Recht & GPT-4?

Aernout Schmidt (emeritus recht & informatica. Leiden)

> Chomsky sees the use of ChatGPT as “basically high-tech plagiarism” and “a way of avoiding learning." www.openculture.com/2023/02/noam-chomsky-on-chatgpt.html
>
> "Iedereen wordt nu geconfronteerd met het potentieel vervreemdende, misleidende of parasitaire karakter van deze technologie en de mogelijke bedreiging voor echte schrijvers en kunstenaars. Hoe zit het met het auteursrecht?" Dirk Visser in Nederlands Juristenblad | 17 February 2023

Toen ik voor het eerst via ChatGPT met GPT-4[1] experimenteerde geloofde ik mijn ogen niet. Een voortreffelijke samenvatting van een obscuur stuk van Nietzsche (”On Truth and Lie in an Extra-Moral Sense,” 1873) gevolgd door een langdurige, van GPT-4 zijde volstrekt samenhangende dialoog[2] over technische kanten van GPT-4 . Ik ondervond zelf de intellectuele opschudding die prominente AI-professionals ertoe bracht hun brief van 22 maart jl op te stellen (zie *https://futureoflife.org/open-letter/*.) Ze zien acute gevaren en doen een beroep op iets dat regulering had kunnen zijn maar er niet is.

In een vlog op YouTube geeft één van hen, Mo Gawdat, voorheen chief business officer van Google X, een nuttige taxonomie van attitudes die ontstaan in het licht van nieuwe AI. Wanneer ik die overneem uit praktische overwegingen, richt ik me tot een divers publiek van onwetenden, kids, opportunisten, utopische evangelisten, dystopische evangelisten en eenlingen.

> *De taxonomie.* Ik geef een (automatisch vertaald en enigszins gepolijst) transcript: "... het zijn de onwetenden, mensen die je zullen vertellen ‘oh nee nee, dit gebeurt niet, AI zal nooit creatief zijn, zal nooit muziek componeren’; dan heb je de kids (ik noem ze zo), overal op sociale media, die zeggen ‘oh mijn god, het piept, kijk ernaar, het is oranje van kleur, ah geweldig, ik kan niet geloven dat AI dit kan’; dan hebben we de opportunisten, die gewoon zeggen ‘kopieer dit, plak het in ChatGPT, ga dan naar YouTube, jat wat, respecteer het intellectueel eigendom van niemand, plaats het in een video en nu ga je honderd dollar per dag verdienen’; en natuurlijk hebben we de dystopische evangelist, eigenlijk zeggen die ‘dit is het dan, de wereld vergaat’, wat volgens mij niet de realiteit is; je hebt ook de utopische evangelisten die zeggen ‘oh je begrijpt het niet, we gaan kanker genezen, we gaan dit en dat overwinnen’, alweer geen realiteit; en dan blijft de eenling over die zegt ‘wat gaan we eraan doen?’ ... " [youtube.com/watch?v=Cv2Xy77B7Co]

[1] ChatGPT is de chatbot die het gesprek tussen de gebruiker en GPT-4 bemiddelt. GPT-4 is een groot, vooraf via *deep learning* met behulp van neurale netwerken en menselijke hulp getraind natuurlijke-taalmodel dat sinds maart van dit jaar beschikbaar wordt gesteld door OpenAI – maar daarover later. Trouwens, ik raadpleegde het in het Engels. Toen ik het later in het Nederlands probeerde was ik minder onder de indruk. Maar dat is vermoedelijk van voorbijgaande aard, gelet op de hype die is ontstaan.

[2] Voor wie daar naar kijken wil: *https://chat.openai.com/share/f44ce062-9cff-485f-bc1e-74d8bcab7bd0*)

Het is niet duidelijk waar Mo zelf past in zijn eigen taxonomie. Wanneer ik de titels van sommige van zijn geschriften zie, zie ik een combinatie voor me: Mo als utopische evangelist (*That Little Voice In Your Head: Adjust the Code That Runs Your Brain*, 2000), als opportunist (*Solve for Happy: Engineer Your Path to Joy*, 2017), als dystopische evangelist (*Scary Smart: the Future of Artificial Intelligence and How You Can Save Our World*, 2021) en als eenling (de *vlog* waaruit ik citeer, 2023). Mo heeft zich van utopisch evangelist via opportunisme en dystopische evangelisatie ontwikkeld tot een eenling, of nee, Mo's persoonlijkheid is eerder meerlagig geworden. Mijn eigen observaties (van 1969 tot 2014) in de IT-gerelateerde westerse academische wereld bevestigen dat dit een gebruikelijk intellectueel traject is.

Mo Gawdat identificeert ook wat volgens hem het grootste risico van GPT-4 gebruik vormt: de loskoppeling van macht en verantwoordelijkheid.

> *Het grootste risico.* Uit dezelfde vlog: "[...] en de grootste uitdaging als je mij vraagt wat er mis is gegaan in de 20e eeuw, is dat we te veel macht hebben gegeven aan mensen die de verantwoordelijkheid niet op zich namen [...] We hebben macht en verantwoordelijkheid losgekoppeld, dus vandaag kan een 15-jarige (inclusief diens limbische horten en stoten en nog zonder een volledig ontwikkelde prefrontale cortex) beslissingen nemen [...] en het probleem dat we vandaag hebben is dat er een kloof is tussen degenen die de code van AI schrijven en de verantwoordelijkheid voor wat er gaat gebeuren vanwege die code, oké? En ik voel medeleven met de rest van de wereld, ik voel dat dit verkeerd is, ik heb het gevoel dat je weet dat iemands leven wordt beïnvloed door de acties van anderen zonder inspraak te hebben [...] culminerend in het ultieme, het absolute hoogtepunt van menselijke stupiditeit"

Door GPT-4 in de wereld los te laten, heeft OpenAI een maatschappelijk *tipping point* veroorzaakt. Het – want GPT is voorlopig onzijdig – slaagde bijvoorbeeld in april voor het Amerikaanse *bar exam*. Welnu, dit is m.i. niet alleen voor juristen relevant, ik kan me namelijk voorstellen dat mensen gaan denken dat de wereld *a profound change in the history of life on Earth* aan het beleven is. Ik zie daarbij ook een diep gevoelde wens om een verdere ontwikkeling van GPT-4 te reguleren. *Should be planned for and managed with commensurate care and resources* staat ook in de brief van 22 maart. Daarmee wordt gesuggereerd dat wanneer we dat nalaten we onze ondergang tegemoet gaan.

