden uitgevoerd.
- Ten tweede kan de overheid haar doelen mogelijk zelfs aanpassen, vernieuwen of omhoog bijstellen door generatieve AI te gebruiken. Bijvoorbeeld doordat de technologie nieuwe inzichten verschaft.

3.2 Toepassingsgebieden

Om de kansen te benutten die generatieve AI biedt is het verstandig deze te classificeren in termen van toegevoegde waarde, risico en veiligheid. Op die manier ontstaat een selectie van toepassingen die toegevoegde waarde hebben voor veel overheidsmedewerkers en weinig risico's met zich meebrengen. Dit zijn toepassingen voor 'dagelijks gebruik'. Ze hebben een relatief lage impact, en zijn daardoor laagdrempeliger om in te zetten na analyse en goedkeuring van de verantwoordelijke (bestuurder) binnen de organisatie. Hierbij valt te denken aan de volgende categorieën:

- Genereren van nieuwe ideeën. Generatieve AI kan helpen bij brainstorms, door suggesties te doen voor nieuwe zienswijzen bij de oplossing van problemen.
- Automatisering van repetitieve taken. Generatieve AI kan bijvoorbeeld helpen met het automatisch invullen van formulieren of, net als traditionele AI, bij het analyseren van (bulk)gegevens.
- Vinden van informatie. Generatieve AI kan helpen om publieke of interne informatie sneller te vinden en lokaliseren. Een voorbeeld hiervan is de sAmmie, de virtuele collega van de provincie Noord-Brabant. Daarnaast kan een tool suggesties geven als antwoord op een gestelde vraag, denk bijvoorbeeld aan WOO-verzoeken.
- Ondersteunen van (beleids-)medewerkers. Generatieve AI biedt een gebruiksvriendelijke interface om eenvoudig relevante en geprioriteerde antwoorden uit complexe (beleids)informatie te achterhalen. Het kan bijvoorbeeld suggesties voor antwoorden geven op vragen als "Met welke exacte regelgeving moet mijn nieuwe AI-systeem rekening houden?". Of "Welke regels binnen wetgeving in de horeca zijn strijdig en op welke onderdelen?".
- Grotere betrokkenheid bij de maatschappij. Generatieve AI kan openbare bronnen scannen en analyseren om ondersteuning te bieden bij het signaleren van opkomende sentimenten, trends en thema's.

[4] Overheidsbrede visie Generatieve AI | Rapport | Rijksoverheid.nl
[5] Snelle doorlichting impact generatieve AI op (IT-) overheidspersoneel | Rapport | Rijksoverheid.nl

- Softwareontwikkeling. Generatieve AI kan ontwikkelaars helpen bij het schrijven en controleren van code, door bijvoorbeeld suggesties te doen of fouten op te sporen. Daarnaast kan generatieve AI geautomatiseerd test-cases genereren.
- Tekstbewerking. Generatieve AI kan helpen om bepaalde stukken tekst te vertalen, versimpelen of te verbeteren. Daarnaast kan generatieve AI helpen bij het herschrijven van teksten voor verschillende doelgroepen en het opstellen van bijpassende FAQ's.
- Beeldbewerking. Generatieve AI kan helpen bij het genereren en bewerken van afbeeldingen, grafieken of infographics voor interne communicatie[6].
- Scenariostudies en prognoses. Generatieve AI kan realistische scenario's bedenken en adviseren hoe medewerkers zich hierop kunnen voorbereiden. Dit kan gaan om simulaties op de werkvloer of trends in de samenleving of sector. Deze toepassing is bijvoorbeeld nuttig in training of opleiding van medewerker.
- Publiceren van informatie. Het inzetten van Generatieve AI kan helpen te voldoen aan wetgeving of het vertrouwen verhogen van de bevolking te verhogen door eenvoudiger actieve openbaarmaking van documenten te realiseren en WOO-verzoeken binnen enkele dagen te beantwoorden.
- Inzet van generatieve AI in wetenschappelijk onderzoek.

Voor een juiste werking van de generatieve AI-tool is de het toegang tot volledige, recente en kwalitatief goede data bepalend. Dan nog is het goed om te onthouden dat de huidige stand van de techniek soms nog te wensen over. Hiernaast zijn er mogelijke toepassingen die meer ingrijpen op het primaire proces of waarbij gevoelige (persoons)gegevens gebruikt worden. Deze toepassingen dienen onderhevig te zijn aan risicoanalyses.7 Ook kunnen ze verwerkt worden in gecontracteerde systemen om mogelijke verspreiding van vertrouwelijke gegevens te voorkomen. Te denken valt aan toepassingen rondom burgerinteractie, selectietechnieken en ondersteuning bij besluitvorming en uitvoering. Tot slot is het van belang om bij de inzet van generatieve AI altijd een vorm van menselijke tussenkomst in te bouwen. Zo blijft de ambtenaar aan het stuur bij de inzet van generatieve AI.

[6] Zie ook: Rijkshuisstijl & beeld - Beeldkompas.

7 Zie hiervoor ook het overheidsbrede standpunt inzet generatieve AI.

4. Aan de slag met generatieve AI

In dit hoofdstuk volgt een globaal en praktisch stappenplan die gebruikt kan worden om generatieve AI op een verstandige en verantwoorde manier te benutten.

4.1 Definieer het doel en de toepassingsgebieden

Generatieve AI is een middel om een doel te bereiken, net als iedere andere technologie. Door de (maatschappelijke) doelen voorop te stellen wordt voorkomen dat we ons blindstaren op technologische innovatie buiten enige context. Er wordt dus altijd vertrokken vanuit uitdagingen en probleemsituaties, niet vanuit de oplossing. Generatieve AI is een krachtige toevoeging aan het repertoire van mogelijke oplossing voor problemen of verdere verbeteringen van reeds goed lopende processen.

Voorbeelden van uitdagingen zijn het aantoonbaar verbeteren van publieke dienstverlening, verhoogde productiviteit, minder werkdruk of lagere projectkosten. Er zijn ook *use cases* voor generatieve AI die ten alle tijden vermeden moeten worden.

4.2 Zorg voor de juiste mensen en vaardigheden

Voor het verantwoord toepassen van deze technologie, is het verstandig om een diverse groep medewerkers te betrekken. In ieder geval zijn de volgende expertises nodig bij het bouwen of aanschaffen van generatieve AI:

- Leidinggevenden die de context en impact begrijpen.
- Juristen, ethici, privacy-, en cybersecurity-officers die kunnen helpen om de oplossing verantwoord en veilig te ontwikkelen en te gebruiken.
- Dataspecialisten die zicht hebben op de kwaliteit en beschikbaarheid van benodigde data.
- Softwareontwikkelaars, UI/UX designers en architecten die weten hoe de oplossingen gebouwd of geïntegreerd moeten worden.
- Inkopers die ervaring hebben met datgene wat ingeregeld moet worden bij de aankoop van software met AI.

