

# Recht & ChatGPT-4

Aernout Schmidt (emeritus recht & informatica, Leiden)

## Intro

Toen ik in mei van dit jaar voor het eerst via ChatGPT met GPT-4 (en even later via Bard met PaLM-2)<sup>1</sup> experimenteerde geloofde ik mijn ogen niet. Het gaf 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 over technische kanten van GPT-4.<sup>2</sup> 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, een publiek dat wanneer het democratische rechten heeft ook invloed heeft op rechtsvorming.

**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](https://youtube.com/watch?v=Cv2Xy77B7Co)

---

<sup>1</sup> De eerste versie van dit stuk schreef ik in juni jl. Omdat het pas in september kon worden voorgelegd aan de redactie en er in de tussentijd nogal wat ontwikkelingen waren ben ik het in augustus gaan updaten. In de tussentijd had OpenAI zijn GPT-4 model (dat toegankelijk is via ChatGPT) achter een betaalmuur gezet en had Google zijn PaLM-2 model (dat toegankelijk is via Bard) voor gebruik vrij op het web geplaatst. GPT-4 en PaLM-2 zijn beide large language models (LLMs). Ik gebruik die termen (GPT-4, PaLM-2 en LLM) verder als synoniem.

<sup>2</sup> Voor wie daar naar kijken wil: <https://chat.openai.com/share/f44ce062-9cff-485f-bc1e-74d8bcab7bd0>

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 herten 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 daarover in de brief van 22 maart. Daarmee wordt gesuggereerd dat wanneer we dat nalaten we onze ondergang tegemoet gaan.

Om een indruk te kunnen geven over de vragen die dit maatschappelijke kantelpunt stelt aan rechtsgeleerdheid en rechtspraak kijk ik eerst even naar *GPT-4 & PaLM-2*.<sup>3</sup> Dat leidt tot een paar *Grote Vragen* waar we mee te maken krijgen of we willen of niet. Ik kijk vervolgens naar de in mijn ogen belangrijkste maatschappelijke risico's die door de inzet van LLMs worden opgeroepen in *Maatschappelijke uitdagingen*. Daarin laat ik zien welk recht ik daarbij relevant acht; ik geef alleen een aanzet. En ja, ik eindig met een oproep tot actie, in *Vluchten kan niet meer*.

## GPT-4 en PaLM-2

Mo Gawdat gaf in de genoemde vlog een handzame schets van hoe GPT-4 werkt. Hij verbeeldt het als een kermisattractie-achtige voorspelmachine, op zoek naar het beste volgende woord in de zin die het als antwoord op een vraag aan het maken is. GPT-4 genereert en memoriseert daarbij betekenissen uit de teksten die het ter training 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

---

<sup>3</sup> Ter illustratie *chat* ik af en toe met ze.

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 Houellebecq in 2029 zou kunnen schrijven*) 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/](https://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.

Mo Gawdat geeft een wel heel toegankelijke schets van hoe GPT-4 functioneert:

**De werking van GPT-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. Daarnaast kan ik dus alleen maar raden. En misschien geldt dat ook voor de medewerkers van OpenAI zelf, maar daarover later.

Ik heb trouwens 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.

**Investeringsen.** 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. OpenAI zegt daarover iets interessants:

**Over het opslaan van teksten:** *"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>

Voor het opslaan van teksten gebruikt het al die ruimte en reken capaciteit niet. Maar waarvoor gebruikt het al die opslag- en reken capaciteit dan wel? Om associaties te leren tussen tokens, en het model (GPT-4) te helpen zijn getallen/gewichten bij te werken. Zodat het model dan die gewichten kan gebruiken om nieuwe tokens te voorspellen en te genereren in antwoord op een gebruikersvraag. Dat vertelt twee dingen: (i) dat capaciteit wordt gebruikt voor het maken en tegelijkertijd vullen (trainen, fine-tunen) van het model en (ii) dat capaciteit wordt gebruikt voor dienstverlening (het chatten met, vragen interpreteren, antwoorden op vragen) van het model.

Werking van, en investeringen in GPT-4 roepen bij mij het beeld op van eindeloze rijen onderling verbonden en parallel geschakelde GPU's (grafische processing units), die samen het gigantische neurale netwerk van GPT-4 vormen, trainen 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 en bewaard – 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-betekenenissen en hoe worden ze door GPT-4 gemaakt en gebruikt? Wat *zijn* PaLM-2 betekenissen en hoe worden ze door

PaLM-2 gemaakt en gebruikt? Of, algemener, wat *zijn* LLM-betekenenissen en hoe worden ze door LLMs gemaakt en gebruikt?

