Samen aanjagen van vernieuwing

Handreiking security en privacy van chatbots

Voor een zo verantwoord als mogelijk gebruik

Auteur(s): Werkgroep Handreiking Chatbots – SURF Security Expertise Centrum
Versie: 1.01
Datum: 31 januari 2024
Kenmerk:

Inhoudsopgave

1 Inleiding

In november 2022 werd ChatGPT 3.5 door OpenAI.org gratis ter beschikking gesteld aan elke internetgebruiker. Het gebruik nam meteen een grote vlucht. Niet eerder kreeg een internetdienst in korte tijd zoveel nieuwe gebruikers.[1]

Al snel kwam er ook discussie over de gevolgen van deze en andere op generatieve AI gebaseerde chatbots: zijn die een bedreiging of biedt het juist kansen?
Binnen het onderwijsveld heeft het geleid tot vraagstukken over controle op misbruik bij toetsen en opdrachten, bescherming van auteursrechten, vertrouwelijkheid van (bedrijfs) informatie, de verwerking van persoonsgegevens, mis- en desinformatie maar ook de mogelijkheden om onderwijs te verbeteren en als hulpmiddel voor onderwijsdeelnemers. Daarnaast zijn er veel juridische vraagstukken bijvoorbeeld over de uitwisseling van data buiten de EU en de (on)mogelijkheid om passende verwerkersovereenkomsten af te sluiten.
Ook heeft de Autoriteit Persoonsgegevens vragen gesteld aan Open AI over het gebruik van persoonsgegevens. De EDPB (European Data Protection Board) heeft een specifieke taskforce ingericht voor ChatGPT om samenwerking te bevorderen over mogelijke handhavingsmaatregelen van gegevensbeschermingsautoriteiten binnen de EU[2].

In Europa is wetgeving in voorbereiding over de kaders van toepassing van AI, de AI act. En, zoals vaak, loopt wet- en regelgeving achter bij de realiteit. Afwachten is geen optie en het gebruik van bijvoorbeeld ChatGPT, of daarop gebaseerde applicaties is er al. Veel onderwijsinstellingen hebben daarom in een korte periode verschillende eigen handreikingen en regels gecommuniceerd.

Maatwerk en specifieke behoeften vanuit het brede onderwijsveld zal de aandacht blijven houden en verwacht wordt dat ook de behoeftevoorziening binnen de verschillende onderwijsvormen verschilt. Zo zal in het basisonderwijs andere vraagstukken spelen als binnen het hoger of middelbaar (beroeps) onderwijs. Dit zal verdere verdieping en maatwerk vragen.

1.1 Wat is het doel van deze handreiking?

Met dit document proberen we de al aanwezige informatie over de security en privacy aspecten van chatbots te structureren en toegankelijk te maken. Hierbij zijn we gezien de aard en omvang van de informatie genoodzaakt dit te beperken tot de hoofdlijn. En komen op basis hiervan met medewerk tips voor een zo verantwoord als mogelijk gebruik[3]. Tot de doelgroep behoren de security en privacy professionals en beleidsmakers binnen de instellingen.

1.2 Wat is deze handreiking niet?

Deze handreiking is geen beschrijving van gedeeld beleid, ook is het geen advies richting instellingen. Daarnaast is er op veel vragen nog geen concreet, of volledig uitgewerkt juridisch antwoord te geven. Het begrip (generatieve) AI vraagt altijd om specifieke en context gebonden beoordeling van de toepassing. Deze beoordeling is aan de instellingen zelf om te maken.

[1] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[2] Autoriteit Persoonsgegevens: https://autoriteitpersoonsgegevens.nl/actueel/ap-vraagt-om-opheldering-over-chatgpt

[3] SURF publicatie responsible tech: https://www.surf.nl/nieuws/responsible-tech-zo-zorgen-we-dat-nieuwe-technologieen-voldoen-aan-publieke-waarden

Bestaande regelgeving zoals de AVG blijft op alle vormen van AI waarin persoonsgegevens worden verwerkt van toepassing.

2 Wat is generatieve AI?

2.1 Beknopte uitleg generatieve AI / LLM in het algemeen

Generatieve AI (Artificial Intelligence) is een vorm van kunstmatige intelligentie die in staat is om op basis van input zoals tekst, afbeeldingen of andere gegevens 'nieuwe' output te genereren. Het maakt gebruik van diepe neurale netwerken en machine learning-algoritmen om contextueel relevante output te produceren.

AI is historisch lastig geweest om te definiëren.[4] De OESO definieert het als volgt: "Een AI-systeem is een machine-gebaseerd systeem dat, voor expliciete of impliciete doelstellingen, uit de input die het ontvangt afleidt hoe outputs moeten worden gegenereerd, zoals voorspellingen, inhoud, aanbevelingen of beslissingen die fysieke of virtuele omgevingen kunnen beïnvloeden. Verschillende AI-systemen variëren in hun mate van autonomie en aanpassingsvermogen na implementatie".[5]

Large Language Models (LLM) zijn een vorm van generatieve AI specifiek voor tekst die worden getraind op enorme hoeveelheden tekst (vandaar de term large). GPT-4 is een LLM (eigenlijk LMM - large multimodal model want het verwerkt zowel taal als afbeeldingen).
Ze kunnen worden ingezet voor een breed scala aan taken, variërend van vertalingen en samenvattingen tot het schrijven van artikelen en het beantwoorden van vragen.

2.2 Chatbots zoals ChatGPT

Chatbots zoals ChatGPT zijn applicaties of systemen die gebruikmaken van generatieve AI en LLMs om interactieve gesprekken met gebruikers te voeren. Ze kunnen vragen beantwoorden, informatie verstrekken, taken uitvoeren en mensachtige gesprekken voeren via tekst- of spraakinterfaces.

2.2.1 ChatGPT

ChatGPT is een chatbot van OpenAI, gebaseerd op de GPT3.5 of GPT4 modellen. Dit model bestaat uit een neuraal netwerk dat getraind is met een enorme hoeveelheid tekst, waarvan grotendeels onduidelijk is welke tekst en waar deze vandaan komt. Het model is in staat om op basis van deze data verschillende vormen van response (tekst, beeld etc.) te geven op een vraag of opdracht van de gebruiker. Je kunt ChatGPT bijvoorbeeld creatieve opdrachten geven, zoals het schrijven van een essay over een door jou aangedragen onderwerp.

2.2.2 Andere toepassingen van generatieve AI

ChatGPT is op dit moment een van de meest bekende en gebruikte generatieve AI-toepassing maar er zijn er veel meer. Leveranciers zoals Microsoft en Google bieden steeds vaker chatfunctionaliteit aan in zoekmachines en experimenteren met integratie in hun applicaties (bijvoorbeeld MS Office). Ook zijn er al toepassingen met image-to-tekst: het automatisch beschrijven van wat er op een beeld te zien is. Andersom: van tekst-to-image (DALL-E en Midjourney) zien we ook steeds nieuwe ontwikkelingen. We zien nu al toepassingen waarin eenzelfde model tekst en beeld tegelijkertijd kan verwerken. Ook werken onder meer tekst naar video en spraak steeds beter en kunnen ook programmeringscodes gegenereerd worden[6].

[4] Zie Bijlage 1 Definities van AI

[5] https://oecd.ai/en/wonk/ai-system-definition-update

[6] SURF Tech Trendsrapport: https://www.surf.nl/files/2023-02/sf_trendrapport_v10_compressed.pdf

Daarbij wordt het steeds moeilijker te beoordelen of deze uitingen van een persoon of AI afkomstig zijn.

3 Wat zijn de relevante kaders voor het gebruik van generatieve AI?[7]

Met betrekking tot het gebruik van generatieve AI zijn er veel relevante kaders van toepassing. Ook is er nog veel in ontwikkeling en is de impact van nieuwe wet- en regelgeving nog niet helemaal duidelijk. Denk bijvoorbeeld aan NIS2. Om deze reden is in dit hoofdstuk met name de focus gelegd op de AVG.