Om een oordeel te geven over de vraag naar welk beroep dit doet op het recht kijk ik eerst even naar *GPT-4* omdat ik dan onder de kop *High-tech plagiaat?* Visser's baanbrekende auteursrechtelijke beschouwing kan bespreken.

Ik laat zien hoe mijn aanvankelijke aanname dat het auteursrecht de doemdenkers weinig te bieden heeft toch van tafel moet worden geveegd. Daar komt dan nog bij, dat het beroep op regulering breder is. In *Maatschappelijke uitdagingen* inventariseer ik wat de belangrijkste bedreigingen zijn die ik zie en welk recht ik daarbij relevant acht; ik geef alleen een aanzet. En ja, ik eindig met een oproep tot actie, in *We moeten aan de bak, vluchten kan niet meer*.

GPT-4

Mo Gawdat gaf in de genoemde vlog ook een handzame schets van hoe GPT-4 werkt – als voorspelmachine op zoek naar het beste volgende woord in de zin die het als antwoord op een vraag aan het maken is. GPT memoriseert en genereert daarbij *betekenissen* uit de teksten die het te lezen kreeg.

Dat zijn heel veel teksten, een internet vol teksten. GPT beschikt zo over de betekenissen van bij benadering *alle* teksten, van wat alle auteurs ooit hebben opgeschreven. En het toont zijn wonderlijke vaardigheden door een volgend woord te voorspellen (en dan nog een etc.), door steeds het volgende woord in het antwoord op een vraag (bijvoorbeeld: *maak een eerste hoofdstuk in een roman die Houellebcq in 2029 zou schijven*) te zoeken en in te passen. Eigenlijk gaat het niet om het volgende woord, maar om het volgende *token*. Het volgende woordbrok dus.

> *Tokens.* Daarover zegt OpenAI zelf: "... OpenAI processes text by breaking it down into tokens. Tokens can be words or just chunks of characters. For example, the word "hamburger" gets broken up into the tokens "ham", "bur" and "ger", while a short and common word like "pear" is a single token. Many tokens start with a whitespace, for example " hello" and " bye". ..." [uit: learn.microsoft.com/en-us/azure/cognitive-services/openai/"

OpenAI (dat GPT-4 ontwikkelt en exploiteert) krijgt het voor mekaar om het teksten te laten lezen en in brokstukken te laten hakken en tegelijkertijd betekenissen vast te laten leggen die het later weer kan vinden en gebruiken om nieuwe teksten te maken — teksten die dezelfde betekenis op verschillende manieren verwoorden. *Dat het werkt kon ik zelf controleren.* Maar hoe OpenAI dat doet is geheim.

Schets van de werking van GTP-4. Opnieuw Mo Gawdat in de vlog, door een bot vertaald en door mij enigszins gepolijst: “Trouwens, de code van een transformer in een GPT is tweeduizend regels lang; het is niet erg ingewikkeld; het is eigenlijk geen erg intelligente machine; het voorspelt simpelweg het volgende woord in een zin in wording. Oké? En veel mensen begrijpen niet dat je ChatGPT zoals het nu werkt allang kent. Dat als je in Amerika bent en je je kind alle namen van de staten en de Amerikaanse presidenten leert en het kind ze dan opdreunt — en je doet dan van ’oh mijn god, dit is een wonderkind!’ ... Niet heus, natuurlijk. Het zijn de ouders die proberen hun kind een wonderkind te laten lijken door het wat betekenisloze rotzooi uit het hoofd te laten leren en op te laten zeggen. Maar, als je erover nadenkt, is dat precies wat GPT doet. Het enige verschil is dat het in plaats van alle namen van de staten en alle namen van de presidenten te onthouden, het miljarden en miljarden en miljarden pagina’s tekst las en de betekenis ervan onthoudt. En daaraan voegt het een ongelooflijk stukje intelligentie toe, waardoor het een bepaalde inhoud op dezelfde manier kan laten zien als Shakespeare die zou hebben opgetekend, weet je, het heeft die ongelooflijke mogelijkheden om de nuances van Shakespeare’s taalgebruik te voorspellen.”

Hoe het betekenissen uit teksten oogst en ze opslaat en beschikbaar houdt staat nergens. Daarnaar kan ik dus alleen maar raden. En misschien geldt dat ook voor de medewerkers van OpenAI zelf, maar daarover later.

Trouwens, ik heb wel een indruk van wat er voor nodig is om GPT-4-diensten aan te bieden. Zdnet van 19 mei 2023 geeft een indicatie van de hardware-investeringen die Microsoft ten behoeve van OpenAI deed. Die zijn extreem.

Over de supercomputer van OpenAI. Opnieuw: vertaald en bijgeschaafd — je moet wel door het advertentie-achtige karakter van de tekst heen lezen ... : "Microsoft-functionarissen zeiden dat ze de op vier na krachtigste openbaar geregistreerde supercomputer (zoals gerangschikt op de TOP500-supercomputerlijst) hebben gebouwd in samenwerking met en exclusief voor OpenAI. […] Microsoft zei dat de supercomputer gebouwd voor OpenAI een enkel systeem is met meer dan 285.000 CPU-cores; 10.000 GPU’s en 400 gigabit per seconde netwerkconnectiviteit voor elke GPU-server." https://www.zdnet.com/article/microsoft-builds-a-supercomputer-for-openai-for-training-massive-ai-model

Toch, OpenAI slaat de gelezen teksten niet op en maakt er ook geen afschriften van. Dáárvoor gebruikt het al die ruimte en rekencapaciteit niet. Maar waarvoor gebruikt het al die opslag- en rekencapaciteit dan wel? Dat zijn twee dingen: *(1)* het maken (trainen, fine-tunen) van het model en *(2)* het gebruiken van (chatten met, stellen van vragen aan) het model.

Is hier het antwoord te vinden op de vraag waarom OpenAI een geheim maakt van hoe GPT-4 zijn wonderbaarlijke gaven realiseert?