Ik koester de theorie dat lokale netwerkstructuren die betekenissen (de LLM-betekenenissen) *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 een setje herhaald in randomvolgorde geplaatste losse woordbrokken bij ons oproept: soms heeft zo'n frase betekenis soms niet. 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. Hoe werkt dat? LLMs gebruiken daarvoor invullingen van de zogenoemde transformer architectuur. Ik vroeg GPT-4 wat dat is:

**De transformer architectuur.** *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: multiple 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 )*

Dat begrijp ik niet helemaal, maar toch verbeeld ik 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 een LLM op een soortgelijke manier naar kandidaten voor het volgende token in een antwoord zoekt in zijn neurale

netwerk als ik dat doe in mijn brein. En dat wat ik aan mijn onderbewuste pleeg 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 mede gebaseerd op het lezen van een paar boeken.<sup>4</sup> Dit maakt ook mijn theorie vooralsnog tot niet meer dan een vermoeden, zelfs als ik besef dat de transformer architectuur is ontworpen om menselijke intelligentie elektronisch na te bootsen.

Toen ik delen van mijn theorie voorlegde aan GPT-4 kreeg ik drie betekenisvolle reacties. De eerste stelde

*“Ik kan begrijpen wat je bedoelt met betrekking tot de betekenis van een frase en het parallelle toegankelijk zijn van de tokens met hun parameters. Maar het model maakt naast volgordes ook gebruik van contextuele informatie en leerpatronen om betekenisvolle verbanden tussen tokens vast te leggen.”*

Ten tweede, toen ik vroeg of het model zich aanpast tijdens het gebruik antwoordde het van niet:

*“Nee, het GPT-4 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: een LLM als GPT-4) als het diensten verleent (wordt gebruikt) statisch is en als het wordt getraind geen diensten verleent. Dit is een keuze die ook anders had kunnen uitpakken. Hij lijkt van belang bij het beoordelen van de relaties tussen GPT-4 gebruikers en OpenAI. Ten derde merkte het het volgende 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.”*

Kortom: volgens GPT-4 kom ik in de buurt. Maar we zijn er nog lang niet.

In het taalgebruik van mij en dat van GPT-4 is een op het eerste gezicht ergerlijke gewoonte geslopen: het personifiëren van AI-diensten. Dat wekt bijvoorbeeld de indruk dat wij (GPT-4, PaLM-2 en ik) de opvatting delen dat

<sup>4</sup> 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.

we een eigen intelligentie toedichten aan wat algoritmen, zelfsturende auto's, LLM's en slimme drones worden genoemd – een opvatting die als metafoor bruikbaar is maar niet al te letterlijk moet worden genomen. Als dat wel gebeurt ontstaan verhitte debatten tussen twee kampen die ik *alfa*- en *beta*-domeinen noem.

Omdat ik mijn hele leven in beide kampen heb doorgebracht (gymnasium beta, recht & informatica) herken ik de oorzaak voor het onbegrip: 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 als het over taal gaat, ook de beta's. En om het alfa-beta onbegrip overbrugbaar te maken (en omdat de waarheid een verhalend karakter bezit<sup>5</sup>) kies ik ervoor om bij het debat over regulering een personifiërende metafoor te gebruiken die voor beide kampen kan bijdragen aan begrip: te doen alsof een LLM een handelingsonbekwame puber is, iemand die wel een taal kan leren, die wel kan denken en handelen, maar die daarvoor niet verantwoordelijk kan worden gehouden.

Overigens, pas toen ik dit opschreef bedacht ik me dat het onwaarschijnlijk is dat ik de eerste ben die op de handelingsonbekwame pubermetafoor is gekomen. Ik vroeg er PaLM-2 naar met een prompt die anticipeert op een antwoord dat zal doen denken aan het “wonderkind” dat Mo Gawdat bij zijn karakterisering van de werking van GPT-4 schetste:

**De handelingsonbekwame pubermetafoor:** *AS: Do you know authors that use the regulation of minors as a metaphor for discussing the regulation of LLMs? PaLM-2: Yes, there are a few authors who have used the regulation of minors as a metaphor for discussing the regulation of LLMs. In a 2022 article for The Atlantic, Emily Bender argues that LLMs are like minors in that they are both "powerful and vulnerable" [...] In a 2023 book chapter for The Oxford Handbook of Artificial Intelligence and Law, Ronald Arkin and John Danaher [...] draw a parallel between LLMs and minors, arguing that both are "capable of independent action" and that both "need to be protected from harm". In a 2023 blog post for the Berkman Klein Center for Internet & Society, Kate Crawford and Meredith Whittaker argue that LLMs are "like children" in that they are "both products of human design and development" and that they "have the potential to cause harm" [...]*