3.1 Wet- en regelgeving

Aanknopingspunten voor de regulering van toepassingen die zijn gebaseerd op AI en algoritmes met een ‘zelfdenkend’ of ‘zelflerend’ vermogen zoals bij een Large Language Model (LLM) het geval is, kunnen gevonden worden in de datasets waar deze toepassingen zich op berusten (‘'Big Data') en in de slimme analyses van deze toepassingen (e.g. ‘data mining’, ‘machine learning’, ‘neurale netwerken’).

Zo krijgen organisaties op het gebied van data te maken met een heel nieuw scala aan verordeningen, naast de huidige AVG.

3.1.1 Europese Digitale Strategie

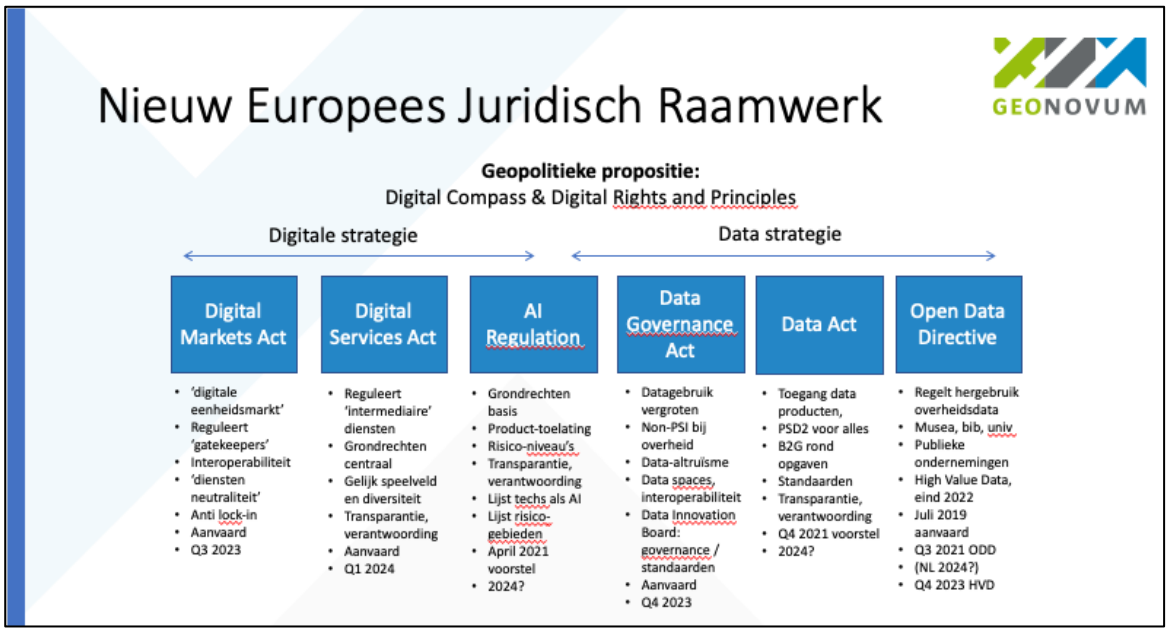
De Europese Digitale Strategie is ontwikkeld om de Europese positie op digitaal gebied te versterken. Als onderdeel van deze digitale strategie zijn drie juridische instrumenten door de Europese Commissie voorgesteld:

1. de Digital Markets Act (DMA);
2. de Digital Services Act (DSA);
3. de AI Act (gericht op de toelating van AI producten).

De Europese Data Strategie streeft naar een eenheidsmarkt voor de beschikbaarheid en het gebruik van data. Binnen deze strategie spelen de genoemde drie juridische instrumenten een rol. Zie in de figuur 1 hieronder een overzicht van het nieuw Europees Juridisch Raamwerk. [8]

[7] https://communities.surf.nl/ai-in-education/artikel/de-ai-act-in-vogelvlucht

[8] Handreiking EU Informatie m.b.t. digitale en data-strategie: https://geonovum.github.io/eu_regelingen_datastrategie/#europese-digitale-strategie



Figuur 1: Europees Juridisch Raamwerk

Hieronder worden de onderdelen van de AVG en de AI Act die van toepassing zijn voor het gebruik van generatieve AI toegelicht, omdat deze verordeningen momenteel het meest actueel zijn ten aanzien van het gebruik van 'foundation models'.

3.1.2 Algemene verordening gegevensbescherming (AVG)

De Algemene verordening gegevensbescherming (AVG) is wetgeving op het gebied van gegevensbescherming en heeft tot doel de bescherming van de grondrechten en vrijheden van personen van wie de persoonsgegevens wordt verwerkt. De AVG ziet dus met name toe op de bescherming van de privacy van individuen wier persoonsgegevens verwerkt wordt door organisaties. De datasets op basis waarvan een AI-toepassing slimme analyses uitvoert bevatten vaak persoonsgegevens waardoor de AVG van toepassing is op de verwerking van deze gegevens door AI-toepassingen in zowel de verzameling van dergelijke gegevens als op de inhoudelijke analyse hiervan. De basisprincipes van de AVG zullen hierdoor in ogenschouw genomen moeten worden wanneer er gewerkt wordt met AI-toepassingen die gebruik maken van persoonsgegevens.

De basisprincipes van de AVG schrijven voor dat:

- bij de verwerking van persoonsgegevens een grondslag moet bestaan ('rechtmatigheid');
- de verwerking van persoonsgegevens noodzakelijk is voor een vastgesteld doel en dat persoonsgegevens in principe niet verder verwerkt mogen worden voor een ander doel waarvoor ze in eerste instantie zijn verkregen of opgehaald. Dit kan het geval zijn wanneer het andere doel niet verenigbaar is met het oorspronkelijke doel ('doelbinding');
- er transparantie wordt geboden in onder meer de persoonsgegevens die worden verwerkt, de doelen waarvoor ze worden verwerkt en in de wijze waarop ze worden verwerkt ('transparantie');
- alleen persoonsgegevens worden verwerkt die noodzakelijk en relevant zijn om een doel te bereiken ('dataminimalisatie');

- persoonsgegevens accuraat en actueel zijn in de context waarvoor deze gegevens worden gebruikt (‘accuraatheid’);
- persoonsgegevens niet langer worden bewaard dan dat ze nodig zijn voor het beoogde doel (‘opslagbeperking’); en
- dat de persoonsgegevens adequaat worden beveiligd met technische en organisatorische waarborgen tegen inbreuken met betrekking tot de beschikbaarheid, integriteit en vertrouwelijkheid van deze gegevens.

In zowel de ontwikkeling als het gebruik van AI-systemen worden deze systemen getraind op basis van Big Data. Big Data kenmerkt zich met een drietal aspecten die mogelijke uitdagingen met zich meebrengen in relatie tot de bovengenoemde basisprincipes van de AVG. Big Data kenmerkt zich namelijk door (i) de omvangrijke volume van de datasets, (ii) de hoge snelheid waarmee de datasets verwerkt worden en (iii) de verschillende variaties waarin data in deze datasets voorkomen. Vaak zijn deze gegevens ongestructureerd van aard.
De kenmerken van Big Data in relatie tot de ontwikkeling en gebruik van AI-systemen kunnen op gespannen voet staan met de bovengenoemde principes uit de AVG. Hieronder wordt in het geval van OpenAI’s ChatGPT geschetst hoe deze spanning zich in de praktijk vertaalt.

