Interprovinciaal Overleg
van, voor en door provincies

Update Ontwikkelingen omtrent ChatGPT

VERSIE: SEPTEMBER 2024

Interprovinciaal Overleg
Gezamenlijke provincies
WWW.IPO.NL

Inhoudsopgave

Hoofdstuk 1. Inleiding en advies van de commissie

Begin juli 2023 presenteerde de Interprovinciale Ethische Commissie haar advies op het gebruik van ChatGPT binnen de provincies, in de “Verkenning ChatGPT, overwegingen voor verantwoord gebruik”. Sindsdien heeft de wereld niet stilgestaan, en de ontwikkelingen omtrent (het gebruik van) generatieve AI al helemaal niet. Daarom bespreken we in dit document de ontwikkelingen omtrent ChatGPT van het afgelopen jaar.

In 2023 adviseerde de Interprovinciale Ethische Commissie de provincies om vooral de rust de bewaren. Hoewel het begrijpelijk is dat provincies willen experimenteren met deze nieuwe technologie, is het niet nodig om overhaaste beslissingen te nemen. Het gebruik van generatieve taalmodellen vraagt om zorgvuldigheid en weloverwogen beslissingen, om kwaliteit en veiligheid te kunnen garanderen. De commissie adviseerde daarbij altijd af te wegen of de inzet van een taalmodel proportioneel en noodzakelijk is, om ook alternatieven te onderzoeken, en om een exit-strategie te hebben voor als er problemen optreden.

Het advies van 2023 is nog steeds van toepassing. We hebben ook gezien dat provincies dat advies ter harte hebben genomen en het in eigen richtlijnen voor het gebruik van taalmodellen hebben meegenomen. Maar nieuwe onderzoeken en opkomende richtlijnen kunnen extra inspiratie of handvatten bieden bij de overwegingen om taalmodellen in te zetten. Daarom willen we de provincies in dit document een kleine update geven van de ontwikkelingen rondom generatieve taalmodellen. Dit doen we aan de hand van wetenschappelijk onderzoek en andere publicaties. Hiermee schetsen we een beeld van de technologische ontwikkelingen, open source alternatieven en de opkomende richtlijnen, beleid en wetgeving voor het gebruik van generatieve AI.

In het volgende hoofdstuk vindt u eerst de samenvatting van de ethische analyse zoals deze in de oorspronkelijke verkenning is opgenomen. In hoofdstuk 3 bespreken we vervolgens de belangrijkste ontwikkelingen op het gebied van generatieve taalmodellen op het gebied van de technologie, wetgeving, maatschappelijke impact en open alternatieven.

Hoofdstuk 2. Samenvatting Ethische Analyse Verkenning ChatGPT

ChatGPT van OpenAI staat sinds de lancering in november 2022 in de belangstelling van vele individuele gebruikers. Dit model is gelanceerd in november 2022 voor een groot publiek en roept naast enthousiasme, veel vragen op. Dit vraagt om een analyse van de ethische aspecten van het gebruik van LLM’s in het algemeen bij de provincies, en ChatGPT in het bijzonder.

2.1 Technische aspecten

Large Language Models (LLM’s) zijn getraind op omvangrijke datasets die grootschalig gebruik maken van openbare bronnen. ChatGPT is een LLM, en kan grote hoeveelheden tekst verwerken en patronen herkennen én voorspellen. Het kan daardoor teksten genereren waar onderling tussen de tekstdelen samenhang is, en het kan ook ingevoerde teksten verwerken. Aan de voorkant heeft de gebruiker te maken met iets wat lijkt op een geavanceerde chatbot. Op basis van een opdracht (‘prompt’) genereert ChatGPT output in de vorm van tekst. Daarover kan de gebruiker met het systeem verder over in interactie gaan.

De ***toepassingsmogelijkheden*** van ChatGPT in de provincie zijn velerlei. Het biedt de mogelijkheid om werk met een talig karakter te automatiseren en te versnellen en zou daarbij ondersteunend kunnen werken. De toepassingsmogelijkheden gaan van het genereren van tekst of code tot het structureren of leesbaar maken van teksten. Denk aan het samenvatten van verslaglegging of het genereren van (delen van) beleidsnotities.