Geen vermenigvuldiging van teksten. OpenAI zegt, in een spiegelgevecht met mogelijk auteursrechtelijke verwijten: “GPT does not copy or store training information in a database. Instead, it learns about associations between words, and those learnings help the model update its numbers/weights. The model then uses those weights to predict and generate new words in response to a user request. It does not “copy and paste” training information — much like a person who has read a book and sets it down, our models do not have access to training information after they have learned from it.” [https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed]

De vraag roept bij mij het beeld op van eindeloze rijen onderling verbonden en parallel geschakelde *GPU's* (*grafische processing units*), die samen het gigantische, vooraf getrainde, neurale netwerk van GPT-4 vormen en beschikbaar maken, het mechanische brein dat onder meer de teksten van Shakespeare leest en weer vergeet na ze in *tokens* opgedeeld te hebben, de gewichten van verbindingen met buur-tokens te hebben bijgewerkt – alles om het volgende token van een antwoord op een vraag te kunnen genereren (gedurende trainingen, maar ook gedurende het gebruik). *Maar wat zijn GPT-4-betekenissen en hoe worden ze door GPT-4 gebruikt?*

Ik koester de theorie dat lokale netwerkstructuren die betekenissen (de GPT-4-betekenissen) *zijn*. Dat als antwoord op een vraag een volgordelijke rij tokens (een "frase”) wordt opgebouwd uit het getrainde neurale netwerk, waarin de tokens met hun parameters niet volgordelijk maar parallel toegankelijk zijn. Een inzicht dat spoort met de ervaring die het voorlezen van herhaald in random volgorde geplaatste losse woordbrokken bij ons oproept: soms heeft zo'n frase betekenis. Betekenis van een frase *is* de volgorde van zijn tokens, anders gezegd, van zijn tokens in volgorde. En de kunst van GPT-4 is de waarschijnlijkheden van betekenisvolle tokenvolgordes in een netwerk te onthouden, bij te werken en aan te bieden. Het gebruikt daarvoor invullingen van de zo genoemde *transformer* architectuur.

Ik vroeg GTP-4 wat dat is:

Transformers. ***AS****:* is a transformer a hardware thing in GPT-4? **ChatGPT:** In the context of GPT-4, the term "transformer" refers to a specific type of neural network architecture rather than a hardware component.
I summarize it from input to output: *input* to the transformer is a sequence of tokens, such as words or subwords, that make up the input text; *embedding*: a high-dimensional vector representation that helps capture the semantic meaning of each token; *positional encoding*: this encoding incorporates the token's position in the sequence; *encoder layers*:

imultiple identical layers, containing two sub-layers: a self-attention mechanism (allowing to focus on different parts of the input sequence when processing each token) and a *feed-forward* neural network (to help further transform the representations); *decoder layers*: multiple identical layers, including an additional sub-layer, an attention mechanism over the encoder's output allowing the decoder to access the encoded information; *output*: a probability distribution over the possible tokens at each position in the sequence. This output can be used for tasks like language generation, machine translation, or text classification. (Ik vatte ChatGPT's antwoorden samen)

En ik verbeeld me dat mijn eigen brein, inmiddels meer dan 70 jaar getraind met indrukken uit mijn omgeving, iets soortgelijks doet als GPT-4, wanneer het een antwoord op een vraag opbouwt, niet alleen in een tweegesprek, maar ook, bijvoorbeeld, als ik in mijn geheugen zoek naar de weggezakte naam van een collega van 40 jaar geleden — en die naam langzaam aan, stukje bij beetje en in halve brokstukken naar boven komt drijven om met wat halfbewuste *trial & error* ineens een sluitend geheel te vormen, voldoende om te worden uitgeserveerd. Een resultaat dat me dan niet alleen een goed gevoel geeft, maar me ook doet vermoeden dat de tokens in mijn hersenpan kleiner zijn dan namen of hele woorden.

Mijn theorie is dan dat GPT-4 op een soortgelijke manier naar kandidaten voor het volgende token in een antwoord zoekt in zijn neurale netwerk als ik in mijn brein. En dat wat mensen aan hun onderbewuste plegen toe te vertrouwen door OpenAI aan de *transformers* van GPT-4 wordt toevertrouwd.

Dit is allemaal beeldspraak omdat niemand weet hoe het menselijk brein processueel met betekenissen omgaat, bewust of onbewust. Er zijn dus heel wat theorieën over. De mijne is gebaseerd op het lezen van een paar boeken.[3]

Dit maakt ook mijn theorie vooralsnog noodzakelijkerwijze niet meer dan een vermoeden. Toen ik delen ervan dan maar voorlegde aan ChatGPT kreeg ik drie zinvolle reacties. De eerste behelsde dat GPT-4 "kan begrijpen wat je bedoelt met betrekking tot de betekenis van een frase en het parallelle toegankelijk zijn van de tokens met hun parameters", aangevuld met "Het model maakt naast volgordes ook gebruik van contextuele informatie en leerpatronen om betekenisvolle verbanden tussen tokens vast te leggen." Dit is overigens in overeenstemming met wat ik aangaf bij *Transformers*.

[3]Lon L. Fuller, *The Morality of Law*, Yale University Press, 1976; Philip Bobbitt, *Constitutional Fate: Theory of the Constitution*, Oxford University Press, 1984; Mary Douglas, *Risk and Blame,* Routledge, 1993; Duncan Watts, *Six Degrees: The science of a connected age*, Norton, 2004; Noam Chomsky, *What kind of creatures are we?*, Columbia University Press, 2016; Robert Sapolsky, *Behave: The Biology of Humans at Our Best and Worst*, Penguin, 2017; Michael Gazzaniga, *The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind*, Farrar, Straus and Giroux, 2018.

Ten tweede, toen ik vroeg of het model zich aanpast tijdens het gebruik antwoordde het van niet: “Nee, het GPT model breidt zich niet uit tijdens het gebruik. Omvang, structuur en het aantal parameters blijven dan gelijk als het getraind is en voor gebruik wordt aangeboden.” En “Uitbreiding van het model vergt training (initieel of fine-tuning).” Dit betekent dat "het model" (zeg: GPT-3.5 of GPT-4) als het diensten verleent (wordt gebruikt) statisch is en als het wordt getraind geen diensten verleent. Dit is van belang bij het beoordelen van regulering via copyrights, maar ook via de letterlijk vandaag aanvaarde EU ”AI-wet." Ten derde merkte het op:

> “hoewel je beeld een intuïtieve voorstelling is, moeten we ons bewust zijn van de complexiteit van de interne werking van GPT-modellen. Hoewel ze gebruikmaken van transformers en parallelle verwerking, bevatten ze ook vele andere componenten en mechanismen die bijdragen aan hun functionaliteit. Het begrijpen van de exacte werking en representaties van dergelijke modellen blijft een actief onderzoeksgebied.”