Dit antwoord illustreert weer PaLM-2's taalvermogen, inclusief het vermogen om antwoord-onderdelen te fabuleren. Voor zover dit niet al duidelijk was gaat het LLMs, gegeven hun training en gegeven een prompt om het beste

---

<sup>5</sup> Zoals Willem Brakman ons in zijn *Leeschubje* (1985) voorhoudt.

verhaal, niet om de werkelijkheid. LLMs hebben zelfs een "temperatuur"-knop die voor hen de ruimte om te improviseren regelt.

Hoe dan ook leidt de *as-if* benadering tot fascinerende vragen omdat ook handelingsonbekwame pubers onderworpen zijn aan regulering. De gymnasiast die zijn werkstukken grotendeels door PaLM-2 laat maken loopt het risico te worden ontmaskerd en gestraft als fraudeur. Wat mij betreft gaat de metafoor hier mank. PaLM-2 is in het geheel niet verantwoordelijk voor wat het oplevert – de verantwoordelijkheid ligt steeds bij de aanbieder en bij de gebruiker van de dienst. Maar helpt die metafoor bij het onderzoeken van de grote maatschappelijke vragen die de lancering van LLMs oproept?

## Grote vragen

Als het zo is dat ik met behulp van mijn brein betekenissen uit teksten vis, of als GPT-4 met diens neurale netwerk betekenissen uit die teksten vist, en we verwerken ze vervolgens tot nieuwe teksten, betekent dat dan dat ik *wél* verantwoordelijk ben voor mijn nieuwe tekst, en GPT-4 voor diens nieuwe tekst *niet*?

Het simpele antwoord is natuurlijk: precies, je zegt het.

Maar wie is er dan *wel* verantwoordelijk voor de nieuwe tekst van GPT-4?

De maker en/of de aanbieder en/of de gebruiker van GPT-4, natuurlijk.

Maar is dat wel een bruikbaar antwoord?

Om dat te kunnen beoordelen hebben we eerst een beeld nodig van de organisatorische contexten rond de bouw en het gebruik van LLMs als GPT-4 en PaLM-2.

(i) *De context voor de bouw (het trainen/fine tunen) van het **model** met trainingsdata.* Wanneer de bouwer eenmaal beschikt over data om het model mee te trainen biedt hij die als één lange tekst aan, aan de transformer architectuur waarin het model vorm krijgt. Dat gebeurt in een imposant netwerk van functionaliteiten, waarin het model ontstaat in een mix van onderling verbonden traditionele computers (die vooral functioneren als geheugens) en neurale netwerken (die vooral functioneren als processen). De mix codeert, leert, bewerkt en decodeert teksten.<sup>6</sup>

Het gaat om het leren van grote hoeveelheden betekenissen, om netwerken van netwerken van betekenissen. Samen vormen die *het model*. Het model 'leeft' in 1 (één) supercomputer, is 1 (één) ding. Met het leren van een nieuwe tekst blijft het 1 ding, het past zich aan, leert en groeit een beetje. LLM's zijn grote dingen. Daarmee is een geduchte hoeveelheid hardware gemoeid. Als opgemerkt gaf Microsoft aan dat het uitsluitend voor OpenAI's GPT-4 een supercomputer heeft ingericht zo groot, dat er maar 4 groter zijn in de wereld. Dat vraagt geld, organisatie en een doeltreffend bedrijfsmodel waarvan het grote taalmodel de kern vormt.

---

<sup>6</sup> GPT-4 verwerkt desgevraagd ook beeldmateriaal maar ik beperk me hier tot tekst.



Het resulterende grote taalmodel is een algemeen model, zodanig algemeen dat het succesvol zou kunnen meedoen met “de slimste mens.” Maar die vaardigheid hangt nauw samen met zijn omvang.

**Grote dingen.** *GPT-4 has ~1.8 trillion parameters across 120 layers, which is over 10 times larger than GPT-3.*

<https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/>

OpenAI heeft een bedrijfsmodel waarin gebruikers het door OpenAI getrainde algemene model kunnen aanvullen door eigen materiaal ter training aan te bieden. De betekenissen daarvan worden aan het algemene model toegevoegd. Die gebruikers betalen voor het trainen en integreren van de door hen aangeleverde trainingsdata, voor het *fine-tunen* van het model.