Een chatbot zoals die van OpenAI kent enerzijds een fysieke component in de vorm van hardware die het mogelijk maakt voor OpenAI om de neurale netwerken in de ‘generative pre-trained transformers’ (GPT’s) te vormen op basis van Big Data. Anderzijds kent de chatbot van OpenAI een softwarecomponent waarbij de neurale netwerken getraind worden in twee fases. De eerste "pretraining"-fase houdt in dat er op basis van beschikbare informatie op het internet, initiële parameters worden ingesteld waarna deze parameters worden bijgesteld tot het gewenste resultaat in de “fine-tuning"-fase.

De datasets die worden gebruikt in de pretraining-fase zijn grotendeels samengesteld uit (ongestructureerde) informatie die publiekelijk beschikbaar zijn op het internet, zoals websites, boeken en andere bronnen.[9] Deze datasets die grootschalig zijn ‘gescraped’, kunnen persoonsgegevens bevatten die dus ook mede gebruikt worden bij het aanleren van chatbots over de structuur van taal in algemene zin ofwel over ‘Natural language processing’-taken (NLP). Door de ongestructureerde aard van gegevens in de verzamelde datasets en de wijze waarop chatbots tot hun output komen, bestaat het risico dat de output onvolledig, gedateerd of onjuist is.

OpenAI berustte zich bij de verwerkingen voor haar GPT-modellen eerder op de gebruikersovereenkomst als grondslag voor de verwerking van persoonsgegevens, maar werd in het voorjaar van 2023 teruggefloten op deze keuze door de Italiaanse toezichthouder die OpenAI sommeerde zich te beroepen op het gerechtvaardigd belang die OpenAI zou hebben of op de expliciete toestemming van gebruikers.[10]
Daar waar het verkrijgen van expliciete toestemming van elk individu voor het gebruik van dergelijke gegevens bij het trainen van het model praktisch onmogelijk zou zijn voor OpenAI, berust OpenAI zich nu op het gerechtvaardigd belang die het zou hebben voor de verwerking van persoonsgegevens in haar model.[11]

[9] "Language Models are Few-Shot Learners", Brown et al, Johns Hopkins University, OpenAI, 28 mei 2020, p. 9.

[10] https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9874751#english

[11] OpenAI: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

De beoordeling van dit gerechtvaardigd belang vereist een driestapstoets: (I) het belang moet gerechtvaardigd zijn, (II) de verwerking van de persoonsgegevens moet noodzakelijk zijn om dit belang te dienen, (III) en de belangen van de organisatie moeten zwaarder wegen dan die van de betrokkenen van wie gegevens worden verwerkt. Hoewel OpenAI een gerechtvaardigd belang kan hebben bij haar algoritmische training activiteiten, zal het ook moeten stilstaan bij de rechten van betrokkenen en hun redelijke verwachtingen van privacy als onderdeel van de driestapstoets.[12] Kan een individu in deze context en als voorbeeld redelijkerwijs verwachten dat diens gegevens opgenomen wordt in een omvangrijke trainingsdataset doordat zij of hij posts op sociale media plaatst die openbaar toegankelijk zijn? Als het antwoord op deze vraag ontkennend is doordat er een gebrek is aan transparantie en informatieverstrekking aan betrokkenen in relatie tot de verwerking van diens persoonsgegevens, dan kan de rechtmatigheid van deze verwerking niet gegarandeerd worden.

Bovendien kan het ook spaak lopen op het gebied van rechtmatigheid, zodra het verdere gebruik van persoonsgegevens – in zowel de pretraining-fase als in de fine-tuning-fase – berusten op de grondslag toestemming en het verdere gebruik niet verenigbaar is met het doel waar deze persoonsgegevens in eerste instantie voor zijn verkregen. Met andere woorden kan er in deze gevallen geen sprake zijn van doelbinding. Hoewel bij het gebruik van persoonsgegevens die door betrokkenen zelf openbaar zijn gemaakt minder snel sprake kan zijn van een gebrek aan doelbinding in de pretraining-fase, kan dit anders zijn bij de fine-tuning-fase van het trainen van de neurale netwerken van OpenAI's chatbot.

Op dit moment staat de instelling die het mogelijk maakt voor OpenAI om diens modellen verder te trainen en te optimaliseren op basis van de input ofwel "prompts" van gebruikers, standaard aan. Dit betekent dus dat een gebruiker mogelijk gevoelige of persoonsgebonden informatie kan delen met OpenAI met het oog op een gewenste output, maar dat deze gegevens óók gebruikt kunnen worden door OpenAI zelf om diens chatbot verder te trainen. Dit laatste gebruik kan onverenigbaar zijn met het oorspronkelijke doel waarvoor de gebruiker deze informatie in eerste instantie heeft gedeeld met de chatbot. Strikt genomen zou een gebruiker hier toestemming voor hebben gegeven door het accepteren van de gebruikersvoorwaarden, maar de aanwezigheid van de mogelijkheid om de instelling voor deze verwerking uit te zetten ('opt-out') veronderstelt een mate van onverenigbaarheid van dit doeleinde door de gebruiker de keuze te geven om deze instelling expliciet aan of uit te zetten. Doordat deze instelling standaard aanstaat, kan ook bepleit worden dat de privacy van de gebruiker niet per definitie en in het ontwerp is meegenomen ('privacy-by-default' & 'privacy-by-design').

Overigens kan de effectiviteit van een opt-out-mechanisme in twijfel worden getrokken. Als een persoon bijvoorbeeld niet wil dat haar of zijn persoonsgegevens gebruikt worden voor trainingsdoeleinden en hiervoor de instelling hiertoe uitzet, hoeft niet te betekenen dat haar of zijn persoonsgegevens niet worden meegenomen voor trainingsdoeleinden wanneer een andere persoon geen gebruik maakt van de opt-out keuze en wél persoonsgegevens over anderen invoert in de chatbot van OpenAI.

De bovenstaande punten – die vooral in het licht van OpenAI's ChatGPT zijn toegelicht – laten zien dat het gebruik van LLM-gebaseerde chatbots niet per definitie stroken met de principes uit de AVG. Andere aspecten van de AVG die relevant zijn om in ogenschouw te nemen bij de beoordeling van een LLM-gebaseerde chatbot zijn onder andere de wijze waarop betrokkenen

[12] Overweging 47 AVG

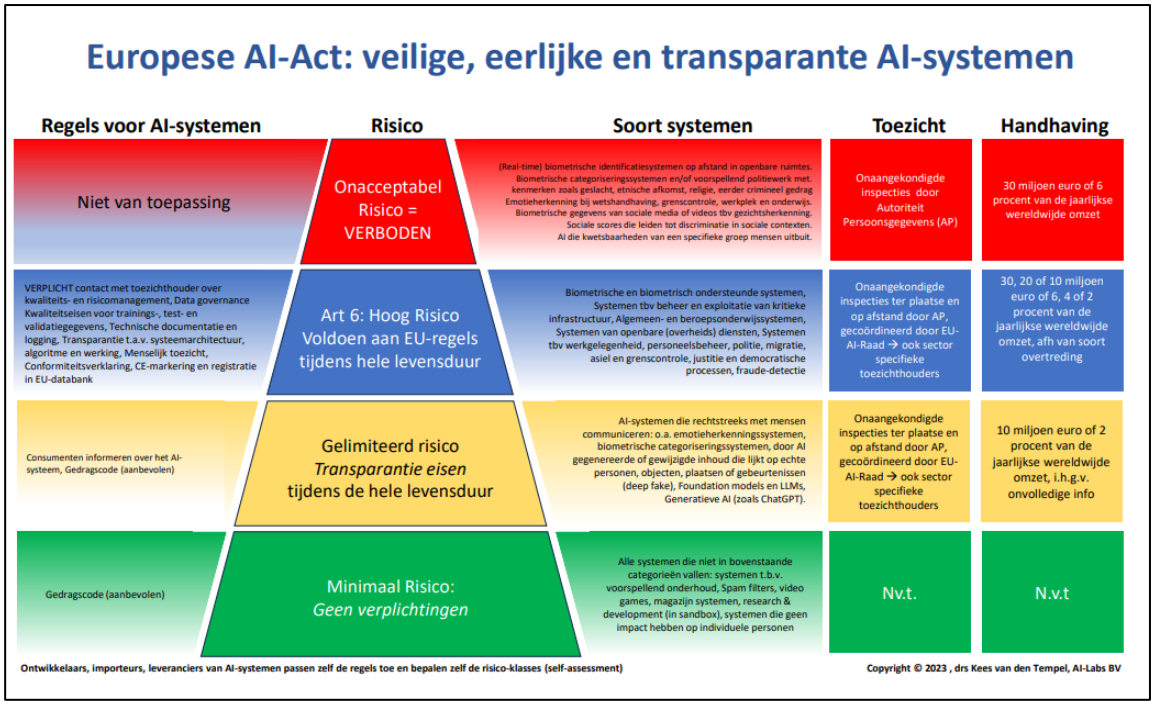
hun rechten kunnen uitoefenen, of de juiste overeenkomsten voor de verwerking van persoonsgegevens zijn gesloten, en wat de impact kan zijn op betrokkenen als er sprake is van geautomatiseerde besluitvorming bij de implementatie van dergelijke technieken middels API's binnen het eigen IT-landschap. Ten aanzien van dit laatste punt is de verwachting dat de AI Verordening meer betekenis kan bieden in de waarborging van mensenrechten bij het gebruik van AI-systemen met een vooraf gedefinieerde risicoclassificatie.