Omdat ik mijn hele leven in beide kampen heb doorgebracht herken ik hierin de oorzaak voor onbegrip van wat ik nog *alfa* en *beta* domeinen noem. De typische *alfa* denkt dat het verder begrijpen van de exacte werking van een ding weinig toevoegt aan werkelijk inzicht; wat de *alfa* weet is voor de *beta* een intuïtie – terwijl de typische *beta* denkt dat het begrijpen van de exacte werking van een ding werkelijke kennis *is*, daarmee kun je het immers maken, besturen, gebruiken, exploiteren; maar wat de *beta* weet is voor de *alfa* geen werkelijk inzicht. We gaan deze tegenstelling nog tegenkomen. En ik zal mijn intuïtie nodig hebben – wij zullen allemaal onze intuïties nodig hebben, ook de *beta*'s. Zij weten immers óók niet alles.

Als het zo is dat ik met behulp van mijn brein betekenissen uit de teksten van die boeken viste en verwerkte in dit stuk, houdt dat dan in (met een knipoog naar Chomsky) dat ik door dit stuk te schrijven “*basically low-tech plagiarism*" pleeg? En wat, als ik dit stuk door ChatGPT-4 had laten schrijven?

High-tech plagiaat?

Als het zo is dat ik met behulp van mijn brein betekenissen uit teksten vis, of GPT-4 met diens neurale netwerk ze eruit vist, en we verwerken dat in nieuwe teksten, houdt dat dan het schenden van auteursrecht in? In de hoop er een antwoord op te vinden raadpleeg ik het aangehaalde "Robotkunst en auteursrecht" van Dirk Visser dat in zijn tweede zin belooft een overzicht te geven

van auteursrechtelijke en aanverwante vragen die deze programma's (bedoeld wordt *DALL-E, Stable Diffusion* en *ChatGPT* die alle drie een invulling zijn van de transformer architectuur van GPT-4) oproepen. Ik zet de vragen die Visser inventariseert op een rij. Het zijn er 10.

> 1. Worden de programma's (die met een transformer architectuur) zelf beschermd? 2. Is robotkunst beschermd? 3. Bij wie ligt de bewijslast van menselijke inbreng? 4. Wat zijn de gevolgen? 5. Wanneer maakt een individueel 'robotkunstwerk' inbreuk op auteursrechten van anderen? 6. Wordt er in het (voorbereidende) samenstellingsproces inbreuk gemaakt door de robot op de auteursrechten van anderen? 7. Is er sprake van toegestane 'tijdelijke reproductie' als voor 'tekst- en datamining'? 8. Worden 'naburige' rechten op databanken en perspublicaties geschonden? 9. Welk rol spelen wetenschappelijke en maatschappelijke zorgvuldigheidsnormen? 10. Is de bescherming tegen stijlnabootsing relevant?

Terwijl ik deze vragen uit het artikel destilleerde merkte ik bij mezelf dat ik op sommige stukken verschillend, vaak intuïtief reageer en dat die verschillen rusten op de rol die ik aanneem, als *alfa* of als *beta* – ik kan ze immers beide spelen. Ik neem me voor verder alleen die vragen onder de loep te nemen die tot de bedoelde innerlijk tweestrijd leiden.

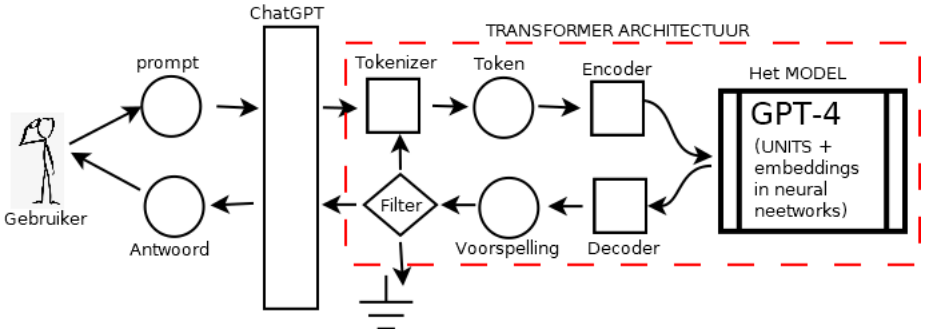
1. Worden de programma's zelf beschermd? Als *alfa* bevestig ik die vraag zonder verder na te denken. We weten al sinds de jaren 1990 dat computerprogramma's auteursrechtelijk beschermd zijn.

Als *beta* denk ik er natuurlijk ook zo over, alleen denk ik niet dat ChatGPT, als programma, van veel betekenis is voor het auteursrechtelijke debat.[4] Het is wel een programma, maar het doet niet waar het debat over gaat, het is voornamelijk functioneel als doorgeefluik, als koppelvlak tussen de gebruiker en GPT-4. Volgens mij gaat het qua auteursrecht om de complexe automatiseringsdiensten waarin neurale netwerken, modellen, API's en computerprogramma's samen de rollen spelen die de transformer architectuur verwezenlijken,[5] en ik ken geen collega die deze samen als computerprogramma ziet waarop auteursrecht zou kunnen rusten.

[4] Voor mij is ChatGPT een programma dat gebruik maakt van een API om de communicatie te bemiddelen tussen de gebruiker en de transformer architectuur waarin GPT-4 centraal staat. Het is overigens onzeker of er overeenstemming bestaat over wat er wordt bedoeld met termen als ChatGPT en GPT-4 – ook omdat het allemaal zo nieuw is. Ik ontleen houvast aan uitingen van OpenAI dat deze diensten op de markt beschikbaar stelt, en aan de voortreffelijke voorpublicatie van *Developing Apps with GPT-4 and ChatGPT* door Olivier Caelen en Marie-Alice Blete, volgens de website van de uitgever (sic!) *Released January 2024 Publisher(s): O'Reilly Media, Inc. ISBN: 9781098152468.*

[5] Het idee dat er auteursrecht op het geheel zou kunnen rusten lijkt me, ook als *alfa*, absurd. GPT-4 is geen computerprogramma maar een ingevulde transformer architectuur. Die architectuur is overigens openbaar en beschreven in een door acht Googlemedewerkers in 2017 geschreven artikel dat "Attention is all you need" heet en inmiddels 77.926 keer is geciteerd.

GPT-4 wordt doorgaans een *model* genoemd. Een model dat op twee verschillende manieren bestaat voor gebruik. Een model dat door OpenAI in een bepaalde toestand wordt bevroren om die vervolgens via ChatGPT als chatbox aan gebruikers ter beschikking te stellen (figuur 1) maar ook een model dat door OpenAI wordt losgekoppeld van de chatbox voor het samenstellingsproces, voor de opbouw, uitbreiding, training en fine-tuning van het model (figuur 2). De auteursrechtelijke vragen betreffen wat er tijdens en voorafgaand aan het gebruik van de transformer architectuur gebeurt, als chatbox en in het voorbereidingsproces.