Op basis van artikel 4 van de AI-verordening moeten organisaties die AI-systemen ontwikkelen en/of gebruiken ervoor zorgen dat hun personeel en andere betrokkenen voldoende AI-kennis hebben (‘AI-geletterdheid’). Ze moeten daarbij letten op de technische kennis, ervaring, opleiding en de context waarin de AI-systemen worden gebruikt, evenals de personen of groepen voor wie de AI-systemen bedoeld zijn.

Overweeg het instellen van een ethische commissie binnen uw organisatie. Zorg ervoor dit een multidisciplinaire, diverse groep is waarin voldoende kennis over (generatieve) AI aanwezig is.[8]

Afhankelijk van het type betrokkene zijn kennis en vaardigheden nodig om met generatieve AI te werken. Deze betrokkenen kunnen als volgt gecategoriseerd worden (deze lijst is niet limitatief):

1. Algemeen: elke ambtenaar die niet bekend is met generatieve AI maar hier wel toegang tot krijgt, wordt aangeraden om een

[8] Zie onder andere ook: https://www.ipo.nl/nieuws/provincies-richten-ethische-commissie-op-voor-digitalisering/

basisopleiding te volgen zodat zij de technologie en de risico´s ervan beter begrijpen. Denk hierbij aan trainingen zoals die van RADIO.

2. Beleid en operatie: ambtenaren die generatieve AI gebruiken voor het zoeken van informatie en genereren van documenten. Voor deze groep wordt geadviseerd dat zij altijd kunnen terugvallen op interne of externe opleiding en handreikingen om de tools verantwoordelijk te gebruiken.
3. Ontwikkelaars: ambtenaren met geavanceerde digitale vaardigheden die werken aan de ontwikkeling of implementatie van generatieve AI-oplossingen. Leermogelijkheden kunnen zich richten op de technische en ethische aspecten en implementatie-uitdagingen die gepaard gaan met generatieve AI.
4. Data-analisten: ambtenaren die werken aan het verzamelen, organiseren, analyseren en visualiseren van gegevens. Een training kan gaan over het gebruik van generatieve AI om geautomatiseerde gegevensanalyse, de synthese van complexe informatie en het genereren van voorspellende modellen te vergemakkelijken.
5. Besluitvormers: bestuurders die verantwoordelijk zijn voor het creëren van een overheidscultuur die klaar is voor het gebruik van generatieve AI. Er wordt aangeraden om hen toegang te geven tot informatie die helpt om inzicht te krijgen in degelijke en verantwoorde inzet van generatieve AI en de mogelijke impact ervan op de organisatiecultuur, het bestuur, de ethiek en de strategie.
6. Compliance-medewerkers: medewerkers die zich richten op de (interne) compliance met wettelijke kaders die betrekking hebben op de inzet van generatieve AI. Van hen wordt verwacht dat zij voldoende kennis hebben van de juridische, ethische en organisatorische maatregelen die nodig zijn om generatieve AI verantwoord in te zetten.

4.3 Creëer een (generatieve) AI-governance structuur

Een effectieve AI-governance binnen een organisatie helpt bij de juiste implementatie van AI-modellen. Daarnaast helpt het bij het identificeren, evalueren en beheersen van eventuele risico's. De governance structuur voor generatieve-AI sluit mogelijk goed aan bij al bestaande AI- of data-governance. Een goede AI-governance bestaat bijvoorbeeld uit:

- Een AI-strategie of beleid: houd je aan het overheidsbrede standpunt en vertaal deze handreiking naar een interne versie. Leg dit beleid ter review voor aan I-collega's binnen je organisatie, zoals de CPO, CDO, CISO,CIO, FG en/of AI-Officer.
    - Zorg dat je beleid of strategie informatie verstrekt over de ethische-, juridische-, en veiligheidsaspecten van generatieve AI.
    - Zorg dat het beleid of de richtlijnen ook aanbevelingen voor eindgebruikers bevat, zodat ze de generatieve AI-toepassing zo veilig en effectief mogelijk gebruiken. Voorkom dat medewerkers de generatieve AI-toepassing voor andere doeleinden gebruiken en maak afspraken over het al dan niet opslaan van promptgeschiedenis, voor transparantie en herleidbaarheid. Zorg dat alle eindgebruikers binnen de organisatie weten waarvoor het systeem wel en waarvoor het systeem niet mag worden ingezet. Maak duidelijke afspraken met medewerkers over

het al dan niet opslaan van chat- of opdrachtpromptgeschiedenis.
    - Zorg dat je in je strategie een 'fall back' scenario hebt, waarin generatieve AI niet meer ingezet mag/kan worden.
    - Wanneer generatieve AI wordt gebruikt voor interactie met burgers, zorg dan dat expliciet wordt aangegeven aan de burger dat deze te maken heeft met generatieve AI. Geef burgers ook altijd een optie om te spreken met een natuurlijk persoon bij vragen of onduidelijkheden.

- Een AI-stuurgroep en verantwoording:
    - Wijs verantwoordelijkheden toe binnen de organisatie. Stel bijvoorbeeld een eindverantwoordelijke aan wie het snijvlak tussen cybersecurity en de inzet van generatieve AI coördineert. Deze persoon staat in nauw contact met bijvoorbeeld de CISO (indien aanwezig). Stel ook een eindverantwoordelijke aan voor privacy en generatieve AI, wie samenwerkt met de Chief Privacy Officer (CPO) en/of de FG. Stel een ondersteunende rol aan, bij wie medewerkers terecht kunnen voor vragen over generatieve AI.
    - Verzeker dat je door ontwikkeling/gebruik van de generatieve AI niet discrimineert. Stel een protocol op als discriminatie toch optreedt. Er is een discriminatieprotocol dat hierbij kan helpen.
    - Ga na of je een duurzaamheidsimpactanalyse kan laten doen en zorg dat je uitstoot zo laag mogelijk blijft. Dit kan bijvoorbeeld door gebruik te maken van bestaande voorgetrainde systemen en deze enkel te finetunen, of door voorkeur te geven aan energie-efficiënte data servers.
    - Breng de veiligheidsrisico's van generatieve AI-toepassingen systematisch in kaart. Gebruik hiervoor bijvoorbeeld de OWASP Risk Rating Methodology. Vraag zo nodig advies van een cybersecurityspecialist binnen of buiten de organisatie.
    - Creëer een structuur waar verantwoording voor fouten van de generatieve AI mogelijk is. Maak het mogelijk om fouten of mogelijke discriminatie te rapporteren.

- Een documentatiestrategie: voor overzicht en transparantie buiten de organisatie.
    - Wees transparant over het gebruik van generatieve AI en laat zien waarom en door wie het model wordt ingezet. Zorg dat je dit documenteert, bij voorkeur via het Algoritmeregister. Maak transparant wat de meerwaarde van de ingezette generatieve AI is, bijvoorbeeld met hoeveel procent een uitkomst verbetert, met hoeveel tijd arbeid is ingekort/beter ingezet, of door performance (regelmatig) te evalueren door variërende prompts/versies modellen.
    - Houd een actueel overzicht bij van alle (generatieve) AI-systemen die binnen jouw organisatie worden ingezet (bijvoorbeeld door publicatie in het Algoritmeregister). Zorg dat de beschrijving van het systeem gedetailleerd is, gebruik hiervoor bijvoorbeeld de Publicatiestandaard.