3.1.3 AI Act[13]

De AI Act is een voorstel voor een regulering van AI binnen de EU, zodat AI-toepassingen veilig op de markt kunnen worden gebracht en er rekening wordt gehouden met fundamentele mensenrechten. De AI Act beoogt AI-toepassingen te reguleren vanuit een risico gebaseerd perspectief. Daarnaast heeft de AI Act verplichtingen opgenomen per rol. Een partij kan een aanbieder zijn, maar ook een importeur, een gebruiker of een derde partij. De AI Act classificeert het gebruik van AI in 3 categorieën met bijbehorende maatregelen:

- Onaanvaardbaar risico AI;
- AI met een hoog risico;
- AI met een beperkt risico.

Zie in de figuur 2 hieronder een overzicht van deze classificaties.



Figuur 2: Europese AI-Act: veilige, eerlijke en transparante AI-systemen[14]

3.1.3.1 AI-systemen met een onaanvaardbaar risico

AI-systemen met een onaanvaardbaar risico zijn systemen die als een gevaar voor de mens worden beschouwd. Deze worden in de wet genoemd en zullen worden verboden. Het gaat hier onder meer om:

[13] Onderstaande is gebaseerd op een nog niet goedgekeurde verordening en onderhevig aan wijzigingen.

[14] De AI Act in vogelvlucht!: https://communities.surf.nl/ai-in-education/artikel/de-ai-act-in-vogelvlucht

1. *cognitieve gedragsmanipulatie* van personen of specifieke kwetsbare groepen: bijvoorbeeld speelgoed dat reageert op een stem en dat bij kinderen gevaarlijk gedrag uitlokt;
2. *sociale scoring:* het classificeren van mensen op basis van hun gedrag, sociaal-economische status of persoonlijke kenmerken;
3. *biometrische systemen* voor identificatie op afstand in real time, zoals gezichtsherkenning.

Hierbij is ook ruimte voor uitzonderingen. Zo zijn bijvoorbeeld systemen voor identificatie op afstand "achteraf", waarbij de identificatie met een aanzienlijke vertraging plaatsvindt, wel toegestaan bij vervolging van ernstige misdrijven. Hiervoor is toestemming van de rechter vereist.

3.1.3.2 AI-systemen met een hoog risico

AI-systemen die een negatief effect hebben op de veiligheid of de grondrechten worden als systemen met een hoog risico beschouwd en worden onderverdeeld in twee categorieën:

1. AI-toepassingen in producten die onder de EU-wetgeving op het gebied van productveiligheid vallen. Deze toepassingen zijn onderworpen aan specifieke regulering. Voorbeelden hiervan zijn speelgoed, luchtvaartuigen, auto's, medische hulpmiddelen en liften.
2. AI-systemen die significante risico's voor mens of maatschappij geeft op acht verschillende gebieden, die in een EU-databank moeten worden geregistreerd:
    a. biometrische identificatie en categorisering van natuurlijke personen;
    b. beheer en exploitatie van kritieke infrastructuur;
    c. onderwijs en beroepsopleiding;
    d. werkgelegenheid, personeelsbeheer en toegang tot zelfstandige arbeid;
    e. toegang tot en gebruik van essentiële particuliere diensten en openbare diensten en uitkeringen;
    f. rechtshandhaving;
    g. migratie, asiel en beheer van grenscontroles;
    h. ondersteuning van rechterlijke instanties bij het uitleggen van feiten en toepassing van het recht.

Alle hoog risico AI-systemen gaan aan een flink aantal eisen moeten voldoen, bijvoorbeeld op gebied van privacy, beveiliging, documentatie, registratie, gegevenskwaliteit, logging, menselijke tussenkomst en doorlopende monitoring van de effecten van het systemen.
De verantwoordelijkheid om hieraan te voldoen zal primair bij de ontwikkelaars liggen.
Bij de tweede categorie AI systemen moet de aanbieder, importeur of gebruiker van het systeem een inschatting maken of de inzet van AI aanleiding geeft tot significante risico's. Als de uitkomst is dat die risico's er niet zijn, moet hiervan melding worden gemaakt bij de toezichthouder en is de AI-toepassing geen hoog risico toepassing. Op het doen van bewust onjuiste meldingen staan boetes. Alle AI-systemen met een hoog risico krijgen een beoordeling voor zij op de markt worden gebracht en worden tijdens de volledige levensduur van het systeem gevolgd.

In recital 35 van de AI Act wordt een en ander toegelicht over het gebruik van AI-toepassingen in onderwijs: AI-systemen die in het onderwijs of voor beroepsopleidingen worden gebruikt, in het bijzonder voor het verlenen van toegang of het toewijzen van personen aan instellingen voor onderwijs of beroepsopleidingen of voor het evalueren van personen aan de hand van tests of

als voorwaarde voor hun onderwijs, moeten als systemen met een hoog risico worden beschouwd, aangezien zij bepalend kunnen zijn voor het onderwijs en de carrière van personen en derhalve voor hun vermogen om in hun levensonderhoud te voorzien. Wanneer dergelijke systemen op ondeugdelijke wijze zijn ontworpen en worden gebruikt, kunnen zij in strijd zijn met het recht op onderwijs en opleiding, alsook met het recht niet te worden gediscrimineerd, en kunnen historische patronen van discriminatie in stand worden gehouden.

3.1.3.3 AI-systemen met een beperkt risico

AI-systemen met een beperkt risico moeten voldoen aan minimale transparantievereisten om gebruikers in staat te stellen weloverwogen beslissingen te nemen. Na interactie met een toepassing is het dan aan de gebruikers om te beslissen of zij de toepassing willen blijven gebruiken. Het moet aan de gebruikers duidelijk worden gemaakt dat zij interageren met een AI-systeem. Dit geldt onder meer voor AI-systemen die beeld-, audio- of videomateriaal genereren, zoals deepfakes.