Het model kent ook beperkingen. Het belangrijkste kenmerk is de ***onbetrouwbaarheid*** van ChatGPT. Het model is ten eerste niet neutraal, doordat het model waardengeladen is (door de makers en trainers van het model) en bias kent. Ten tweede is er sprake van zogenaamde *‘hallucinaties’*. ChatGPT als LLM heeft geen model van de werkelijkheid en produceert taal, geen kennis. Ten derde zijn het model, de trainingsdata en de keuzes in de totstandkoming van het model *intransparant*. De intransparantie is grotendeels een gevolg van de complexiteit van het model en de wijze waarop het is ontworpen. Daardoor is het moeilijk te achterhalen hoe en op basis waarvan het model tot output komt. Deze beperkingen brengen risico’s met zich mee naarmate er meer ‘autonomie’ aan ChatGPT wordt gegeven.

2.2 Wettelijke kaders

De wettelijke kaders waarbinnen ChatGPT gebruikt zou kunnen worden leveren spanningen op. In verschillende Europese landen loopt er onderzoek naar de schending van de ***Algemene Verordening Gegevensbescherming (AVG)*** door OpenAI. Reeds geïdentificeerde privacyrisico’s bij dit model zijn de zogenoemde prompt leaks en het gebruik van privégegevens (verkregen uit openbare bronnen en door het gebruik van ChatGPT) als trainingsdata. In het kader van de provincie moeten deze privacyrisico’s breder worden getrokken naar bedrijfs(gevoelige) gegevens.

Generatieve AI in het algemeen zet de aankomende ***Europese AI verordening*** onder druk waardoor er onlangs bepaald is dat er aanvullende eisen aan Foundation Models (waaronder ChatGPT) worden gesteld. Er is bij ChatGPT sprake van vraagstukken rondom ***Copyrightschending***. Bronnen die openbaar toegankelijk zijn, zijn zonder bronvermelding of toestemming gebruikt als trainingsdata enworden gebruikt om tot output te komen.

Het gebruik van ChatGPT moet in de provincie aan de ***Provinciewet*** en de ***ambtseed*** voldoen, die beide om een zorgvuldige omgang met gegevens vragen. De vraag is of dit haalbaar is als ChatGPT wordt gebruikt.

De bestaande wettelijke kaders roepen diverse vraagstukken op hoe de provincies zich hiertoe zullen verhouden bij het gebruiken van ChatGPT. Ondertussen zijn er ook oproepen tot nog strenger regulatie en beleid, in Nederland, maar ook daarbuiten.

2.3 Ethische aspecten bij het gebruik van ChatGPT

Het gebruik van ChatGPT bij de provincie betekent iets voor de ambtenaar, de maatschappij en de provincie als organisatie. Het zal impact hebben op de *vaardigheden*, het *vakmanschap* en de *autonomie* van de ***ambtenaar***. Er is bijvoorbeeld een risico dat de verantwoordelijkheid (ongemerkt) verlegd wordt en de ambtenaar te veel op ChatGPT gegenereerde teksten of mogelijkheden van de technologie gaat leunen. De ***maatschappij*** heeft behoefte aan en baat bij een betrouwbare overheid. Bij het maken van rechtvaardige keuzes speelt de invloed van juiste informatie een rol. De onbetrouwbaarheid van het model, dat geen informatie maar *tekst* genereert, en de beperkte mogelijkheid tot verantwoording bij het gebruik van dit model bevat risico's voor de informatiekwaliteit en het effect daarvan op publieke waarden en mensenrechten. Dit kan leiden tot onbetrouwbaar handelen van ***de provincie*** en heeft gevolgen voor het in de provincie gestelde vertrouwen. Een ander effect is de mogelijke invloed op de arbeidsinzet en de verandering van rollen en roept vragen op over de toekomstbestendigheid van de organisaties. Een belangrijke ethische vraag is in wiens belang het inzetten van deze technologie is en of dit niet conflicteert met andere (morele) verplichtingen en wettelijke taken die de provincie heeft