Beschrijf in ieder geval het publieke doel en de beoogde doelgroep. Maak hierbij duidelijk of de interactie gericht is op interne of externe gebruikers.
    - Documenteer overwegingen over toegang en gebruik van data. Beschrijf bijvoorbeeld of het generatieve AI-model (op termijn) verder getraind zal worden en met welke data. Beschrijf of deze data wordt opgeslagen, wie toegang heeft, en hoe de integriteit hiervan wordt gewaarborgd.
    - Verzamel uitgevoerde risicoanalyses als DPIA's en algoritme impact assessments op een centrale plek. Zorg dat deze gemakkelijk teruggevonden kunnen worden, en voor anderen als naslagwerk kunnen dienen om dubbel werk te voorkomen.

4.4 Risicoanalyse(s) uitvoeren

Wanneer er is nagedacht over de doelstelling, de juiste mensen zijn betrokken zijn en er een aanzet is voor een uitgezette generatieve AI-governance structuur, is de volgende stap om een risicoanalyse uit te voeren (zie standpunt generatieve AI). Dit doe je voordat je een generatieve AI-systeem inkoopt of tijdens het zelf ontwikkelen.

- Zorg bij het zelf ontwikkelen van generatieve AI-toepassingen dat je risico's op gebied van privacy, veiligheid en mensenrechten meeneemt in het design. Raadpleeg hiervoor bijvoorbeeld het Algoritmekader.
- Voer de benodigde risicoanalyses uit. Een overzicht van een aantal instrumenten is te vinden in het Algoritmekader. Dit kan bijvoorbeeld gaan om een Impact Assessment Mensenrechten en Algoritmen (IAMA) of AI Impact Assessment (AIIA), maar mag ook een eigen instrument zijn. Het doel van zo'n analyse is de interdisciplinaire discussie tussen verschillende rollen professionals in een organisatie over het gebruik en de risico's van een algoritme aan te wakkeren. Een Data Protection Impact Assessment (DPIA) moet worden uitgevoerd indien dat wordt vereist vanuit de AVG. Een pre-scan DPIA is altijd verplicht als er persoonsgegevens verwerkt worden.[9] Dit geldt voor alle generatieve AI-toepassingen, zowel voor de toepassingen die worden ingezet op projectbasis als voor generatieve AI die door medewerkers structureel ingezet wordt bij dagelijks werk. De uitkomsten van deze assessments dien je bestuurlijk te maken, zoals ook beschreven in het standpunt voor Rijksorganisaties voor het gebruik van generatieve AI. Deze assessments zijn extra van belang bij de inzet van generatieve AI voor hoog-risico toepassingen. De AI-verordening bepaalt de risicoclassificatie van een AI-systeem aan de hand van het gebruik en de inzet. Wordt het systeem ingezet in een hoog-risico toepassing? Zie dan Hoofdstuk 5.1 voor meer informatie over de AI-verordening.
- AI-systemen gebruiken onder consumentenvoorwaarden levert momenteel nog onacceptabele risico's op. Deze vorm van licenties kunnen enkel voor privégebruik of voor verkenning van of experimenten met nieuwe AI-systemen in een gecontroleerde niet-productieomgeving, mits er geen gevoelige of vertrouwelijke data

[9] Tijdens het proces van een DPIA dient de Chief Privacy Office (CPO) meegenomen te worden alvorens de uitkomsten van de assessment worden voorgelegd aan bestuurders. Ook moet de Functionaris Gegevensbescherming (FG) geconsulteerd worden.

ingezet worden. Deze beperking zal mogelijk opgeheven worden als blijkt dat providers gaan voldoen aan de vereisten van de wet- en regelgeving. Gesprekken zijn hierover gaande om dit overheidsbreed op te lossen.

4.5 Generatieve AI inkopen of bouwen

Na de risicoanalyses volgt de inkoop, het of de ingebruikname van de toepassing. Generatieve AI kan op diverse manieren worden toegepast en geïntegreerd. Elke methode heeft unieke voordelen en veiligheidsrisico's, wat invloed heeft op hoe je met deze technologie omgaat. De meest voorkomende toepassingen van generatieve AI zijn:

1. Openbare generatieve AI-toepassingen en webdiensten. Dit zijn kant-en-klare AI-tools die via het internet toegankelijk zijn voor het grote publiek, zoals ChatGPT.
2. Ingebedde generatieve AI-toepassingen. Hierbij wordt AI-functionaliteit geïntegreerd in bestaande software of apparaten, zoals AI-ondersteunde tekstverwerkers. Te denken valt aan Github Copilot die helpt bij het ontwikkelen van code door suggesties te geven.
3. Openbare generatieve AI-API's. Deze interfaces stellen ontwikkelaars in staat om AI-functionaliteit in hun eigen applicaties te integreren zonder de onderliggende technologie te hoeven bouwen.
4. Lokale ontwikkeling. Dit betreft het creëren en trainen van AI-modellen op lokale systemen, wat meer controle biedt maar ook meer middelen vereist.
5. Cloudoplossingen. Deze benadering maakt gebruik van cloud-infrastructuur om AI-modellen te hosten, wat schaalbaarheid en flexibiliteit biedt.

Neem de volgende stappen in acht bij het inkopen of bouwen van generatieve AI:

- Leer van andere overheidsorganisaties die generatieve AI-oplossingen hebben geïmplementeerd.
- Betrek commerciële collega’s vanaf het begin wanneer je generatieve AI-toepassingen inkoopt. Besteed aandacht aan leveranciersmanagement om zeker te zijn van de naleving van juridische kaders. Informatie over inkoopvoorwaarden worden gebundeld in het Algoritmekader.
- Overweeg alle mogelijkheden om *vendor lock-in* te voorkomen. Verken mogelijkheden van Nederlandse of Europese dienstverleners van generatieve AI. Deze voldoen vaak eerder aan onze (Europese) veiligheidsstandaarden dan niet-Europese dienstverleners. Daarnaast versterken we zo de Nederlandse en Europese digitale soevereiniteit.
- Stel duidelijke *requirements* op. Neem hierbij een probleemstelling op, benadruk je datastrategie en maak afspraken over de toegang tot en het gebruik en opslag en logging van gegevens. Dwing daarbij zekerheid af over de manier waarop de dienst gemonitord wordt door de leverancier en zorg voor controle over eigen data.
- Bij voorkeur wordt een open-sourceapplicatie of -model gebruikt.[7] Deze modellen bieden meer inzicht, bijvoorbeeld op het gebied van transparantie. Hierbij geldt wel dat transparantie niet ten koste mag gaan van de veiligheid van generatieve AI-modellen.Verken de

mogelijkheden van *offline* en *on-premise* generatieve AI-modellen. *Offline en/of on-premise* AI-modellen zijn in beginsel niet met het internet verbonden. Hierdoor blijft de generatieve AI-inzet binnen de organisatie en wordt data via het gebruik van een lokale server verwerkt. De kwaliteit van (vaak *open source*) modellen die offline in te zetten zijn, wordt steeds beter en benadert al regelmatig de bekende, grote online modellen. Om dit te realiseren zijn echter aanzienlijke investeringen nodig die onder meer verband houden met de hardware die nodig is.