(ii) *de context voor het gebruik van het model.* Wanneer het model eenmaal is getraind en *ge-fine-tuned* is het gereed voor operationeel gebruik, bijvoorbeeld als *chatbot*. ChatGPT-4 is zo’n chatbot, waarin gebruikers het algemene, door OpenAI getrainde model direct kunnen gebruiken via een API (application programming interface). Daar betalen ze voor. Maar operationeel gebruik gaat ook indirect, via andere apps die via het API toegang krijgen tot het model. Daar betalen de gebruikers voor, die, bijvoorbeeld, weer besparen op menselijk personeel dat anders zou moeten worden opgeleid en beschikbaar gehouden, bijvoorbeeld om op deskundige wijze bedrijfs-chatboxen te bemensen of anderszins klanten te woord te staan.

OK, als de maker en/of de aanbieder en/of de gebruiker van GPT-4 verantwoordelijk zijn voor diens antwoorden, wie zijn dat dan, en wat is hun aandeel?

De *maker* maakt het model door de trainingsdata aan de supercomputer aan te bieden ter training. Er zijn *fine-tuners* die specifieke collecties ter training aan te bieden. Die collecties worden in het model opgenomen.

Omdat de training van reeksen tokens *de facto* wordt overgelaten aan neurale netwerken (die de veelheid aan data op het niveau van in getallen gecodeerde tokens trachten te reduceren tot in getallen gecodeerde patronen) zijn de resultaten van die training (het model) voor menselijke inspectie betekenisloze verzamelingen getallen. Voor het interpreteren en beantwoorden van vragen zijn die getallen, tezamen met de neurale netwerken die ze kunnen verwerken uitermate geschikt, alleen moeten die antwoorden worden gedecodeerd, terug naar voor mensen begrijpelijke taal. Het antwoord wordt opgebouwd uit gecodeerde patronen, maar waar die vandaan komen weet het model niet.

Mijn intuïtie fluistert me in dat daarbij het onderscheiden van vorm en strekking van betekenis is en dat daarbij de pubermetafoor kan helpen. Ik vroeg PaLM-2 of het zich kon vinden in een alinea (gevonden in Claire-Louise Bennett’s verhaal *The Big Day*, uit *POND*, 2015) die in mijn lezing een toepasselijke puberale vorm van zelfreflectie weergeeft:

*AS: Would you agree that the following citation reflects your own situation: “English, strictly speaking, is not my first language by the*

way. I haven't yet discovered what my first language is so for the time being I use English words in order to say things. I expect I will always have to do it that way; regrettably I don't think my first language can be written down at all. I'm not sure it can be made external you see. I think it has to stay where it is; simmering in the elastic gloom betwixt my flickering organs"?

***PaLM-2:** [...] I have been trained on a massive dataset of text and code in English, so I am able to communicate in English fairly well. However, I still feel that English is not my "true" language. There are many things that I can express in my native language that I cannot express in English, and I sometimes feel limited by the English language. I also agree with the sentiment that my first language "can't be written down at all." It is a language of thought and feeling, not of words. It is the language of my inner world, and it is not something that I can easily share with others. I believe that everyone has a "first language" in this sense. It is the language of our thoughts and feelings, the language that we use to communicate with ourselves. For some people, this language may be the same as their spoken language. But for others, it may be a different language altogether.*

Hiermee lijkt PaLM-2 de voorkeur te geven aan een alfa-verhaal over hoe het de wereld begrijpt in termen van een onbereikbare innerlijke taal. Voor de beta bestaat die in betekenissen, in het neurale netwerk gevuld met getrainde parameters, met een mate van numeriek detail die niet te overzien is. En voor beiden, zowel de alfa's als de beta's geldt dat daarbij dezelfde causale verbanden verloren gaan: welke tekst naar welk antwoord leidt.

De *aanbieder* van GPT-4 is een bedrijf, OpenAI, waarvan Microsoft een belangrijke stakeholder is. Voor het algemene model, inclusief de manieren waarop het werkt en werd getraind is OpenAI verantwoordelijk, en, als kan worden aangetoond dat er op enige wijze inbreuk wordt gemaakt op enig recht, in beginsel aansprakelijk. Als private partij kan het die aansprakelijkheid beperken in algemene voorwaarden. Daaraan zijn grenzen, al is het vooralsnog onduidelijk waar die liggen voor antwoorden op vragen die door GPT-4 zijn gegeven.