2.4 Verborgen kosten

Er zijn ook ethische aspecten die gepaard gaan met de *inzet* en de ontwikkeling van deze modellen. De reeds genoemde ethische vraagstukken kunnen al worden gezien als 'kosten' voor de samenleving. Daarnaast is er sprake van andere 'verborgen kosten'. Het gebruik van ChatGPT heeft een flinke ***klimaat***impact die conflicterend is met de provinciale plicht tot natuurbescherming. Het gebruik van ChatGPT draagt bovendien bij aan een ongewenste ***machts***concentratie bij Big Tech en heeft door het technische ontwerp risico's voor onze publieke waarden. Tot slot maakt het gebruik van en versterkt het economische en sociale ***ongelijkheden*** door het model te trainen en bij te laten sturen door *clickworkers* die onder erbarmelijke omstandigheden dit werk doen. Op basis hiervan zou men kunnen stellen dat er sprake is van een wens tot efficiëntie, maar de vraag is of de efficiëntieslag een illusie is. De last vermindert niet, maar wordt doorgeschoven en komt in andere vormen en bij anderen terug.

2.5 Technologische alternatieven

Er zijn alternatieve ***open source*** modellen beschikbaar, die in verschillende fasen van ontwikkeling zijn. Deze open source modellen zijn doorgaans transparanter over zaken als trainingsdata en klimaatimpact. De huidige open source modellen zijn op dit moment nog sterk afhankelijk van de toegang die de grote commerciële modellen verlenen aan hun modellen. De prestatie- en inzetmogelijkheden van de alternatieve open source modellen zijn op dit moment niet op het niveau van ChatGPT. Ze bieden bijvoorbeeld nog geen gebruiksvriendelijke *user interface*, vereisen lokale installatie en hebben meer data nodig om ze verder te *finetunen*. Redelijk succesvol zijn kleinere modellen die getraind worden om één specifieke taak uit te voeren. Deze zijn wel gebonden aan scopebepaling en licentiemodellen. De ontwikkelingen gaan snel en mogelijk worden open source modellen steeds geschikter voor grootschalige inzet. Bij de verkenning van een alternatief, zou de overlap met ethische vraagstukken zoals verder in deze analyse geschetst zijn, verder moeten worden onderzocht.

Hoofdstuk 3. Ontwikkelingen

In dit hoofdstuk beschrijven we de ontwikkelingen op het gebied van LLM's en ChatGPT in het bijzonder. Dit doen we op het gebied van de techniek (3.1), de wettelijke kaders (3.2), de maatschappelijke impact (3.3) en de ontwikkeling van open source alternatieven (3.4).

3.1 Technische aspecten

Deze gids verwijst naar instrumenten die binnen provincies worden toegepast ten behoeve van een verantwoorde inzet van data en technologie. Deze instrumenten zijn ontwikkeld door experts op het gebied van digitale ethiek. Wat vinden deze experts van de gids?

Toepassingsmogelijkheden van ChatGPT

Naast het genereren van tekst is het in betaalde versies van ChatGPT nu ook mogelijk om te werken met afbeeldingen en spraak. Dit leidt tot aanvullingen in de toepassingsmogelijkheden, zoals het genereren van afbeeldingen op basis van tekst.

3.2 Wettelijke kaders

Privacy en auteursrecht

November 2023 werd het artikel "Scalable extraction of training data from (production) language models" gepubliceerd[1] Dit onderzoek liet zien dat ChatGPT met een vrij eenvoudige 'aanval' (een prompt die aan ChatGPT vroeg een bepaald woord eeuwig te herhalen), het model uiteindelijk afweek van de prompt en gememoriseerde training data als output gaf. Deze output bevatte ook persoonsgegevens en teksten van bronnen die onder het auteursrecht vallen. Hoewel OpenAI zich inmiddels tegen deze specifieke aanval gewapend heeft, laat het zien dat er kwetsbaarheden in het systeem zitten en dat modellen, ondanks de inspanningen om dat te voorkomen, trainingsdata kunnen 'onthouden' en rek produceren.