- Controleer de data-instellingen van het generatieve AI-systeem en zet datadeling voor externe optimalisatie van het model uit. Zo kan worden voorkomen dat de door eindgebruikers ingevoerde data wordt gebruikt om het model mee te trainen. Hiermee voorkom je dat persoonsgegevens worden gedeeld, wat bijvoorbeeld in strijd kan zijn met de AVG.
- Houd bij het selecteren van LLM's rekening met de capaciteiten die verschillende *Foundation Models* bezitten. De Holistic Evaluation of Langue Models van het Stanford Center for Research on Foundation Models (CRFM) levert een goede vergelijking van modellen op verschillende criteria.

5. Juridische aspecten

Dit hoofdstuk bespreekt de belangrijkste regelgeving en juridische kaders die van belang zijn bij de inzet van generatieve AI.

5.1 AI-verordening

De AI-verordening vormt een belangrijk wetgevend kader voor de ontwikkeling en het gebruik van AI in de Europese Unie (EU). De AI-verordening verdeelt AI-systemen onder in verschillende risicocategorieën. Afhankelijk van de categorie waarin een AI-systeem valt, gelden zwaardere of minder zware regels. Met name hoog-risico toepassingen vergen extra afweging wanneer overheidsorganisaties deze in willen zetten. De verordening, die gefaseerd in werking treedt, stelt ook eisen aan generatieve AI (in de AI-verordening valt dit onder "general purpose AI, GPAI").

Overheidspartijen zullen in de praktijk veelal gebruikmaken van GPAI-modellen en/of systemen. Deze kunnen bijvoorbeeld worden ingekocht bij een leverancier en vervolgens voor domein-specifieke toepassingen worden ingezet. Hierbij is het altijd van belang om na te gaan of je als overheidspartij dan aanbieder en/of gebruiksverantwoordelijke bent volgens de AI-verordening. Dit bepaalt namelijk aan welke vereisten je moet voldoen.

5.2 Privacywetgeving

In de ontwikkelfase worden generatieve AI-modellen vaak getraind door middel van '*scraping*' van grote hoeveelheden data van het internet. Het is waarschijnlijk dat hierbij (bijzondere) persoonsgegevens zonder toestemming worden verwerkt. Omdat voor veel generatieve AI-modellen momenteel niet kan worden uitgesloten of persoonsgegevens zijn verwerkt, voldoen deze waarschijnlijk niet aan privacywetgeving[10] zoals de Algemene Verordening Gegevensbescherming (AVG) en de Uitvoeringswet AVG (UAVG). Hier is ten tijde van het opstellen van dit document nog geen rechterlijke uitspraak over gedaan in EU-verband. Hiernaast moeten generatieve AI-modellen die worden ingezet door de politie en justitie zich houden aan de Wet politiegegevens (Wpg) en de Wet justitiële strafvorderlijke gegevens (Wjsg). Dit is geen uitputtende opsomming, per context kan de wetgeving van toepassing verschillen.

Daarnaast is het voor de gebruiker belangrijk om kritisch te zijn op de invoer bij het gebruik van generatieve AI. Als er (met grondslag) persoonsgegevens aan de AI gevoed worden, kan het zo zijn dat deze gegevens bij de ontwikkelaars van het model opgeslagen worden om opnieuw voor trainingsdoeleinden te worden gebruikt. Dit is standaard het geval bij niet-gecontracteerde diensten als *ChatGPT*. Zo kunnen er datalekken ontstaan. Het in eigen beheer opereren van *open source* of zelf ontwikkelde generatieve AI-modellen gaat dit risico tegen maar stelt technische barrières. Bij inkoop van generatieve AI of integratie van generatieve AI in bestaande software (zoals MS *Copilot*) is het van belang inzichtelijk te hebben hoe gebruikersdata wordt gebruikt, opgeslagen en gedeeld, en hier afspraken over te maken met leveranciers.

[10] AP: scraping bijna altijd illegaal | Autoriteit Persoonsgegevens

5.3 Auteursrecht

Aangezien het niet altijd duidelijk te achterhalen is waar de trainingsdata van generatieve AI-modellen vandaan komen, is de kans zeer groot dat deze auteursrechtelijk beschermd materiaal bevatten. Denk daarbij aan kunstwerken, boeken, foto's en publicaties. Een medewerker kan er niet vanuit gaan dat voor alle trainingsdata die voor een generatief AI-model zijn gebruikt toestemming is verkregen. Veel aanbieders zijn momenteel niet transparant over welke auteursrechtelijke werken zij gebruiken in het trainingsproces. De AI-verordening brengt hier verandering in, door het stellen van transparantievereisten aan aanbieders. Denk hierbij bijvoorbeeld aan een gedetailleerde samenvatting van de trainingsdata en informatie over auteurrechtelijke kwesties.

Op dit moment bestaat er nog geen aanpak waarmee getoetst kan worden of er inbreuk wordt gemaakt op het auteursrecht door generatieve AI. Er is weinig jurisprudentie over of, hoe en wanneer generatieve AI-modellen het auteursrecht schenden. Momenteel lopen meerdere rechtszaken om te verhelderen of door generatieve AI-modellen gebruikte trainingsdata rechtmatig verkregen zijn. Vanuit de rechtspraak is er nog geen expliciet oordeel over of, en in welke vorm, *scraping* ten behoeve van het trainen van generatieve AI-modellen auteursrechtelijk gezien toelaatbaar is.

5.4 Inkoop

Door middel van publieke inkoop wordt door overheidsinstellingen software met generatieve AI ingekocht. Voor publieke inkoop gelden algemene en specifieke voorwaarden. Dit kan gaan om kant-en-klare oplossingen, een bestaande oplossing met aanpassingen, of een nieuwe oplossing die ontwikkeld wordt. Bij de aankoop van hoog-risico (generatieve) AI-toepassingen is het aannemelijk dat de opdrachtgever een actievere bijdrage levert aan de samenwerking met de leverancier. De opdrachtgever zal bijvoorbeeld moeten aangeven wat de juridische en ethische grenzen zijn van de uiteindelijk werking van het AI-systeem. Als een kant-en-klare oplossing wordt afgenomen, moet de leverancier aan kunnen tonen dat de ontwikkelde AI voldoet aan alle vereisten. Naast vereisten is het goed om stil te staan bij ethiek, duurzaamheid, en soevereiniteit. Voor meer informatie over inkoop, raadpleeg het Algoritmekader en de Algemene Rijksvoorwaarden bij IT-overeenkomsten (ARBIT). Binnen de NDS zal gezamenlijke aanpak van inkoop van digitaliseringsoplossingen, inclusief generatieve AI, op de agenda staan om onze marktpositie als Nederlandse overheden te versterken en om efficiëntie-voordelen te behalen.

5.5 Overige wet- en regelgeving

Andere wetten waar een organisatie mee in aanraking komt bij het gebruik van generatieve AI zijn bijvoorbeeld de Awb, de WOO, de Archiefwet en de WDO. Deze wetten zijn van toepassing op IT-systemen, en daarmee ook op (generatieve) AI-toepassingen. De vereisten uit deze wetten zijn ook gebundeld in het Algoritmekader.