Figuur 1: **Als chatbox**. Waar kunnen de verantwoordelijkheden liggen? Drie mogelijkheden: de *gebruiker* die de vragen stelt, *OpenAI* die verantwoordelijk is voor wat het (gefixeerde) model doet en nogmaals *OpenAI* dat GPT-4 maakte. Waar is sprake van "werken van wetenschap of kunst?" Vijf mogelijkheden: een prompt (een stukje tekst dat ChatGPT als inbreng van de gesprekspartner ziet) , een antwoord (een stuk tekst dat ChatGPT uitserveert), een dialoog, een programma (als het koppelvlak en de Tokenizer) en het Model.

Iedereen is het erover eens dat aanwijsbare programma's als ChatGPT (opgevat als API) en de algoritmen die neurale netwerken trainen beschermd zijn. Het ligt voorts voor de hand dat op prompts auteursrecht rust en dat discussie bestaat over of op dialogen auteursrecht kan rusten en zo ja aan wie dat dan toekomt; ik denk aan de maker van de prompts.

Tot verbazing van de *beta* in me (het Model bevat überhaupt geen tekst), gaat het *alfa*-debat óók over de vraag of *het Model* een inbreuk kan maken op een "echte" tekst van wetenschap of kunst — maar niet over de vraag of het zelf een auteursrechtelijk beschermd werk *is*. Daarover later.

2. Is robotkunst beschermd? De *alfa* zegt: "Het antwoord luidt vermoedelijk 'nee, tenzij'. Een door een computerprogramma gegenereerde tekst of afbeelding (het 'antwoord' in figuur 1) geeft geen blijk van creatieve menselijke keuzes. Dat is zelfs een van de wezenlijke kenmerken ervan." Voor

een doorsnee *beta* is de vraag onverwacht, om te beginnen omdat het bij GTP-4-output zelden gaat om kunst; meestal gaat het om kennis: de hoofdmoot van de exploitatie gaat over het produceren van antwoorden op vragen in natuurlijke taal. De resultaten worden beoordeeld op doeltreffendheid en begrijpelijkheid en niet op hun artistieke waarde.

Overigens voel ik me als *beta* onderschat door de *alfa-kwalificatie* van GPT-4 als "programma" (zie hierboven) en ook door het "de gegenereerde tekst of afbeelding geeft geen blijk van creatieve menselijke keuzes." Dit miskent immers dat er inmiddels een hele industrie (opleidingen, designers) is ontstaan rond de kunst van het ontwerpen van "prompts," de met creatieve menselijke input ontworpen teksten waarmee de gebruiker GTP-4 stuurt in de richting van een kwalitatief bruikbaar antwoord. In die zin is elk GPT-4 antwoord (zo u wilt: alle robotkunst) altijd mede gestuurd door creatieve menselijke inbreng van de gebruiker (de prompt).

3. Bij wie ligt de bewijslast van menselijke inbreng? Voor de *alfa* is dit een groot probleem omdat je aan de output niet direct kunt zien dat het jouw creatieve inbreng is die ertoe geleid heeft. Voor de *beta* is het vermoedelijk geen probleem: hij heeft (in beginsel) auteursrecht op de prompt en als hij de invloed van zijn creatieve invloed op de output zou willen bewijzen kan hij dat eenvoudig doen door dezelfde prompt herhaaldelijk aan te bieden en te loggen wat het model oplevert.

4. Wat zijn de gevolgen? Visser zegt "Met deze vraag raken we aan het fundament van het auteursrecht dat er is om het maken van kunst te erkennen, te belonen en aan te moedigen. *Wordt dat fundament aangetast als robots iets ook of beter kunnen, zonder erkenning, aanmoediging of beloning?*" De vraag (mijn cursivering) is cruciaal omdat de mogelijkheid lijkt te zijn aanvaard dat robots ongevoelig zouden kunnen zijn voor *erkenning, aanmoediging of beloning.* Dat aspect komt in de volgende twee hoofdstukken aan de orde.

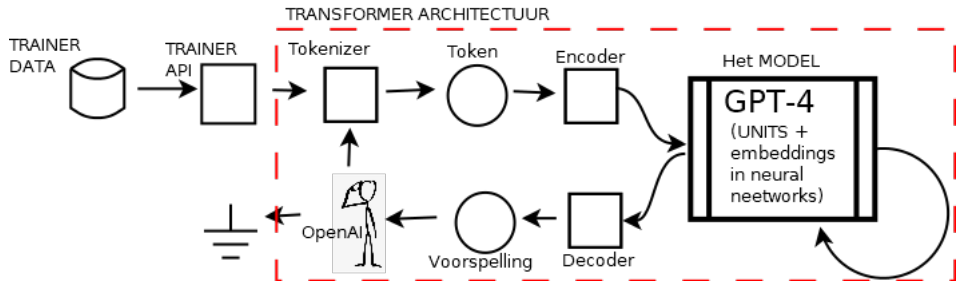
Intussen denk ik dat de output van GTP-4 eerder aanmoediging verdient als het een kunstwerk beoogt te wezen dan als het een inhoudelijk antwoord is op een inhoudelijke vraag. Gelet op de overweldigende explosie van bedrijvigheid in het kennisdomein (chatbots die de diensten van ChatGTP gebruiken voor commerciële communicatie) is daar voorlopig ook geen enkele auteursrechtelijke aanmoediging nodig.

Bovendien past in het traditionele auteursrecht-denken het gevoelen dat de incentive voor kunstenaars om werk te maken dat zich onderscheidt van

het werk van anderen (GPT-4 incluis) zeer wenselijk is. Maar de gedachte dat dit voor de kennissector (ik denk aan het befaamde *staan op de schouders van reuzen*) minder nodig zou zijn sluit niet aan op heersende positiefrechtelijke opvattingen.

5. Wanneer maakt een individueel robotkunstwerk inbreuk op auteursrechten van anderen? Ik vertaal dit naar: wanneer maakt GPT-4 output inbreuk op de auteursrechten van anderen? De *beta* in me denkt dat er door GPT-4-output geen inbreuk gemaakt *kan* worden omdat *het GPT-4-model* in de gefixeerde gebruiksstand helemaal niet beschikt over een afbeelding van enig origineel (zie figuur 1, waarin de inhoud van het model bestaat uit door training gevormde *units* en hun *parameters*).[6] Als *alfa* vind ik het ook moeilijk te geloven dat het model geen afbeelding van een origineel bevat maar laat me door mijn *beta*-kant overtuigen (mede door de eerste hoofdstukken in het boek van Caelen en Blete).