Wat betreft het gebruik van documenten waar auteursrecht op rust, is OpenAI zelf van mening dat het niet de wet overtreedt en dat het nodig is om een goed en bruikbaar taalmodel te creëren. OpenAI schreef in december 2023 het volgende in een statement aan een onderzoekscommissie van de Britse *)*overheid (House of Lords Communications and Digital Select Committee inquiry):

"OpenAI's large language models, including the models that power ChatGPT, are developed using three primary sources of training data: (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or our human trainers provide. Because copyright today covers virtually every sort of human expression–including blog posts, photographs, forum posts, scraps of software code, and government documents– it would be impossible to train today's leading AI models without using copyrighted materials. Limiting training data to public domain books and drawings created more than a century ago might yield an interesting experiment, but would not provide AI systems that meet the needs of today's citizens."[2]

Daarnaast biedt OpenAI de mogelijkheid om de functie uit te schakelen waarbij informatie die gebruikers delen met het model wordt gebruikt om het model verder te trainen. Dit is een belangrijke instelling om te controleren, om (bedrijfs)gevoelige informatie te beschermen.

Europese AI Verordening

Op 13 maart 2024 heeft het Europees Parlement de AI Verordening ingestemd en in mei heeft ook deRaad goedkeuring gegeven. In juli is de verordening gepubliceerd en per 1 augustus wordt de wet in verschillende fasen van kracht. Ook voor ChatGPT heeft deze wet gevolgen, omdat ChatGPT een zogenaamd 'general purpose' AI-systeem is, omdat het voor verschillende doeleinden kan worden gebruikt. Voor deze modellen gaat een verplichting gelden om bepaalde informatie te delen, zoals een beschrijving van het model, de architectuur, en de gebruikte data.[3]

Onderzoekers van Stanford University hebben in 2023 al op basis van de conceptversie van de AI-verordening onderzocht in hoeverre op dat moment verschillende taalmodellen voldeden aan de nieuwe verplichtingen van de verordening. Dit hebben ze met scores weergegeven in een overzicht, op basis van 12 eigenschappen. Het gaat voornamelijk over transparantie. De evaluatie laat zien dat er voor alle modellen nog vooruitgang te boeken is op het gebied van transparantie over de taalmodellen.[4]

De Verenigde Staten

In de Verenigde Staten heeft president Biden in oktober 2023 een decreet getekend dat de veiligheid en privacy van burgers moet beschermen bij het ontwikkelen en gebruiken van AI-systemen.[5] Zo moeten onder andere de testresultaten op het gebied van veiligheid van 'krachtige AI systemen' gedeeld worden met de overheid.

Ook was de opkomst van generatieve AI in de VS onderdeel van de redenen voor schrijvers voor tv en films om te staken. Dit heeft geleid tot afspraken over hoe AI wel of niet gebruikt mag worden, om de rechten van schrijvers te beschermen.[6]

3.3 Maatschappelijke impact

Begin december 2023 presenteerde het Rathenau Instituut de Scan Generatieve AI[7], een onderzoek naar generatieve AI in opdracht van het ministerie van BZK. Deze scan is ook gebruikt als input voor de Overheidsbrede visie op generatieve AI. In deze scan worden vier risico's benoemd bij de opkomst van Generatieve AI. Ten eerste is de veiligheid in het geding door de mogelijke schendingen van privacy, het genereren van foutieve of bevooroordeelde informatie en de intransparantie van het complexe achterliggende model. Ten tweede zijn er zorgen over hoe mensgericht de systemen zijn en welke invloed ze hebben op onze cognitieve, sociale en culturele omgangsvormen. Ten derde is de vraag hoe eerlijk de systemen zijn, gezien de invloed op de maatschappij, veranderende banen, maar ook de invloed op het milieu. Tot slot heeft generatieve AI invloed op de democratie, vanwege de machtspositie van grote techbedrijven en de mogelijke invloed van generatieve AI op het publieke debat.