6. Ethische aspecten

Dit hoofdstuk schetst de belangrijkste waarden en thema's op het gebied van ethiek (kritisch bezinnen op wat wenselijk is) die bij de inzet van generatieve AI meespelen. Dit hoofdstuk focust zich met name op de ethische aspecten die direct komen kijken bij de inzet van generatieve AI, en niet op bredere maatschappelijkere vraagstukken zoals de lange termijnimpact van generatieve AI op de arbeidsmarkt. Hiervoor verwijzen wij graag naar de overheidsbrede visie op generatieve AI.

6.1 Bias en discriminatie

Generatieve AI-modellen, of de data waarmee de modellen worden getraind, kunnen vooroordelen of vooringenomenheden bevatten. Dit wordt *bias* genoemd. Dit kan zitten in de dataverzameling, zoals *selectiebias*. Hier zorgt beperkte trainingsdata voor een onjuiste afspiegeling van de werkelijkheid met als gevolg blinde vlekken. Maar een generatief AI-model heeft vooral als risico dat discriminerende of onwenselijke patronen overgenomen worden van de trainingsdata. Hierom moet bij de ontwikkeling van generatieve AI-modellen trainingsdata kritisch worden onderzocht, en waar mogelijk mitigerende systemen worden ingezet om de modellen heen (ook wel "guard rails" genoemd). Voor modellen zonder guard rails is het verstandig om duidelijke afspraken te maken met eindgebruikers over de manier waarop uitkomsten worden gebruikt. Hiernaast bestaat het risico op *groepsdenken*: een generatieve AI-model kan komen tot consensus zonder alternatieve denkwijzen in overweging te nemen of te tonen. Van belang is als medewerker kritisch te blijven reflecteren en open staan voor niet-conventionele gedachten.

Volgens de Nederlands en Europese wetgeving is *discriminatie* op basis van geslacht, ras, godsdienst of op welke grond dan ook, niet toegestaan. De AI-verordening stelt eisen aan aanbieders, van AI-systemen om discriminatie te voorkomen[11]. Deze eisen gelden ook voor ontwikkelaars van generatieve AI-modellen. Enige mate van bias is echter vaak een gegeven in AI-modellen die gebaseerd zijn op echte data.

Overheidsorganisaties wordt daarom aangeraden goed na te gaan welke anti-discriminatie eisen ze stellen. Hiervoor kan bijvoorbeeld de handreiking non-discriminatie by design geraadpleegd worden. Het is van belang dat resultaten uit generatieve AI-systemen waarvan de inzet mensenrechten raakt tijdig en regelmatig worden getest op ongewenste bias, niet alleen tijdens de ontwikkeling van het model maar ook nadat het in gebruik genomen is. Daarnaast is het verstandig als overheden onderzoeken wat de risico's zijn als het blijkt dat een dergelijk generatief AI-model niet aan deze anti-discriminatie eisen kan voldoen.

6.2 Transparantie

Het is van belang dat er transparantie is over de ontwikkeling en het gebruik van een generatief AI-model, om na te gaan of het model op ethisch verantwoorde wijze is getraind. Bij het trainen van een model dient bijvoorbeeld oog te zijn voor mensenrechten en het risico op onderbetaling van werknemers. In het model zelf kan transparantie worden bevorderd

[11] Ook deze eisen en bijpassende maatregelen zullen worden opgenomen in het Algoritmekader: Algoritmekader - Algoritmekader [concept] (minbzk.github.io)

door open te zijn over de datasets die gebruikt worden om het model mee te trainen en een duidelijke documentatie van het systeem aan te bieden, waaronder informatie over de gebruikte technieken, broncode en parameters. Daarnaast gelden er transparantieregels vanuit de AVG wanneer persoonsgegevens worden verwerkt. Bij ingekochte modellen verdient dit een plek in de inkoopvoorwaarden. Bij het finetunen of zelf ontwikkelen van generatieve AI, kun je dit met technische documentatie bijhouden.

Daarnaast is het belangrijk om transparant te zijn over het gebruik van generatieve AI, zeker als dit invloed heeft op besluitvorming. Personen waarvan hun data wordt verwerkt of die direct in contact staan met het AI-systeem, omdat zij bijvoorbeeld met een *AI-chatbot* interacteren, moeten hiervan op de hoogte worden gebracht, zo volgt ook uit de AI-verordening. Dit kan bijvoorbeeld worden gedaan door een privacyverklaring op te nemen over het verzamelen en gebruik maken van data[12].

6.3 Vertrouwen

Overheidsorganisaties hebben de verantwoordelijkheid om de burger van betrouwbare informatie te voorzien. Wanneer generatieve AI bij dit proces wordt ingezet, kan dit negatieve gevolgen hebben voor de geloofwaardigheid en reputatie van de overheid. Een generatieve AI-systeem weegt niet de betrouwbaarheid van informatie mee bij het beantwoorden van een prompt. Dit schaadt mogelijk het vertrouwen van burgers. Dit risico bestaat ook als er transparant wordt gecommuniceerd over het gebruik van generatieve AI, aangezien men de informatie mogelijk met een kritischere blik tot zich neemt. Het gebruik van generatieve AI richting de burger vraagt om gedegen afweging van de risico's. Sommige risico's kunnen ondervangen worden door een disclaimer dat de output van het generatieve AI-systeem fouten kan bevatten, of door antwoorden te beperken tot enkel verwijzingen naar officiële publicaties. Verder draagt het hebben van een menselijk element dat input levert als onderdeel van het generatieve AI-model (´human-in-the-loop´), bij aan de betrouwbaarheid van de output.

Wanneer AI wordt ingezet bij contentcreatie, zoals het genereren van foto's, video's en *voice-overs*, kan dit ingaan tegen bestaande beleidskaders, zoals dat visuele communicatie altijd een realistische weergave moet zijn van de Nederlandse samenleving en terughoudend dient te worden omgegaan met beeldmanipulatie.13 Het gebruik van dit type 'nepbeelden', in zowel in- als externe communicatie, wordt om deze redenen vooralsnog ten zeerste afgeraden.

6.4 Uitlegbaarheid

Uitlegbaarheid gaat over het kunnen uitleggen van de werking van een algoritme zodat mensen de werking begrijpen. Generatieve AI-modellen zijn *'blackbox'* modellen. Dit betekent dat de uitkomsten van een model niet goed terug beredeneerd kunnen worden, waar dit bijvoorbeeld bij een 'eenvoudig' algoritme als een simpele beslisboom wel volledig kan. De miljarden calculaties die voor iedere output nodig zijn maken menselijk

[12] De vereisten van een privacy verklaring zijn opgesteld door Autoriteit Persoonsgegevens: Recht op informatie | Autoriteit Persoonsgegevens

13 Zie bijvoorbeeld:
https://www.beeldkompas.nl/kennisbank/rijkshuisstijl-beeld

begrip van het generatieve AI-model op het laagste niveau niet mogelijk. Dit is één van de redenen waarom de inzet van generatieve AI niet bruikbaar is voor directe besluitvorming, alleen ter ondersteuning. Een manier om toch (deels) uitleg te kunnen geven over de werking van generatieve AI is uitleg verschaffen over de capaciteiten, beperkingen, en trainingsdata van het AI-model zodat gebruikers weten wat het model wel en niet (goed) kan. Transparantie van nog groter belang bij de inzet van generatieve AI.