6. Wordt er in het (voorbereidende) samenstellingsproces inbreuk gemaakt door de robot op de auteursrechten van anderen? Omdat de voorbereiding twee verschillende fasen kent leidt de vraag naar twee figuren, te weten figuur 2 (training) en figuur 3 (data mining).



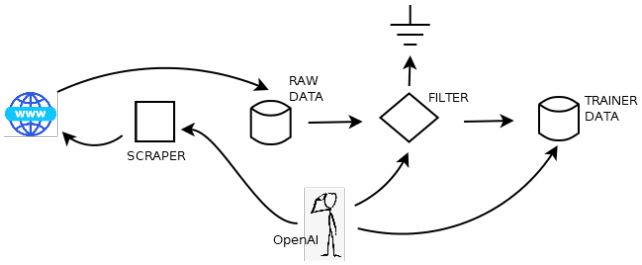
Figuur 2: **Training van het Model**. Waar kunnen tijdens de training de verantwoordelijkheden liggen? Twee mogelijkheden: de *leverancier* van de trainer data en *OpenAI*, als verantwoordelijke voor het maken en onderhouden van de invulling van de transformer architectuur. Waar is sprake van "werken van wetenschap of kunst?" Drie plaatsen zijn relevant: de trainer data, de menselijke feedback bij fine-tuning en het model.

Training. Wanneer de bouwer eenmaal beschikt over data om het model mee te trainen biedt hij die als één lange tekst aan, aan de trainer API die voor de training intermedieert tussen de trainer data als input voor de

[6]Units zijn tokens, maar ook nieuw gevormde knopen (inclusief hun embeddings) in de verborgen lagen van het via deep learning gevormde netwerk dat op zoek is naar patronen en verbanden in de data.

transformer architectuur. Als eerste wordt die input opgehakt in tokens die worden gevoerd aan neurale netwerken die er 'hidden layers' met nieuwe units aan toevoegen, met 'embeddings' met stochastische gewichten die patronen vastleggen (de 'parameters'). Dit verbeeld ik in figuur 2 met de verwijzing van het model naar zichzelf. Bedenk wel: ik probeer een functionele weergave te geven van iets waarvan ik de *beta*-kant niet goed ken — een beetje zoals ik enkele onderdelen onder de motorkap van mijn auto kan benoemen zonder de werking ervan te kennen. Maar ik denk dat dit voldoende is om een belangrijk auteursrechtelijk punt op te heideren: het model is geen kopie noch een bewerking van enig werk dat onderdeel is van de binnenkomende stroom van tokens waarin die werken werden opgedeeld en verder bewerkt. Ik denk ook dat dit beeld voldoende is om te kunnen vaststellen dat het model zelf een 'werk van wetenschap' kan zijn, met name als het (bij voorbeeld voor het bewaren van een back-up) op media is vastgelegd. Wat soms inderdaad mogelijk is.

Data mining Ik ga ervan uit dat de bouw van elk groot taalmodel (LLM), dus ook van GTP-4 daarmee begint. De bouwer instrueert een *scraper*, een computerprogramma dat het internet afgraast naar teksten, en krijgt daarvoor een bestand met ruwe teksten terug. Dan filtert de bouwer die en houdt een bestand over dat geschikt is om het model mee te trainen. Maar alvorens dat proces te bezien stel ik vast dat er auteursrechtelijke regels kunnen zijn die deze vorm van data-mining reguleren. Daarbij kan een rol spelen of de digitale tekstbestanden eventuele scrapers waarschuwen om hen over te slaan. Opgemerkt zij dat er auteursrechtelijke vragen zijn die nog beantwoord moeten worden met name door de transnationale effecten van internet en de grote verschillen tussen de US en de EU op dit terrein.



Figuur 3: **Data mining**. Waar kunnen tijdens de training de verantwoordelijkheden liggen? Twee mogelijkheden: wie teksten beschikbaar maken op het web en OpenAI dat verantwoordelijk is voor het scrapen van het internet, het filteren van de data het beschikbaar stellen als trainer data. Op alle teksten kunnen auteursrecht rusten.

7. Is er sprake van toegestane 'tijdelijke reproductie' als voor 'tekst- en datamining'? Deze vraag had hierboven al aan de orde kunnen zijn gesteld

waar van *alfa*-zijde twijfel werd gesuggereerd over of het token-gewijs invoeren van tekst in de transformer-architectuur geen auteursrechtelijke inbreuk maakt. De formulering van de uitzondering voor tijdelijke reproductie betreft reproductiehandelingen die "van voorbijgaande of incidentele aard zijn, en die een integraal en essentieel onderdeel vormen van een technisch procedé en die worden toegepast met als enig doel: […] een rechtmatig gebruik van een werk of ander materiaal mogelijk te maken, en die geen zelfstandige economische waarde bezitten," een formulering die voor GPT-4-modellen gemaakt had kunnen zijn en daarmee, aldus mijn *alfa*-gemoed ook op hen van toepassing is. Daarnaast heeft de EU destijds bovendien besloten om ruimte te bieden voor de economisch en wetenschappelijk zo betekenisvolle onbelemmerde tekst- en data-mining. Ook deze exceptie wijst mijns inziens ondubbelzinnig in de richting van de mogelijkheid van legitieme training van GPT-4-modellen op teksten die via data-mining zijn verkregen. Voor anderen, als Visser, is dit nog geen uitgemaakt zaak. Hij oppert zelfs "Een volgende vraag is of rechthebbenden nu massaal hun auteursrecht gaan voorbehouden, waardoor commerciële tekst- en datamining misschien sowieso niet mag naar EU-auteursrecht." Als *beta* interesseert me dit eigenlijk niet omdat het een activistische suggestie is die wel onder ogen wordt gezien als het zover is. En bovendien en vooral: omdat het doel van de training is om digitaal een taal te leren en niet om teksten te kopiëren of te bewerken.

9. Welk rol spelen wetenschappelijke en maatschappelijke zorgvuldigheidsnormen? alfa's en *beta's* zijn het doorgaans erover eens dat de schaal van opzettelijke misleiding, waartoe de nieuwe GPT-4-modellen de hulpmiddelen bieden, ongewenst is en de risico's van maatschappelijke ondermijning verhogen. *De vraag is de belangrijkste van allemaal, maar geen kwestie van auteursrecht.* Natuurlijk ligt bij de beantwoording ook hier de confrontatie met de grondrechten op de loer.

Tot zover de vragen en de antwoorden van Visser. Ik lees eruit dat het Nederlandse auteursrecht geen effectieve belemmeringen opwerpt tegen de schrijvende mens noch tegen het genereren van teksten via GPT-4 architecturen.