3.1.3.4 General Purpose AI, Generatieve AI and foundation models

Hoe Generatieve AI en zgn foundation models wordt opgenomen in de AI Act is nog onzeker. De verwachting is dat de onderhandelingen in december 2023 worden afgerond, en dan zal duidelijk worden hoe foundation models moeten worden ingeschaald. Vooralsnog lijkt het erop dat de volgende spelregels voor deze groep zal gaan gelden:

- Foundation models gedefinieerd als AI-toepassingen die in staat zijn om verschillende taken vakkundig uit te voeren;
- Een gelaagde aanpak met onderscheid tussen gewone, grootschalige en ‘high impact’ models;
- De aanbieders van deze models moeten kunnen aantonen dat er maatregelen zijn genomen om auteursrechtschendingen tegen te gaan;
- Een AI-bureau dat toezicht houdt (niet per lidstaat maar Europees).

De algemene toepassingsmogelijkheden verhouden zich in beperkte mate tot de op concrete systemen en specifieke toepassingen gerichte basis van de risicoclassificatie van de AI Act.

In de huidige voorstellen zal Generatieve AI, zoals ChatGPT, minimaal moeten voldoen aan een aantal transparantievereisten:

- De toepassingen moeten de vermelding bevatten dat de inhoud is gegenereerd door AI;
- Het model moet zodanig worden ontworpen dat er geen illegale inhoud kan worden gegenereerd;
- Er moet een overzicht worden gegeven van auteursrechtelijk beschermde gegevens die voor het trainen van het systeem zijn gebruikt.

De rijksoverheid werkt momenteel aan een visie op generatieve AI. Momenteel worden er in de samenleving visies en meningen opgehaald. Naar verwachting wordt de definitieve visie begin 2024 gepubliceerd. [15]

[15] Verder op weg naar een visie op generatieve AI: https://www.rijksoverheid.nl/actueel/nieuws/2023/10/11/verder-op-weg-naar-een-visie-op-generatieve-ai

3.1.4 Intellectueel eigendom en copyright

Naast specifieke regulering zoals de AVG en de aankomende AI verordening, zijn er meer juridische uitdagingen die met generatieve AI komen. Bijvoorbeeld vraagstukken rondom intellectueel eigendom en copyright.
Dit geldt ook voor bedrijfsgevoelige informatie. Zo hebben medewerkers van Samsung een onopzettelijk lek ontdekt van gevoelige interne broncode die een ingenieur in ChatGPT had geüpload.[16] Ook Amazon heeft eerder ondervonden dat er ChatGPT teksten genereerde die erg leken op interne bedrijfsgegevens.[17] Daarom is het een punt van aandacht dat gescrapete of geüploade (persoons)gegevens letterlijk terug kunnen komen in de output, wat in theorie een datalek oplevert. Hier is ook al een voorbeeld van, waarbij persoonsgegevens uit ChatGPT gehaald worden.[18]

Daarnaast is de toegang tot gebruikersaccount onvoldoende beschermd door bijvoorbeeld twee-factor-authenticatie bij zowel de gratis als de betaalde 'Plus' variant van ChatGPT waardoor een handel op het dark web is ontstaan in gehackte ChatGPT-accounts. Een gehackte account kan veel (persoons)gegevens vrijgeven, omdat de interactiegeschiedenis met ChatGPT 'by default' bewaard wordt.[19] Om dergelijke (bedrijfs)risico's tegen te gaan, heeft OpenAI ook een betaalde 'Enterprise' variant van haar chatbot uitgebracht waarbij de prompts van gebruikers niet worden gebruikt voor het trainen van haar modellen en de beveiliging van accounts beter zijn geborgd middels onder andere 'Single Sign On' en een centraal beheerdersportaal.[20]

[16] https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

[17] https://futurism.com/the-byte/amazon-begs-employees-chatgpt

[18] https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html?ref=404media.co

[19] https://sea.mashable.com/tech/24452/the-dark-web-is-overflowing-with-stolen-chatgpt-accounts

[20] https://openai.com/blog/introducing-chatgpt-enterprise

4 Analyse van risico's

Zoals bij ieder proces waarbij (persoons-)gegevens betrokken zijn, is een risicoanalyse een essentiële stap om grip te houden op gegevens die worden verwerkt. Deze paragraaf heeft als doel instellingen te helpen een beeld te vormen op welke wijze deze grip verloren kan gaan, zodat het eenvoudiger wordt om tot een op maat gemaakte risicoanalyse te komen. Hiertoe wordt middels een aantal categorieën een overzicht gegeven van de dreigingen, inherente zwakheden en bijbehorende consequenties gerelateerd aan het gebruik van AI en/of de wijze waarop deze wordt aangeboden.

Het type gegevens dat via AI wordt verwerkt en ook de eisen die aan deze gegevens worden gesteld, zullen per instelling en zelfs per proces verschillen. Bij onderstaande beschrijvingen zal dus door de instelling zelf een nadere invulling moeten worden gegeven aan de impact en het uiteindelijke risico.

De risico's aangaande chatbots kunnen worden verdeeld in een viertal categorieën; risico's ten aanzien van de:

1. Beschikbaarheid van de functionaliteit;
2. Integriteit (juistheid, volledigheid en tijdigheid);
3. Vertrouwelijkheid ten aanzien van de gedeelde gegevens;
4. Compliance, aan met name de AVG en andere relevante verordeningen behorend aan de Europese Digitale en Data Strategie, zoals de AI-verordening.

1. Voorbeelden van risico's ten aanzien van **beschikbaarheid** zijn:

- Er kan nieuwe regelgeving van kracht worden waardoor leveranciers geen AI-diensten meer aanbieden. Als organisaties in haar onderwijs- en ondersteunende processen (via applicaties) zijn gaan steunen op alleen deze diensten, raken kritieke processen verstoord.
- Rechtszaken over de geldigheid van de verwerkingsgrondslag van trainingsdata kunnen ertoe leiden dat leveranciers hun AI-diensten meer aanbieden, met dezelfde mogelijke consequenties als hierboven beschreven.

2. Voorbeelden van risico's ten aanzien van **integriteit** zijn:

- Chatbots werken op basis van verzamelde en gecomprimeerde data, waarbij de bron verloren kan gaan. Hierdoor kan niet of moeizaam worden vastgesteld of output juist, volledig en/of actueel is. Bij verkeerd gebruik kan deze onbetrouwbare informatie leiden tot onjuiste beleidskeuzes, verlies van accreditatie of verlies van reputatie door verstoorde onderwijsprocessen.

3. Voorbeelden van risico's ten aanzien van **vertrouwelijkheid** zijn:

- Chatbots kunnen ingevoerde gegevens bewaren - bijvoorbeeld voor trainingsdoeleinden - en scheiden deze vervolgens mogelijk niet voldoende af voor derden. In zijn algemeenheid is het vaak niet mogelijk vast te stellen of de leverancier verwerkte gegevens passend beschermt tegen ongeautoriseerde toegang. Wanneer medewerkers vertrouwelijke gegevens bij gebruik van de tool delen, kan dit leiden tot gerichte cyberaanvallen, afpersing, reputatieschade en boetes.
- Leveranciers van chatbots voldoen mogelijk niet aan de eisen rond de bescherming van vertrouwelijkheid die instellingen hanteren. Als aanbieders van andere applicaties niet

helder communiceren over integratie met voorgenoemde chatbots, kan de instelling onbewust blootstaan aan deze beveiligingsrisico's.
- Criminelen, statelijke actoren, maar ook bedrijven en leveranciers zelf, willen mensen en organisaties beïnvloeden of zelfs manipuleren t.b.v. financieel, politiek of ander gewin. Het is niet goed vast te stellen of leveranciers van Chatbots voorkomen dat gevoelige gegevens samen met een/het algoritme worden ingezet ten behoeve van deze doelen. Hierdoor is niet uitgesloten dat medewerkers en instellingen door een bijzonder effectieve, veelomvattende en verfijnde beïnvloedingcampagne gedeeltelijk of geheel de controle over hun leven en de organisatie verliezen.