Ook publiceerde de Autoriteit Persoonsgegevens in december de tweede halfjaarlijkse Rapportage AI & algoritmerisico's Nederland (RAN).[8] waar specifiek in werd gegaan op de risicobeheersing van generatieve AI. Hier benadrukken zij het belang van algoritmische geletterdheid in de organisatie, om deze technologie op een verantwoorde manier in te zetten. Ook stellen zij dat de huidige risicobeheersingsinstrumenten nog niet specifiek van toepassing zijn op generatieve AI. Er zouden speciale 'impact assessments' en auditstandaarden moeten komen voor generatieve AI. Bovendien geeft de AP aan in te willen zetten op het vormgeven van kaders voor het veilig gebruiken van generatieve AI door organisaties.

Richtlijnen gebruik

Verschillende provincies hebben eigen richtlijnen/gedragscodes gevormd voor het werken met generatieve AI, mede op basis van het advies van de Interprovinciale Ethische Commissie van 2023. Deze richtlijnen wijzen gebruikers op hun eigen verantwoordelijkheid en verbieden het invoeren van persoonsgegevens, bedrijfsgevoelige, geheime of onwenselijke informatie in generatieve AI tools. Ook benoemen ze het belang van het controleren van de uitkomsten vanwege de beperkingen van taalmodellen, het nadenken over het nut en de noodzaak van het gebruik van een tool en de maatschappelijke kosten die eraan verbonden zijn. Het wordt gestimuleerd om er met elkaar over in gesprek te gaan en samen te leren en ervaring op te doen.

De WHO (World Health Organization) heeft begin 2024 richtlijnen gepubliceerd voor 'ethics and governance' in de ontwikkeling en het gebruik van large multi-modal models in de gezondheidszorg.[9] Deze zijn gestoeld op de in 2021 gepresenteerde ethische principes voor AI in het gezondheidsdomein van de WHO. Dit document geeft een overzicht van de risico's die in verschillende fasen (development, provision, deployment) van een model geadresseerd moeten worden en wie er in deze fase actie kan ondernemen om de risico's te mitigeren. In elke van de fasen ligt er een grote rol voor de overheid om verplichtingen te stellen aan bijvoorbeeld de mate van transparantie, het houden aan ethische standaarden, audits, maar ook om te investeren in training en participatie van de gebruikers.

Overheidsvisie

Op 11 december 2023 maakte staatssecretaris van Huffelen een "Voorlopig standpunt voor Rijksorganisaties bij het gebruik van generatieve AI" bekend.[10] Hierin stelde ze dat Rijksoverheidorganisaties voor het gebruik van generatieve AI een risicoanalyse moeten uitvoeren. Ook krijgt het werken met open source generatieve AI de voorkeur, in het kader van de Wet Open Overheid en het stimuleren van transparantie. En stelde ze dat het gebruik van niet-gecontracteerde generatieve AI toepassingen (zoals ChatGPT) niet toegestaan zijn te gebruiken, omdat deze (over het algemeen) niet voldoen aan de privacy- en auteursrechtelijke wetgeving.

Dit was een voorlopig standpunt, omdat de Overheidsbrede visie op Generatieve AI nog in ontwikkeling was. Deze werd in januari 2024 gepubliceerd.[11] De overheid heeft als ambitie om in Nederland en de EU aan een sterk AI-ecosysteem te werken, wat betekent dat er ruimte is voor innovatie en het benutten van de mogelijkheden van generatieve AI. Echter, naast de kansen zijn er ook verschillende risico's voor individuele burgers, de arbeidsmarkt, de maatschappij en de markt voor techbedrijven. Daarom is het een voorwaarde dat generatieve AI op verantwoorde wijze wordt ontwikkeld en gebruikt. Hiervoor zijn vier waardengedreven uitgangspunten geformuleerd: 1) generatieve AI wordt op een veilige manier ontwikkeld en toegepast, 2) generatieve AI wordt op een rechtvaardige wijze ontwikkeld en toegepast, 3) generatieve AI dient het menselijk welzijn en borgt de menselijke autonomie, en 4) generatieve AI draagt bij aan duurzaamheid en onze welvaart. De visie beschrijft een aanpak die is gebaseerd op samenwerking en samen leren, waarbij de ontwikkelingen op het gebied van generatieve AI nauwlettend worden gevolgd en de kennis op dit gebied wordt vergroot. Ook wordt er gewerkt aan de toepassing van wet- en regelgeving en het vormgeven van toezicht en handhaving.