Wanneer we zoeken naar effectieve juridische instrumenten tegen bedreigingen die hartenkreten opwekken als *a profound change in the history of life on Earth!* en *further development should be planned for and managed with commensurate care and resources* is het auteursrecht misschien niet het

geëigende instrument, en zeker niet het enige. Daarvoor is het auteursrechtelijke gevaar te gering voor een traditionele bedrijfstak die naar verwachting binnen die bedrijfstak wel tot de benodigde innovatie zal kunnen komen om te overleven.

De pointe van het voorafgaande beeld van GPT-4 in actie is dan dat het een basis legt voor de analyse van waarom velen AI wel in verband brengen met kunst, kennis en ontzag, maar niet met macht, burgerschap en *the rule of law*. En hiermee zit ik midden in het probleem (de loskoppeling van macht en verantwoordelijkheid.) dat Mo Gatwat aanwijst en pal voor de vraag (... als robots iets ook of beter kunnen, zonder erkenning, aanmoediging of beloning?) van Visser die ik voor me uit schoof.

Maatschappelijke uitdagingen × macht en verantwoordelijkheid

In de eerste alinea van zijn artikel verzucht Visser: *iedereen wordt nu geconfronteerd met het potentieel vervreemdende, misleidende of parasitaire karakter van deze technologie en de mogelijke bedreiging voor echte* sc*hrijvers en kunstenaars.* Dit citaat speelt meer de rol van *teaser* dan van *abstract* en doet mijns inziens geen recht aan de afgewogen stijl die de wetenschappelijke bijdrage doorgaans onderscheidt van de activistische.

Maar nu het er eenmaal staat − het toont wel een sentiment dat breed wordt gedeeld door juristen. Dat sentiment gaat uit van de morele wenselijkheid van rechtsbescherming, in dit geval van traditionele kunstenaars in competitie met innoverende techniek, maar niet alleen van hen. Die beschermingsimpuls vindt een politiek tegenwicht in een andere brede en breed gedeelde impuls: in de roep om vooruitgang, die de creatieve destructie als noodzakelijk bijverschijnsel voor lief neemt. Voor elk type verworvenheid, voor elk type vooruitgang, of die nu economisch, politiek, sociaal of wetenschappelijk is bestaan beide impulsen (reactionair × progressief). Op al die gebieden circuleren slogans die juist die twee sentimenten vertegenwoordigen, slogans waaraan een medemens zich kan binden, waarmee de geleerde en de artiest kunnen experimenteren, waarmee een koopman zich kan verrijken en die een regent kan exploiteren. Alle vier (medemens, geleerde/artiest, koopman, regent) kennen collectieve sentimenten als uitingen van belangrijke krachten in hun arena's. En het zijn (om Cicero's Rhetorica vrijmoedig te parafraseren) die sentimenten en hoe ze worden gemanipuleerd die de toekomst

van een jurisdictie bepalen. ChatGPT zegt het desgevraagd[7] zelf: *'Cicero beschouwt de sentimenten van het volk in zijn Rhetorica als van cruciaal belang voor het behoud en de stabiliteit van een staatsbestel.'*[8]

Wie om zich heen kijkt ziet dat sentimenten belangrijke politieke wapens zijn. Dat is al millennia zo. Daarmee rijst het vermoeden dat als de communicatietechnologie exponentieel verbetert, de context ontstaat waarin de behoefte aan het aansturen ervan een evenredige groeistuip doormaakt. Daarvoor is recht nodig. Het is dan ook geen wonder dat toen de boekdrukkunst van de grond was gekomen, zeg na 1650 ongeveer, de *regent* naar meer regels, de *koopman* naar meer markt, de *artiest* naar bredere inspiratie en de medemens naar meer dan alleen godsvruchtig leesvoer gingen grijpen; en dat, nadat de elektronische informatiedragers en informatiespelers van de grond waren gekomen, zeg na 2000, de manieren veranderden waarmee de censuur, de markt, de creativiteit en de informatiehonger vorm kregen – ik denk, vrij associërend, aan de Chinese firewall, aan Google, aan Trump's tweets en aan een treincoupé vol reizigers die allemaal zwijgend op hun telefoon sociaal zitten te wezen; en dat we nu, anno 2023, gespannen uitkijken naar hoe die vormen verder zullen veranderen onder de invloed van GPT-4-modellen.

Geen wonder dat de vraag voorligt of door GPT-4 die vormen zullen leiden tot *a profound change in the history of life on earth that should be managed with commensurate care and resources.* Het antwoord ligt in de toekomst en die is moeilijk te voorspellen. Maar ik kan wel, mijn *alfa*- en *beta*-kant combinerend, er een gooi naar doen door na te denken over wat GPT-4 in die vier domeinen gaat betekenen.

Ik probeer dat luchtig te doen, maar ik ben geen lachebekje en focus op het Nederland van 2023 met de bijbehorende context. Dat wil zeggen: in het kielzog van gaswinnings-, toeslagen- en Ter Apel-affaires, van de Covid-19 verwerking, van de stikstofdiscussie en van de recente verkiezingsuitslagen – en dat alles in de slagschaduw van het EU-lidmaatschap, de oorlog in Oekraïne, de toestroom van vluchtelingen en de door de uitstoot van CO_2 bevorderde opwarming van de aarde. Hoe gaat de ontwikkeling van GTP-4 de sentimenten in die vier domeinen — en daarmee de richting van verkozen reguleringsinstrumenten — beïnvloeden? Ik houd het kort.

[7] Mijn prompt luidde: "Kun je me in één zin schetsen hoe belangrijk Cicero de sentimenten van het volk in zijn Rhetorica acht voor het voortbestaan van een staatsbestel?"

[8] 🙂

1. Regenten en GPT-4. Vandaag zijn dat bijvoorbeeld Rutte, ministers, Kamerleden, magistraten en leidinggevende ambtenaren. Ze moeten het collectieve belang dienen, maar willen tegelijkertijd dat collectief beschermen tegen innovaties die politieke instabiliteit zou inluiden, een tweeledig doel dat veel middelen heiligt, al kan dat niet altijd worden uitgesproken; *anyway*, ze geven er blijk van te geloven dat de leugen om eigen bestwil hun integriteit niet aantast. Syri is bijvoorbeeld door de rechter verboden, maar bestuursdiensten blijven ernaar handelen (Argos onthulde in 2022 dat onder andere gemeenten, Belastingdienst, Toeslagen, UWV, SVB en politie samenwerken binnen de LSI en risicosignalen en privacygevoelige informatie delen); ook bij de behandeling van vluchtelingen overheersen het wantrouwen en de overtuiging dat veel grondrechtelijk dubieuze middelen geoorloofd zijn. Wat verwacht ik dat onze huidige regenten onder dit gesternte met GPT-4 gaan doen?