Het is daarom aangeraden dat overheidsorganisaties goed nagaan of het gebruik van een *blackbox* model de beste keus is, en of een ander model of algoritme dezelfde oplossing kan bieden. Het is daarnaast verstandig als organisaties onderzoeken wat de risico's zijn als de uitkomsten van een generatief AI-systeem niet uitgelegd kunnen worden.

6.5 Kwaliteit

Hoewel door generatieve AI gegenereerde content er op het eerste gezicht vaak geloofwaardig oogt, geven generatieve AI-modellen niet altijd een feitelijk juist antwoord. Dat komt omdat generatieve AI geoptimaliseerd is voor plausibiliteit, niet voor nauwkeurigheid. Het genereren van output die onwaar is wordt 'confabuleren', of beter bekend 'hallucineren', genoemd. Zo kan een generatief AI-model onbewust een onjuist gegeven presenteren als feit. De kwaliteit van de trainingsdata van een generatief AI-model én de kwaliteit van de prompt beïnvloeden in grote mate de kwaliteit van de output. Zo kunnen de uitkomsten van generatieve AI-modellen een onvolledig beeld geven of niet meer actueel zijn, vanwege incomplete of verouderde trainingsdata.

6.6 Duurzaamheid

Het trainen en gebruiken van generatieve AI-modellen vereist veel rekenkracht, hardware en data, afhankelijk van het type model. Deze benodigde computerkracht en dataservers hebben impact op het milieu. Met name het energieverbruik, de CO_2-uitstoot en het waterverbruik van AI-modellen kunnen hoog oplopen. Ook zijn er schaarse grondstoffen nodig voor het produceren van de hardware. Op dit moment bestaan er nog geen gestandaardiseerde methodes om milieu-impact van generatieve AI te meten en lopen de schattingen uiteen. Veel AI-dienstverleners hebben dan ook geen duidelijke cijfers over de uitstoot van hun modellen.

Nieuwe technieken zoals *quantumcomputing*, efficiëntere architecturen of optimalisatie van bestaande architecturen zouden modellen zuiniger kunnen maken. Tegelijkertijd kunnen we er niet vanuit gaan dat technologische vergroening voldoende zal zijn. Organisaties zullen zelf keuzes moeten maken om de digitale economie te verduurzamen. Het is daarom aangeraden dat organisaties nagaan waarom er voor een generatief AI-model wordt gekozen, en of een duurzamer alternatief dezelfde oplossing kan bieden.

7. Veiligheidsaspecten

Dit hoofdstuk geeft inzicht in de mogelijke veiligheidsrisico's die bij de inzet van generatieve AI kunnen optreden.

7.1 Beschermen van en tegen AI: Cyberaanvallen

Waar generatieve AI enerzijds veel kansen biedt op het terrein van cyberveiligheid, brengt de technologie ook veiligheidsrisico's met zich mee. Zo kunnen generatieve AI-modellen worden aangevallen, of juist worden ingezet als middel om een aanval mee uit te voeren. Binnen de Rijksoverheid wordt er via een apart traject acties en maatregelen uitgewerkt om cyberrisico's van de inzet van AI te beheersen.

Wanneer generatieve AI-systemen onvoldoende robuuste veiligheidsmechanismen bevatten, kunnen deze worden misbruikt door kwaadwillende actoren. Een voorbeeld hiervan is het uitvoeren van *promptinjectie-aanvallen*. Hierbij wordt een generatief AI-systeem gemanipuleerd door middel van een bepaalde prompt of instructie. Aanvallers proberen het model hiermee te dwingen om schadelijke output te produceren, zoals het vrijgeven van vertrouwelijke informatie, het genereren van manipulatieve content of het verspreiden van desinformatie. Om dit soort aanvallen te voorkomen, is het van belang dat generatieve AI-systemen voldoende veiligheidsmechanismen bevatten en regelmatig worden gecontroleerd op mogelijke kwetsbaarheden[14] [15].

Generatieve AI-modellen kunnen tevens worden ingezet om *deepfakes* te creëren. Ook kan generatieve AI helpen bij het efficiënter en op grotere schaal creëren van gepersonaliseerde *phishing* e-mails of *voice* berichten.

7.2 Privacy en datalekken

Zoals beschreven in hoofdstuk 3 kan de training van generatieve AI-modellen leiden tot schending van privacyrechten. Omdat generatieve AI-modellen getraind worden op enorme hoeveelheden data, kunnen deze data persoonsgegevens bevatten, of persoonlijke informatie delen als output[16]. Daarnaast kan het invoeren van vertrouwelijke informatie in een publiek beschikbaar generatief AI-systeem leiden tot datalekken. Dit brengt veiligheidsrisico's met zich mee. Zie ook de publicatie van de EDPS voor advies over informatiebeveiliging en privacy bij het gebruik van generatieve AI. Bovendien kan zelfs het invoeren van niet-gevoelige informatie tóch gevoelig worden, doordat we als overheid maar een beperkt aantal IP-adressen gebruiken en publieke AI-systemen invoerdata kunnen combineren en op deze manier een profiel bouwen.

7.3 Informatiebeveiliging

Voor alle overheidslagen geldt één basisnormenkader: de Baseline Informatiebeveiliging Overheid (BIO). Dit kader stelt informatiebeveiliging standaarden en eist van organisaties dat er een BIO *quickscan* en risico-analysering wordt uitgevoerd. De uitkomsten hiervan dienen besproken te

[14] AI-systemen: ontwikkel ze veilig | Nieuwsbericht | AIVD
[15] De OWASP heeft een lijst opgesteld van de top 10 meest kritieke kwetsbaarheden in generatieve AI-systemen, om hierbij ontwikkelaars en organisaties te informeren over de potentiële veiligheidsrisico's bij het inzetten van deze technologie: OWASP Top 10 for Large Language Model Applications | OWASP Foundation
[16] Bron: Rathenau techscan

worden met degene die verantwoordelijk is voor de geïdentificeerde risico´s. Daarnaast is het mogelijk om de Chief Information Security Officer (CISO) om advies te vragen.

8. Governance

Zoals beschreven in de voorgaande hoofdstukken, kan de inzet van generatieve AI verschillende risico's op het gebied van regelgeving, ethiek en veiligheid met zich meebrengen. Een effectieve AI-governance binnen een organisatie helpt bij het identificeren, evalueren en beheersen van deze risico's. Dit hoofdstuk tracht niet om een blauwdruk te schetsen van een AI-governance. Wel schetst het kort de onderwerpen waar een overheidsorganisatie aan kan denken bij de governance van generatieve AI.