4. Risico's ten aanzien van **compliance**, aan met name de wet AVG, omvat onder andere het volgende:
- Niet ieder land beschermt persoonsgegevens zoals beoogd in de AVG. Als persoonsgegevens van medewerkers, studenten of anderen door gebruik van chatbots op servers in 'onveilige' gebieden worden verwerkt, dan kan dit leiden tot datalekken, boetes door toezichthouders, reputatieschade en zelfs ingrijpen door de rechter of het ministerie van OCW.
- De meeste leveranciers van chatbots sluiten geen verwerkersovereenkomst af, terwijl wel persoonsgegevens worden verwerkt. Hierdoor kunnen instellingen niet voldoen aan eisen die de AVG aan verwerkingsverantwoordelijken stelt, wat kan leiden tot boetes door toezichthouders en reputatieschade.

4.1 Gebruik van data

Van wie is de informatie die de generatieve AI in jouw opdracht heeft gegenereerd? Uit welke bron komt deze informatie? Is die rechtmatig verkregen? Mag dat ook hergebruikt worden? Hoe chatbots, of GPT in het algemeen, zijn getraind en komen tot het antwoord op jouw vraag is nog altijd onduidelijk. De kans dat het om beschermde bronnen gaat is reëel. Maar dat betekent omgekeerd ook dat je in veel gevallen geen aanspraak kan maken op je auteursrecht als jij gebruik hebt gemaakt van chatbots en daarbij data (gegevens) hebt ingevoerd.

Een veel voorkomende zorg is dat chatbots van de door jou ingevoerde gegevens zou kunnen 'leren' en vervolgens weer hergebruikt. Enige bezorgdheid is terecht, maar om een andere reden. Momenteel worden chatbots getraind en vervolgens wordt het resulterende model bevraagd. Een chatbot voegt (op het moment van schrijven) niet automatisch gegevens uit zoekopdrachten toe aan zijn model zodat anderen deze kunnen opvragen. Dat wil zeggen dat het opnemen van gegevens in een zoekopdracht er niet toe zal leiden dat die gegevens in de chatbot worden opgenomen.

De ingevoerde gegevens uit zoekopdrachten zijn echter wel zichtbaar voor de leverancier achter de chatbot, bijvoorbeeld OpenAI. Deze zoekopdrachten worden opgeslagen en zullen vrijwel zeker op een gegeven moment worden gebruikt voor het (door)ontwikkelen van de chatbot-service of het data-model. Dit zou kunnen betekenen dat de leverancier (of zijn partners/contractanten) zoekopdrachten kan lezen en deze op de een of andere manier in toekomstige versies kan opnemen. Om deze reden is het van belang om de gebruikersvoorwaarden en het privacybeleid van de chatbot / leverancier goed door te nemen voor het gebruik.

4.2 Verwerking persoonsgegevens

Bij het gebruik van chatbots worden gegevens uitgewisseld en dus verwerkt waarvan de gebruikers zich lang niet altijd bewust is. Onderstaand volgt een opsomming van gegevens die mogelijk verwerkt worden[21].

- Voornaam gebruiker;
- Achternaam gebruiker;
- E-mailadres gebruiker;
- Telefoonnummer gebruiker;
- Wachtwoord gebruiker;
- Dat de gebruiker ouder is dan 18 jaar;
- Input in de conversaties (als je daar PII[22] in doorgeeft);
- IP-adres;
- Browser user agent;
- Diepgaande en unieke informatie over het gebruikte OS/browser/systeem;
- Tracking identifiers;
- Cookies.

Ook worden gegevens gedeeld met andere partijen buiten de EER[23].
Wat verder opvalt is dat er voor een groot deel van de tracking en cookies geen expliciete / transparante toestemming wordt gevraagd, laat staan dat de gebruiker een keuze krijgt.

4.3 Mis- en desinformatie, bias

Het gebruiken van AI-tools (zoals chatbots) brengt ten aanzien van de kwaliteit en betrouwbaarheid van bronnen een risico met zich mee. De antwoorden die de tool genereert kunnen geen garantie bieden ten aanzien van accuraatheid en betrouwbaarheid. Het omvat in feite een reproductie van de data die erin gestopt wordt en deze valt niet op echtheid te controleren. De datamodellen gebruiken statistische analyses om het correcte ‘antwoord’ te bepalen, maar kunnen niet de context daarbij schetsen. De gebruikersinterfaces kunnen daarbij vormgeven aan de manier waarop gebruikers de geldigheid van de antwoorden waarnemen. Daarnaast kan het voorspellende karakter van AI ook zorgen voor het in stand houden van vooringenomenheid en/of vormen van racisme door algoritmes.
Hieronder volgt een uiteenzetting van de grootste risico’s (met betrekking tot misinformatie, desinformatie en bias) binnen het onderwijs ten aanzien van het gebruik van generatieve AI-tools.

4.3.1 In-accuraatheid en bias

- In sommige gevallen kunnen gegenereerde antwoorden compleet fout zijn of misleidend (de zogenaamde ‘hallucinaties’).
- Vooringenomenheid is een risico die het gebruik van een dergelijke tool mee kan brengen. Bijvoorbeeld antwoorden die enkel getraind zijn op Westerse literatuur of enkel vanuit een mannelijk perspectief kan sociale ongelijkheid vergroten. Deze gevolgen kunnen groot zijn wanneer dit in een bepaalde omgeving wordt toegepast.

[21] Op basis van beknopte analyse ChatGPT 3.5 met gratis account, uitgevoerd door S. Veeke – SURF.

[22] PII staat voor Persoonlijk Identificeerbare Informatie.

[23] EER: https://www.rijksoverheid.nl/onderwerpen/europese-unie/vraag-en-antwoord/welke-landen-horen-bij-de-europese-economische-ruimte-eer

4.3.2 Misinformatie

- Generatieve AI-tools kunnen misbruikt worden door het genereren van hoogwaardige, goedkope en gepersonaliseerde inhoud, dit kan echter ook voor schadelijke doeleinden ingezet worden. Informatie en tools die door modellen worden ontwikkeld, worden al gebruikt voor het genereren van deep fakes (kunstmatige afbeeldingen, video en spraak van hoge kwaliteit voor desinformatie, onder andere door staats actoren) die bijna niet van echt te onderscheiden zijn voor de gemiddelde persoon.
- Bestaande uitdagingen met betrekking tot de verspreiding van desinformatie kunnen worden versterkt als AI-gegenereerde inhoud naast andere informatie circuleert.
- De door AI-tools gegenereerde inhoud zou ook misbruikt kunnen worden in democratische processen door in de AI-tool een stortvloed aan data in te voeren om de publieke opinie te misleiden.

4.3.3 Pre- en postprocessing

Applicaties en diensten gebaseerd op generatieve AI gebruiken een reeks pre- en post-processing activiteiten, zoals het herformuleren van de vraagstelling op een manier die de tool het beste kan beantwoorden, en het afwijzen van 'ongeschikte' verzoeken op basis van vooraf gedefinieerde regels. Deze verwerking codeert echter waarden die niet altijd transparant zijn voor de gebruiker of potentiële toezichthouders, en kunnen worden ondermijnd door middel van praktijken waarbij de functies van een systeem worden gebruikt om de door de provider opgelegde beperkingen op het gebruik ervan op te heffen.[24]

[24] Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). Rapid Response Information Report: Generative AI – language models (LLMs) and multimodel foundation models (MFMs). Australian Council of Learned Academies.

5 Richtlijn ‘verantwoord’ gebruik chatbots

In de voorgaande hoofdstukken is beschreven wat generatieve AI en chatbots zijn en welke kaders relevant zijn bij het gebruik hiervan. Hierbij is vooral dieper ingegaan op de AI-act en de AVG. Ook zijn de risico’s van het gebruik van chatbots toegelicht. In dit hoofdstuk is een handreiking beschreven om ondanks de risico’s chatbots toch op een zo verantwoord als mogelijke wijze te gebruiken.