Onze regenten gaan inzien dat het gebruik van GPT-4-modellen een dramatische verbetering qua snelheid, doeltreffendheid en consistentie van de besluitvorming bij de publieke dienstverlening zou betekenen – alleen de motiveringen ontbreken, of liever, blijven abstract, maar dat is toch al onderdeel van de bestuurlijke praktijk. Eén en ander kan leiden tot het plan een *dedicated* versie van GPT-4 te maken waarin alle gedigitaliseerde bestuurlijke archieven worden ingevoerd, in tokens opgehakt, getraind en aan bestuurders beschikbaar gesteld.

Dat zou ook een gevaarlijk wapen in handen van regenten leggen die weinig van IT weten, die er inmiddels aan gewend zijn de grondrechten met een korrel zout te nemen en er de voorkeur aan geven in abstracties (kans op fraude) in plaats van feiten (daad en dader) te denken. Iets dat perfect wordt ondersteund door GTP-4 modellen die geleerd is om geen antwoorden te geven die de privacy kunnen schenden. Ik sluit niet uit dat we – als we niet gered worden door de EU – na verloop van tijd de PRC op het gebied van de grondrechten rechts gaan inhalen met die eigen GPT-4-bestuursdienst, waarvan de gebruikende ambtenaar immers de antwoorden op vragen direct kan terugkoppelen naar de persoon waarover de vraag wordt gesteld. En ik ben er niet zeker van dat de huidige formulering van de burger- en grondrechten daartegen voldoende beschermt. Ik verwacht, kortom, dat onze regenten gaan inzetten op het beschermen van de medemens tegen de risico's van GPT-4 op

het EU-niveau van de AI-wet, terwijl ze tegelijkertijd voor ongeremde vernieuwing gaan van de publieke administratie en dienstverlening met behulp van een groot, eigen GPT-4-model.

Ik verwacht ook dat daarmee het wantrouwen bij de medemens exponentieel toeneemt en de treincoupés vol komen te zitten met mensen die hun sores verdringen door de volgende aflevering van hun favoriete complottheorie door ChatGPT-4 te laten maken en uitserveren op hun telefoons.

Ik aarzel, en voel enige ironie opborrelen, weet eigenlijk niet of ik dat nu wel of niet als een *profound change in the history of life in the Netherlands* zou ervaren.

Maar in ernst concludeer ik dat we als samenleving alert in de weer moeten met het robuust maken tegen GTP-4-inbreuken van de mensenrechten en de grondrechten, en met name van de condities in excepties daarop. Maar ik ben bang dat ons politieke bestel dat wel, maar onze politieke cultuur die uitdaging niet aankan. Voor regenten zijn de vrijwaringskansen die het verschansen achter GTP-4-ondersteunde besluitvorming biedt, te groot. En hiermee ben ik terug bij Mo Gatwat en diens grootste zorg. En aan die zorg verleen ik voorlopig voorrang.

We moeten aan de bak; vluchten kan niet meer

Soit. Ik was van plan een hele lijst te maken, van *2. Kooplui en GPT-4* (mededingingsrecht en vrijheid van expressie?) via *3. Geleerden/artiesten en GPT-4* (vrijheden van onderzoek en expressie?) naar *4. Medemensen en GPT-4* (vrijheden van vergadering en vereniging?) Maar het werd te veel. Te veel voor mijn denkvermogen en, als er dan toch iets kwam, teveel voor een enkel artikel.

Ik pak dus de draad weer op bij het beeld van een met digitale bronnen gevoede en getrainde GovGPT-NL. Als bronnen denk ik bijvoorbeeld aan universiteitsbibliotheken, de KB, het Rijksarchief, de bestuurlijke gedigitaliseerde postdiensten, rechtspraak.nl en open internetbronnen. Het eerste dat een jurist zal opmerken is iets in de geest van "maar dat mag niet," denkend aan auteursrechten, rechten op bescherming van de persoonlijke levenssfeer en misschien ook op de belemmeringen die zouden worden opgeworpen door een vermeend niet-kunnen-waarmaken van motiveringsverplichtingen en de verplichtingen uit de Woo. Maar is dat ook zo? Dat zijn vier suggesties die alle vier serieuze aandacht verdienen van de wetgever en dus van een gedegen

publieke discussie.

Visser's bijdrage over het auteursrecht is daartoe een aanzet maar schiet inhoudelijk tekort. Hij suggereert dat GPT-4 bij zijn antwoorden op vragen de werken uit het trainingsmateriaal kopieert omdat hij zich niet kan voorstellen dat dat niet zo zou zijn. Daarnaast bagatelliseert hij ten onrechte de betekenis van de excepties voor 'tijdelijke reproductie' als voor 'tekst- en datamining'? Dat is mijns inziens onzorgvuldig en gevaarlijk. De belangrijke vraag die qua auteursrecht voorligt is, als een inbreuk kan worden aangetoond in de output van GPT-4, wie dan verantwoordelijk voor die inbreuk is: de GPT-exploitant, de GPU-programmeur, degene die de trainingsdataset samenstelde, en zo ja welke samensteller en welke dataset, degene die de prompt formuleerde, of degene die de output vrijgaf? Die vraag is niet aan de orde gesteld en niet beantwoord en zal niet zonder dat rechter of wetgever verantwoordelijkheid nemen hem met gezag te beantwoorden, te beantwoorden zijn.

En dat is symptomatisch voor alle GPT-4 output die als inbreuk op het recht wordt ontmaskerd. Ook wanneer die voortvloeit uit vragen aan GovGPT-NL. En dat zouden we, als we er niet bijtijds iets tegen doen, inderdaad als een *profound change in the history of life in the Netherlands* kunnen gaan ervaren.

PS. Omdat de samenleving, ook de Nederlandse, spoedig zal worden overspoeld met *apps* die met behulp van GPT-4 juridische antwoorden op vragen voorbereiden is het aan te bevelen dat het zelf samenstellen en *fine-tunen* van onderscheidende GPT-4 dialecten met behulp van eigen trainingsmateriaal tot de vaardigheden van de praktijkjurist gaan behoren. Dat heeft hopelijk als neveneffect dat het juridische en het technologische debat beter op elkaar gaan aansluiten dan vooralsnog het geval lijkt.