8.1 Verantwoording

Bij de inzet van generatieve AI is het van belang na te denken over het afleggen van verantwoording. Verantwoording afleggen behelst het geven van informatie aan de juiste gremia over de wijze van inzet van generatieve AI en de daarbij genomen risico's, genomen beslissingen, behaalde resultaten en de gespendeerde kosten. Hierbij worden controlemechanismen ingericht om op een juiste wijze en op de juiste plek deze verantwoording af te leggen. Naast (jaar)rapportages zijn het inzetten van interne en externe audits denkbaar.

8.2 RASCI

Om duidelijkheid te krijgen over rollen en verantwoordelijkheden bij de inzet van generatieve AI, kan een RASCI-model helpen. Dit model helpt bij het vastleggen van verantwoordelijkheid (*responsible*) voor het gebruik van generatieve AI, over de hoofdverantwoordelijkheid (*accountable*), over de ondersteuning (*support*), over welke functie wordt geraadpleegd (*consulted*) en welke functie wordt geïnformeerd (*informed*).

8.3 Klokkenluiden, bezwaar en beroep

Gezien de gevoeligheid en mogelijke grote impact die generatieve AI kan hebben op burgers, bedrijven en medewerkers is het van belang om de juiste processen in te richten en/of aan te sluiten[17] voor klokkenluiders en bezwaar en beroep, zie ook Richtlijn 2019/1937. Voor medewerkers moet het mogelijk en duidelijk zijn hoe intern of extern anoniem te rapporteren over negatieve effecten van de inzet van generatieve AI. Voor burgers, bedrijven en medewerkers moet het duidelijk zijn op welke wijze zij bezwaar en beroep kunnen aantekenen naar aanleiding van (directe of indirecte) besluiten naar aanleiding van de inzet van generatieve AI.

[17] Aansluiting kan bijvoorbeeld bij een datalekken procedure of incidentenproces binnen de organisatie.

9. Aanbevelingen voor eindgebruikers

Dit hoofdstuk beschrijft een aantal praktische aanbevelingen voor overheidsorganisaties om aan haar medewerkers te communiceren, om zo het verantwoordelijk gebruik van generatieve AI te bevorderen. De aanbevelingen zijn op CIO offices om richtlijnen voor eindgebruikers op te stellen. CIO-offices kunnen deze aanbevelingen vertalen naar organisatie specifiek context, of toespitsen op een bepaalde generatieve AI-toepassing. De aanbevelingen zijn gecategoriseerd op basis van de onderwerpen uit de voorgaande hoofdstukken.

9.1 Juridische aspecten

- Wees ervan bewust dat de resultaten van generatieve AI mogelijk gebaseerd zijn op auteursrechtelijk beschermde informatie. Neem de door generatieve AI gegenereerde resultaten niet klakkeloos maar gebruik deze ter ondersteuning of als startpunt. Indien je vermoedt dat er een inbreuk wordt gemaakt op auteursrecht, gebruik de uitkomsten van een generatief AI-model dan niet. Neem bij twijfel contact op met een auteursrechtspecialist binnen of buiten je organisatie.
- Het zonder wettelijke grondslag verwerken van persoonsgegevens in een generatief AI-systeem is volgens privacywetgeving (zoals de AVG) onrechtmatig. Deel daarom niet zomaar gevoelige informatie of persoonsgegevens met een AI-model. Ingevoerde informatie kan (onbedoeld) weer in de uitkomsten van het model terugkomen.

9.2 Ethische aspecten

- Denk voordat je ervoor kiest om generatieve AI in te zetten altijd goed na over wat je hiermee wil bereiken en of deze technologie hier het meest geschikte middel voor is. Gebruik generatieve AI niet als uitbesteding voor vaardigheden die je zelf niet bezit of begrijpt.
- Gebruik generatieve AI slechts als hulpmiddel, niet als vervanging. Neem gegenereerde content niet één op één over. Vermenselijk deze technologie niet en formuleer eerst je eigen antwoord.
- Beoordeel de uitkomsten van generatieve AI kritisch op mogelijke vooroordelen.
- Wanneer je gegenereerde content gebruikt in communicatie richting burgers, zorg dan dat de tekst of het beeld gesigneerd of gewaarmerkt is zodat duidelijk is dat dit door een AI-model is gemaakt.

9.3 Tips voor het schrijven van effectieve (en ethisch veilige) prompts

- Experimenteer met verschillende instructies en stel vervolg- en verduidelijkingsvragen. Dit kan helpen om erachter te komen of het model een inhoudelijk juist antwoord geeft. Houd er rekening mee dat LLMs geen echt geheugen hebben binnen een sessie. Zij imiteren dat door (delen van) de inhoud van de vorige prompts en antwoorden achter de schermen mee te nemen in de nieuwe prompt. Om zeker te zijn dat zij de verduidelijking geven, die je zoekt kan je het beste de voor jou belangrijkste teksten meenemen in je prompt als context.
- Wees voorzichtig met het stellen van vooringenomen en suggestieve vragen of opdrachten. Wanneer je een suggestieve vraag stelt aan een generatief AI-model, is de kans groot dat je een bevestigend antwoord ontvangt. Blijf dus altijd kritisch in het

beoordelen van de antwoorden en controleer zo nodig bij een collega of raadpleeg een (wetenschappelijke) bron.

- Wees specifiek en nauwkeurig door middel van gebruikelijke taal en voldoende details. Gebruik geen jargon als dit niet nodig is voor de beantwoording en verdeel een complexe taak op in verschillende simpelere sub-taken. Valideer de (sub-)taak los van elkaar en in context met de resultaten van andere taken.
- Generatieve AI-modellen kennen standaard geen context. Geef dus voldoende achtergrondinformatie mee in je instructies, en gebruik voorbeelden waar mogelijk. Als je een instructie schrijft met een specifiek doel, vermeld dan wat dit doel (en eventueel de doelgroep) is.
- Geef regels mee in je prompt. Geef duidelijk aan wat je verwacht van het generatieve AI-model. Geef bijvoorbeeld aan hoe lang het antwoord moet zijn en in wat voor stijl en format het antwoord gegeven moet worden. Voor tekst generatie kan dit bijvoorbeeld in een stap-voor-stap uitleg, tabel of lijst. Voor beeld generatie kun je een bepaalde stijl of kleureneffect meegeven, en voor geluid generatie een genre of tone-of-voice. Soms kan het helpen om het model een bepaalde rol toe te wijzen in je instructie.
- Om de uitlegbaarheid van output van generatieve AI-systemen te vergroten, kun je ´chain-of-thought prompting´ toepassen. Dit is een wijze waarop instructies stapsgewijs worden geformuleerd om zo betere resultaten uit een taalmodel te verkrijgen. De gebruiker vraagt dan aan het model om een beredenering achter het antwoord.
- Vraag hulp bij ervaren collega’s of experts en verdiep je verder in de (on)mogelijkheden van deze nieuwe technologie. Volg bijvoorbeeld een educatieve training bij RADIO of cursus prompt engineering.

Begrippenlijst

Algoritme Impact Assessment
Een algoritme impact assessment is een hulpmiddel voor het maken van afwegingen bij het inzetten van kunstmatige intelligentie. Voorbeelden van algoritme impact assessments zijn het Impact Assessment Mensenrechten en Algoritmes (IAMA) en het AI Impact Assessment (AIIA).