5.1 Verbieden is geen optie

Iedereen gebruikt chatbots. Maar weten collega's, medewerkers of studenten hoe ze chatbots het beste kunnen inzetten én zijn ze ook voorbereid op risico's van het gebruik? Compleet verbieden is geen serieuze optie. Mensen duidelijke richtlijnen geven wel. Deze problematiek is niet nieuw, gebruikers brengen vaak ict-toepassingen van buiten af de organisatie binnen. Denk bijvoorbeeld aan WhatsApp groepen voor commmunicatie en Google Docs met persoonlijke accounts. Onderzoek suggereert dat schaduw-ICT toepassingen een inherent onderdeel zijn van het ICT landschap van een onderwijsinstelling, het verbannen is zelden wenselijk.[25]

Opties om daarmee om te gaan zijn:

1. Zorg dat basismaatregelen worden gehanteerd.
2. Zorg dat medewerkers ergens terecht kunnen met hun ICT gerelateerde ideeën.
3. Maak ICT zo gebruiksvriendelijk mogelijk.

Welk handelingsperspectief is er wel voor de security en privacy-professional? In de volgende secties worden verschillende inspiratiebronnen voor het zo verantwoord als mogelijk gebruik beschreven.

5.2 Tien gouden regels uit “Richtlijn gebruik ChatGPT” van ICTrecht

De tien gouden regels voor het ‘verantwoord’ gebruik van chatbots zijn overgenomen uit de ‘Richtlijn gebruik ChatGPT[26]’ van ICTrecht.

1. GPT is een hulpmiddel.
   Controleer en bewerk de uitvoer van GPT altijd voordat je deze met collega’s of derden deelt. Uiteindelijk blijf jij verantwoordelijk voor wat er met de uitvoer gebeurt.
2. GPT is saai.
   Pas de uitvoer van GPT altijd aan naar jou of je organisatie voordat je deze gebruikt. Het laatste dat je wilt is dat je teksten veel lijken op die van de concurrent.
3. GPT is een redacteur.
   Gebruik GPT voor het opzetten of structureren van teksten, of ter inspiratie bij het schrijven van eigen werk. Hele lappen tekst van GPT letterlijk gebruiken is niet alleen saai, maar kan ook plagiaat opleveren.
4. GPT is geen zoekmachine.
   Gebruik GPT niet voor het opvragen of controleren van feiten. GPT levert je dingen die mooi klinken, ongeacht of ze kloppen.
5. GPT kent geen context.

[25] https://communities.surf.nl/cybersecurity/artikel/schaduw-ict-in-onderwijs-en-onderzoek-wat-moet-je-er-als-instelling-mee

[26] Richtlijn gebruik ChatGPT van ICTrecht: https://www.ictrecht.nl/chatgpt

Als je een vraag stelt, neem dan zelf de achtergrondinformatie of voorbeelden op zodat GPT daarop in kan haken.

6. GPT doet niet aan privacy.
   Noem geen persoonlijke informatie van jou of anderen in een vraag aan GPT. En controleer persoonlijke informatie die eruit komt, altijd nog een keer extra. Heb je een betaald account, gebruik dan de “data-instellingen” om gesprekken privé te houden.
7. GPT doet niet aan geheimhouding.
   Alles dat je invoert, gebruik men bij het bedrijf achter GPT voor eigen doeleinden. Omzetcijfers, klantnamen, stappenplannen of strategie om de concurrent uit te schakelen: dat hoort allemaal niet in GPT. In een betaald account kun je dit uitzetten via de “data-instellingen”.
8. GPT doet niet aan controle.
   Controleer dus zelf alle uitvoer op feitelijke onjuistheden of ongewenste / ongepaste uitspraken.
9. Gebruik je gezond verstand.
   Zou je diezelfde vragen ook aan een student, collega en/of leverancier voorleggen? Zo niet, dan hoef je het GPT ook niet te vragen.
10. Gebruik een zakelijk account.
    Maak met je werk- of student mailadres een account aan, gebruik dit alleen voor werk en/of studie gerelateerde zaken en vermijd het invoeren van privé- en/of bedrijfsgevoelige informatie. Wis geen conversaties; de chatgeschiedenis is belangrijk als bewijs.

5.3 Impact Assessments

Om aan de verplichtingen uit de AVG te voldoen, kan het verstandig (en vaak zelfs verplicht) zijn om een Data Protection Impact Assessment (DPIA) uit te (laten) voeren. Zo wordt stapsgewijs nagegaan of de privacy van betrokkenen voldoende beschermd wordt.

Indien een risicoanalyse breder moet toezien dan alleen op privacy van betrokkenen door ook de technische en ethische aspecten integraal en in samenhang met andere mensenrechten te laten toetsen, dan zijn andere assessment methodes, zoals die van de IAMA of AIA of DEDA, wellicht passender. Onlangs heeft het College van de mensenrechten in een uitspraak gezegd dat een instelling zorgplicht heeft naar de student om te controleren dat toepassing van AI inderdaad niet discrimineert.[27]

De IAMA (Impact Assessment Mensenrechten Algoritmen) is een instrument dat organisaties ondersteunt bij besluitvorming rondom het ontwikkelen en inzetten van algoritmes. Naast dat er gekeken wordt naar het effect van de inzet van de AI-toepassing op mensenrechten, komen ook transparantie, uitlegbaarheid en governance aan bod.

Zowel in de DPIA en IAMA worden risico’s benoemd en is er ruimte te beschrijven welke mitigerende maatregelen worden ingericht. Dit kunnen, naast de in 5.2 genoemde regels met betrekking tot verantwoord gedrag, ook technische en organisatorische maatregelen zijn. Advies met betrekking tot deze maatregelen valt buiten scope van dit document.

[27] https://www.mensenrechten.nl/actueel/nieuws/2023/10/17/student-niet-gediscrimineerd-door-tentamensoftware-proctorio-maar-vu-had-de-klacht-zorgvuldiger-moeten-behandelen

Het is in alle gevallen aan te raden een register in te richten waarmee in kaart wordt gebracht voor welke processen AI wordt ingezet, vergelijkbaar met het bekende Register van Gegevensverwerkingen.

6 Conclusie/samenvatting

Nieuwe technologische ontwikkelingen zoals generatieve AI zullen beleidsterreinen binnen de instelling blijven uitdagen. Dit document is een eerste stap in het samenbrengen van informatie hoe om te gaan met de opkomst van AI gebaseerde chatbots. Dit document zal meerdere iteraties nodig hebben om relevant te blijven in een veranderende wereld.

Welke lessen we in ieder geval mee willen geven:

- Compleet verbieden van chatbot gebruik door eindgebruikers is niet de oplossing
- Chatbots brengen serieuze privacy en security risico's met zich mee, maar ook op breder ethisch, juridisch, en strategisch vlak.
- Er is nog geen overeengekomen standaard best practice, maar er ontstaan meerdere good practices.
- Kom tot een richtlijn of beleid voor een zo verantwoord als mogelijk gebruik van chatbots. De tien 'gouden regels' zijn hiervoor een goed startpunt.
- Voer een impact assessment uit, chatbots en andere AI toepassingen moet je gewoon behandelen net zoals alle andere applicaties.