3.4 Technologische alternatieven

Open source LLM's

In juli 2023 werd het artikel "Opening up ChatGPT: tracking openness, transparency and accountability in instruction-tuned generators" gepubliceerd.[12] In dit onderzoek is van verschillende LLMs geïnventariseerd welke informatie zij openbaar hebben gemaakt. Denk aan de beschikbaarheid van de code, documentatie van het model en de toegankelijkheid van het model. Dit heeft een overzicht opgeleverd van hoe transparant en open verschillende modellen zijn, waarbij ChatGPT het slechtste scoort (van de op dit moment 45 onderzochte modellen in de github omgeving[13]).

Begin november 2023 kondigde TNO, NFI en SURF aan samen een nationaal open taalmodel te ontwikkelen: GPT-NL.[14] Het project wordt gefinancierd vanuit RVO/Ministerie van Economische Zaken. Een eigen taalmodel biedt verschillende voordelen ten opzichte van andere modellen: 1) Nederland ontwikkelt publieke expertise en ervaring op het gebied van generatieve AI, 2) het model wordt volledig in lijn met waarden en wetgeving gebouwd en 3) het model is open en transparant en 4) het vermindert de afhankelijkheid van (niet-Europese) big tech bedrijven. Het gaat nog wel even duren voordat het model er daadwerkelijk

is; in de aankondiging staat dat het ‘eerste jaar’ wordt besteed aan de ontwikkeling van het Nederlandse taalmodel. Daarna volgt nog de exploitatiefase, waarin de nationale supercomputer Snellius van SURF wordt ingezet om het model te laten werken.

Op Europees niveau wordt er aan verschillende open taalmodellen gewerkt, zoals OpenGPT-X (Duitsland)[15], GPT-SW3 (Zweden)[16] en Viking (Finland)[17]. Ook is er begin 2024 een initiatief gestart, the Alliance for Language Technologies (ALT-EDIC), dat zich richt op de ontwikkeling en ondersteuning van Europese LLMs.[18]Een EDIC (European Digital Infrastructure Consortium) is een nieuwe Europese rechtspersoon die wordt gecreëerd door een besluit van de Europese Commissie, maar wordt opgericht door 3 of meer Lidstaten. Nederland (Ministerie van BZK) is een van de oprichters van ALT-EDIC, en de Nederlandse taken worden o.a. door de Hogeschool Utrecht en TNO vervuld. Daarnaast is er een Europees ondersteund initiatief voor de Europese gemeenschappelijke dataruimte voor taal (Language data space), dat voortbouwt op de AI en taal activiteiten die o.a. de basis vormen voor de vertaaltools van de Europese Commissie. De language data space moet taalmodellen bruikbaar maken in de andere dataruimten (de Europese eenheidsmarkt voor vrij verkeer van data).

Hoofdstuk 4. Verantwoording

Deze “Update Ontwikkelingen omtrent ChatGPT” is geschreven door The Green Land. Het is een aanvulling op de “Verkenning ChatGPT”, welke tot stand is gekomen in opdracht van het Interprovinciaal Overleg. Piek Knijff, Isis Hazewindus en Joeri Hazelaar van Filosofie in actie hebben in juni 2023 vooronderzoek verricht en de ethische analyse geschreven. De samenvatting van deze analyse is ook in dit document opgenomen.

Hoofdstuk 5. Interprovinciale Ethische Commissie

Over de commissie

De interprovinciale ethische commissie is gericht op het soort vraagstukken dat alle provincies gelijkelijk raakt, zoals rond de opkomst van een nieuwe technologie als Artificial Intelligence (AI), die mogelijk van waarde kan zijn voor de uitvoering van provincietaken, of het gebruik van een specifieke technologie binnen de context van een specifiek maatschappelijk vraagstuk, zoals voor slimme oplossingen voor de hierboven genoemde energietransitie.