Artificiële Intelligentie (AI)
Een technologie die, voor expliciete of impliciete doelen, afleidt hoe het op basis van ontvangen input (zoals data), outputs genereert. Outputs betreffen onder meer voorspellingen, content, aanbevelingen en/of beslissingen. De technologie kan leren, redeneren en taken uitvoeren op een manier die normaliter menselijke intelligentie vereist.

AI-verordening
De Europese AI-verordening vormt een van 's werelds eerste uitgebreide wetten specifiek voor kunstmatige intelligentie (AI). De AI-verordening legt kaders en eisen vast voor zowel overheden als bedrijven met betrekking tot de ontwikkeling en het gebruik van AI-systemen.

Chief Information Officer (CIO)
De Chief Information Officer is belast met de ontwikkeling en coördinatie van het informatievoorzienings- en digitaliseringsbeleid en het zorgdragen voor de ontwikkeling en het beheer van de informatiesystemen van de organisatie conform dit beleid.

Data Protection Impact Assessment (DPIA)
Een DPIA is een instrument om vooraf de privacyrisico's van een gegevensverwerking in kaart te brengen, zodat een organisatie maatregelen kan nemen om deze risico's te verkleinen.

Eindgebruikers
Medewerkers die generatieve AI gebruiken in hun werk.

Functionaris Gegevensbescherming (FG)
De Functionaris Gegevensbescherming houdt binnen een organisatie toezicht op de toepassing en naleving van de privacywetgeving.

Generatieve AI
Een vorm van AI waarbij complexe algoritmes worden ingezet om content te generen zoals tekst, afbeeldingen, computercode of video's.

Hoog-risico AI
Hoog risico AI-toepassingen zijn gedefinieerd als toepassingen die een hoog risico voor de gezondheid en veiligheid of de fundamentele rechten van personen inhouden. Om toegelaten te worden op de Europese markt, moeten deze toepassingen aan strenge verplichtingen voldoen.

Large Language model (LLM)
Een gespecialiseerd type generatieve AI-model dat getraind is op grote hoeveelheden tekst om bestaande content te begrijpen en tekstuele content te generen.

(AI-)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. Een model wordt getraind op een taak, bijvoorbeeld het classificeren van documenten of bij een LLM het genereren van zinnen die lijken op de trainingsdata. *GPT-4* is een voorbeeld van een AI-model.

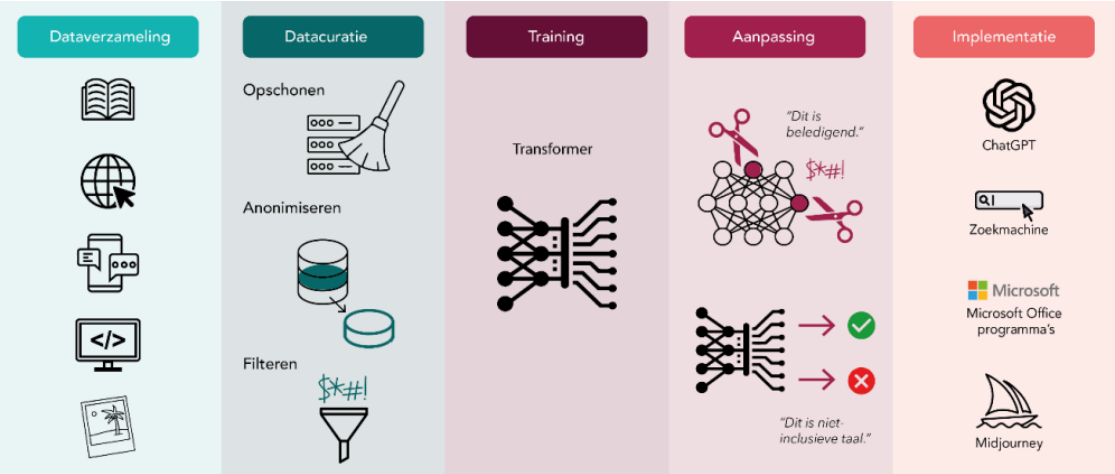
(AI-)Systeem
Een AI-systeem omvat niet alleen het model, maar ook de gehele infrastructuur eromheen. Dit omvat de hardware, software, gegevensverwerking, input- en outputinterfaces, en alle componenten die nodig zijn om het model effectief te laten werken. Voorbeelden van een AI-systemen zijn *ChatGPT, Google Gemini en Co-Pilot*.

Training
Het proces waarbij een AI-model wordt geleerd om patronen in gegevens te herkennen.

Bijlage: Hoe komt een generatief AI-systeem tot stand?

Om een gedegen begrip van een generatief AI-systeem te krijgen, is het van belang om te begrijpen welke fasen in de totstandkoming kunnen worden geïdentificeerd. De ontwikkeling van generatieve AI-systemen is te verdelen in verschillende fases. Het Rathenau Instituut onderscheidt de volgende 5 fases (zie figuur 1)[18]:

Figuur 2: ontwikkelfases van een generatief AI-systeem (bron: Rathenau Instituut[3])



1. Dataverzameling: in de eerste fase worden er grote hoeveelheden data verzameld om het model mee te trainen, zoals (wetenschappelijke) artikelen, boeken, foto's en video's. Deze data zijn vaak openbaar toegankelijk op het internet.
2. Datacuratie: in de tweede fase vindt filtering van de data plaats. Denk hierbij aan het anonimiseren van persoonsgegevens en het verwijderen van duplicaten.
3. Training: deze fase wordt ook wel pre-training genoemd. Tijdens deze fase leert het model om patronen te herkennen in de data. Dit proces vereist aanzienlijke computerkracht en wordt uitgevoerd op gespecialiseerde *hardware*.
4. Aanpassing: in deze fase wordt het model nader verfijnd en aangepast. Zo kan het model worden bijgestuurd met specialistische (organisatie specifieke) kennis en wordt het model mogelijk geoefend om sociaal aanvaardbare antwoorden te geven. Hierbij worden speciale technieken toegepast, zoals *Reinforcement Learning from Human Feedback*[19] (RLHF).
5. Implementatie: in de laatste fase wordt het model beschikbaar gesteld aan gebruikers en toegepast in de praktijk. Het model kan worden gedupliceerd om vervolgens via een consumenteninterface (bijvoorbeeld in de vorm van een chatbot of beeldengenerator) te kunnen worden geraadpleegd door tienduizenden gebruikers tegelijk. De toepassingsvormen van generatieve AI ontwikkelen zich zeer snel. Twee noemenswaardige trends zijn: *Retrieval Augmented*

[18] Generatieve AI | Rathenau Instituut
[19] 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.

Generation (RAG), waarbij het AI-systeem toegang heeft tot een database aan documenten om te kunnen doorzoeken en mee te geven of nemen in het antwoord. *Agentic AI* geeft het AI-systeem ook het vermogen tot zelfstandige handelingen, om bijvoorbeeld bestanden aan te maken. Let op dat nieuwe toepassingsvormen van generatieve AI eigen (mogelijk nog onbekende) risico's met zich meebrengen.