Bijlage 1 Definities van AI

Definities van AI zijn met de tijd aan verandering onderhevig geweest. Een korte passage van het SURF report 'Promises of AI in Education' beschrijft die geschiedenis:

> "Defining 'artificial intelligence' is a complex activity, as its systematic impact results in lots of different people having different perspectives on the technology. When it comes to AI, there is a strong history of speculating about the nature of intelligence and attempting to build parts of it. Often AI is defined as the research that develops technologies that have a capability to do things that would require intelligence if done by humans. As a consequence, our perception on what AI is continuously shifts as we often see AI as the "cool things that computers can't do" and humans can. It is important to realise that AI is more of a discipline than a 'thing', which means it is not 'an AI' but 'an application of AI' or 'an AI method' that we are talking about[28]. Two often used definitions are those by Nilsson and by Russell and Norvig.
>
> > "*AI, broadly (and somewhat circularly) defined, is concerned with intelligent behaviour in artifacts. Intelligent behaviour involves perception, reasoning, learning, communication and acting in complex environments*."[29]
>
> And
>
> > *"In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving"."*[30]
>
> Both share the placing of intelligent capability within artefacts or systems and an outwards oriented element where the system seems to demonstrate intelligent behaviour. This behaviour might be learning, communicating, or acting; but at least has a proactive element where a system is 'doing' something. A more general definition can be found in the Dutch National AI Course, which defines AI as "*Intelligent systems that can perform tasks independently in complex environments and improve their own performance by learning from experience*"[31].
>
> Another definition of artificial intelligence comes from a recent report by the OECD, which refers to AI as, "*the capacity for computers to perform tasks traditionally thought to involve human intelligence or, more recently, tasks beyond the ability of human intelligence*."[32]
>
> The key word in the OECD definition is 'traditionally', as AI methods and applications have begun to challenge what people should be doing versus what machines should be doing in task-oriented work. As we shall see, the same challenges about what tasks

[28] "A Free Online Introduction to Artificial Intelligence for Non-Experts," Elements of AI, accessed January 19, 2022, https://course.elementsofai.com/.

[29] Nilsson, N. J., & Nilsson, N. J. (1998). *Artificial intelligence: a new synthesis*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-049945-1.50005-8

[30] Russell, Stuart, and Peter Norvig. "Artificial intelligence: a modern approach." (2002). http://aima.cs.berkeley.edu/

[31] "Gratis Online Cursus Kunstmatige Intelligentie," accessed April 25, 2022, https://app.ai-cursus.nl/.

[32] OECD, *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, OECD Digital Education Outlook (OECD, 2021), https://doi.org/10.1787/589b283f-en.

should be done by computer applications and what tasks should be done by people has also begun to ring through the halls of educational institutions.
These definitions might be broad, somewhat circular, and open to interpretation. In legal and policy discussions, AI is often defined more strictly. In the current draft version of the European Union AI Act, the definition specifically refers to an annex list of technologies and approaches that fall under 'AI'.
"Artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments it interacts with"[33]
This annex includes technologies based in machine learning, logic- and knowledge based, but also statistical and probabilistic approaches. Policy definitions such as this are oriented on the effect of the technology on people, as can also be seen with the mention of 'content, predictions; where recommendations, or decisions' central to AI have a severe impact on peoples' lives and wellbeing. In the current draft of the AI Act, AI systems used in education are seen as high-risk where AI systems determine access to education or evaluate persons on tests, as they can significantly determine a person's future."

Verschillende internationale organisaties hebben eigen definities opgesteld als basis voor samenwerking en beleidsontwikkeling. Zoals:

OECD.AI:

"An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."

De Europese Unie:
De EU heeft voor de aankomende AI verordening verschillende definities, in de uiteindelijke wetgevingstekst zal er een uitgewerkte definitie zijn:

European Parliament draft (2023)
'artificial intelligence system' (AI system) means a machine-based system that

- is designed to operate with varying levels of autonomy and that
- can, for explicit or implicit objectives, generate outputs
- such as content (generative AI systems), predictions, recommendations or decisions,
- that influence physical or virtual environments;

European Council draft (2022)
'artificial intelligence system' (AI system) means a system that is designed to operate with a certain level elements of autonomy and that, based on machine and/or human- provided data and inputs, infers how to achieve a given set of human-defined objectives using machine learning and/or logic- and knowledge based approaches, and produces system-

[33] EC. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts (2021/0106 (COD).

generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts;
European Commission draft (2021)
'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;
'artificial intelligence system' (AI system) means a system that

i. receives machine and/or human-based data and inputs,
ii. infers how to achieve a given set of human-defined objectives using learning, reasoning or modelling implemented with the techniques and approaches listed in Annex I, and
iii. generates outputs in the form of content (generative AI systems), predictions, recommendations or decisions, which influence the environments it interacts with;

Bijlage 2 Relevante en gebruikte bronnen

Aanvullend op de bronnen en andere verwijzingen in de voetnoten geven we hieronder nog een overzicht van andere relevante literatuur. Met hierbij ook de disclaimer dat veel nog in ontwikkeling is en deze opsomming dus niet volledig en/of actueel is.

- Inholland Statement ChatGPT_def.pdf
- Privacy Library Of Threats 4 Artificial Intelligence
- Guidelines for secure AI system development
- Verbod op generatieve AI voor rijksambtenaren in de maak
- Guidelines on the use of online generative artificial intelligence tools.pdf
- EU ai act first regulation on artificial intelligence
- Large Language Models and AI: What's the Risk to Higher Education?
- Guidance for generative AI in education and research
- NIST AI RISK MANAGEMENT FRAMEWORK
- The Promise and Peril of the AI Revolution: Managing Risk
- FPF RELEASES GENERATIVE AI INTERNAL POLICY CHECKLIST TO GUIDE DEVELOPMENT OF POLICIES TO PROMOTE RESPONSIBLE EMPLOYEE USE OF GENERATIVE AI TOOLS
- Eerste indrukken vragenlijst gebruik ChatGPT in het onderwijs
- Hogeschool Rotterdam Leer hoe ChatGPT betrouwbaar en verantwoord te gebruiken
- KU Leuven Gebruik van generatieve artificiële intelligentie als onderzoeker
- KU Leuven Verantwoord gebruik van Generatieve Artificiële Intelligentie
- U-Gent ChatGPT en andere generatieve AI-tools
- Fontys richtlijnen, overwegingen en adviesen t.a.v. GAI
- Onderwijs en AI - Universiteit van Amsterdam
- AI Technology for People - Universiteit van Amsterdam (uva.nl)

Bijlage 3 Samenstelling werkgroep

Deze handreiking is tot stand gekomen door samenwerking in de werkgroep 'Handreiking chatbots' met een brede vertegenwoordiging vanuit de sector.

Aan de werkgroep is een bijdrage geleverd door o.a.:

- Anthony Gadson – NWO
- Arvin Khozooei – TU Delft
- Barbara Gerretsen – Universiteit van Amsterdam
- Bas Dekker – ICT Recht
- Bertine van Deyzen – SURF
- Carla Brinkman – SURF
- Dominique Dalloesingh – ROC van Amsterdam
- Duuk Baten – SURF
- Ed de Vries – SURF (voorzitter van de werkgroep)
- Elske posthuma – NHL Stenden
- Ewald Beekman – Amsterdam UMC
- Floor May-Smit – NHL Stenden
- Heleen van der Laan – ROC van Amsterdam
- Helma de Boer – SURF
- Inge Leenheer – Avans
- Jan van den Berg – Hogeschool van Amsterdam
- Joan Schrijvers – WUR
- Job Geven – Hogeschool Arnhem Nijmegen
- Martijn van Hoorn – Yuverta
- Matijs van Hilst – Hotelschool Den Haag
- Merijn Nicolai – ROC Nijmegen
- Mick Deben – MBO Digitaal
- Nicole van Deursen – SURF
- Patrick van Aalst – NHL Stenden
- Remy van den Boom – TNO
- Ruurdje Procee – Hogeschool van Amsterdam
- Sanderijn Kuijvenhoven – SURF
- Saskia van den Hout – Universiteit Utrecht
- Satwik Singh – TU Eindhoven
- Thomas van Osch - SURF
- Zumreta Schuurman – NWO

Bijlage 4 Tien gouden regels