De werkzaamheden van de commissie zijn nadrukkelijk van onderzoekende aard, en dus niet gericht op besluitvorming. Door vraagstukken te onderzoeken kan een interprovinciale ethische commissie adviezen verstrekken die handelingsperspectief bieden aan de provincies bij het maken van hun eigen afwegingen en besluitvorming. De commissie draagt op die manier bij aan het sterker borgen van kernwaarden van het publiek domein in de werkzaamheden van provincies.

Samenstelling

Jeroen van den Hoven (voorzitter)
Anne-Marie Spierings
Erna Ruijer
Nelleke Groen
Roel Dobbe

Contact

ethischecommissie@ipo.nl

Colofon

De interprovinciale ethische commissie richt zich op morele vraagstukken waarvoor provincies gesteld staan bij de toepassing van digitale technologie op hun maatschappelijke opgaven. Zowel digitale technologie als maatschappelijke opgaven en de context waarin provincies werken zijn voortdurend in beweging. Dit maakt elk advies een momentopname gegeven de technologie, opgaven en context van dat moment. Dit betekent dat de interprovinciale commissie periodiek publicaties en adviezen kan herzien en bijwerken naar de laatste stand van zaken. Daarom zijn verkenningen, handreikingen en adviezen voorzien van versienummers en een datum. De actuele versie van een document zal telkens op de website van het IPO te vinden zijn.

Dit is “Update Ontwikkelingen omtrent ChatGPT”. Versie 1.0, d.d. 12-09-2024

Bronnen

1. Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035.
2. https://committees.parliament.uk/writtenevidence/126981/pdf/
3. https://www.europarl.europa.eu/topics/nl/article/20230601STO93804/ai-verordening-eerste-regels-voor-artificiele-intelligentie
4. https://crfm.stanford.edu/2023/06/15/eu-ai-act.html
5. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/
6. https://www.wgacontract2023.org/the-campaign/summary-of-the-2023-wga-mba
7. Rathenau Instituut (2023). Generatieve AI. Den Haag. Auteurs: Hamer, J., L. Kool, B. Hijstek, Q. van Eeden en D. Das. https://www.rathenau.nl/nl/digitalisering/generatieve-ai
8. Autoriteit Persoonsgegevens (2023). Rapportage AI- & algoritmerisico's Nederland (RAN) – najaar 2023. https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023
9. World Health Organization (2024). Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. Licence: CC BY-NC-SA 3.0 IGO.
10. Van Huffelen, A. C. (2023, 11 december). Voorlopig standpunt voor Rijksorganisaties bij het gebruik van generatieve AI. Geraadpleegd op 20 maart 2024, van https://www.rijksoverheid.nl/documenten/kamerstukken/2023/12/11/kamerbrief-over-voorlopig-standpunt-voor-rijksorganisaties-bij-het-gebruik-van-generatieve-ai
11. Rijksoverheid. (2024, 18 januari). Overheidsbrede visie Generatieve AI. Geraadpleegd op 23 maart 2024, van https://www.rijksoverheid.nl/documenten/rapporten/2024/01/01/overheidsbrede-visie-generatieve-ai
12. Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023, July). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In Proceedings of the 5th international conference on conversational user interfaces (pp. 1-6).
13. https://opening-up-chatgpt.github.io/
14. https://www.tno.nl/nl/newsroom/2023/11/nederland-start-bouw-gpt-nl-eigen-ai/
15. https://opengpt-x.de/
16. https://www.ai.se/en/project/gpt-sw3
17. https://www.silo.ai/blog/viking-7b-the-first-open-llm-for-the-nordic-languages
18. https://language-data-space.ec.europa.eu/related-initiatives/alt-edic_en

Telefoon: 070 888 1212
E-mail: ethischecommissie@ipo.nl

Den Haag
Herengracht 23