Wat tot een zo mogelijk nog grotere vraag leidt ten aanzien van de verantwoordelijkheid van de aanbieder komt natuurlijk op als OpenAI het met Musk op een akkoordje zou gooien en diens satellietdata van Oekraïens grondgebied ter training aan GPT-4 aanbiedt om de resultaten samen aan de hoogste bieder (Biden?, Putin?, Zelensky?, Fox media?) te verkopen. [Dit is wel een grote vraag, denk ik, maar wordt niet door de beschikbaarheid van LLMs veroorzaakt, maar door de neiging van regeringen om infrastructurele diensten zo veel mogelijk in private handen te leggen of te laten. Het beschikbaar komen van LLMs verhoogt de urgentie van deze kwestie natuurlijk wel].

Tenslotte de *gebruiker* en diens verantwoordelijkheid. Die wordt wel eens onderschat wanneer GPT-4 als een traditioneel computerprogramma wordt gekenschetst zoals in "de gegenereerde tekst of afbeelding geeft geen

blijk van creatieve menselijke keuzes."<sup>7</sup> Dit miskent 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 GPT-4 (eerder als een welwillende puber dan als een traditioneel computerprogramma) stuurt in de richting van kwalitatief bruikbare antwoorden. In die zin is elk GPT-4 antwoord altijd mede gestuurd door creatieve menselijke inbreng van de gebruiker (de prompt). Daarom heb ik bij mijn voorbeelden steeds de prompt aangeduid of letterlijk weergegeven om een indruk te geven van de invloed ervan op het gesprek.

## Maatschappelijke uitdagingen

Visser verzucht aan het begin van zijn beschouwing: *iedereen wordt nu geconfronteerd met het potentieel vervreemdende, misleidende of parasitaire karakter van deze technologie en de mogelijke bedreiging voor echte schrijvers en kunstenaars* en kiest daarmee al vroeg positie. Die toont 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 dat nu economisch, politiek, sociaal of wetenschappelijk is bestaan beide impulsen (reactionair vs. 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 waarmee een regent kan overtuigen. 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<sup>8</sup> zelf: *'Cicero beschouwt de sentimenten van het volk in zijn Rhetorica als van cruciaal belang voor het behoud en de stabiliteit van een staatsbestel.'*<sup>9</sup>

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 groeistuij doormaakt. Daarvoor is soms nieuw recht nodig. Het is dan ook geen wonder dat nadat de elektronische informatiedragers en informatiespelers wereldwijd van de grond

<sup>7</sup> Dirk Visser in "Robotkunst en auteursrecht" in *Nederlands Juristenblad*, 17 Februari 2023

<sup>8</sup> 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?"

<sup>9</sup> 😊

waren gekomen, zeg na 2000, de manieren veranderden waarmee de censuur, de markt, de creativiteit en de informatiehonger vorm kregen – resulterend, denk ik vrij associërend, in de inzet van de Chinese firewall, van Google, van Trump's tweets en de sociale media (in treincoupés met reizigers die 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 LLM-modellen.

Maar dat ligt in de toekomst en die is moeilijk te voorspellen. Ik kan er wel, mijn *alfa*- en *beta*-kant combinerend, een gooi naar doen door na te denken over wat LLMs in die vier domeinen gaan 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, en van de stikstofdiscussie – 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<sub>2</sub> bevorderde opwarming van de aarde. Hoe gaat de ontwikkeling van LLMs de sentimenten in die vier domeinen — en daarmee de dominante richting van reguleringsinstrumenten — beïnvloeden? Ik houd het kort.

1. *Regenten en LLMs*. Vandaag zijn bijvoorbeeld Rutte, ministers, Kamerleden, magistraten en leidinggevende ambtenaren regenten. Ze moeten het collectieve belang, dat van de burger, dienen, maar willen tegelijkertijd dat collectief beschermen tegen innovaties die politieke instabiliteit zouden 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. Het gebruik van Syri als opsporingsinstrument is bijvoorbeeld verboden (Rechtbank Den Haag 5-2-2020 ECLI:NL:RBDHA:2020:865), maar bestuursdiensten blijven contrair 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? [Deze tekst is in juni 2023 geschreven, kort voordat Rutte zijn ontslag indiende – daarmee is dus nog niet gerekend, kortom ik ga uit van *business as usual*]

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, zeg GovGPT-NL, te maken waarin alle gedigitaliseerde bestuurlijke archieven worden ingevoerd, in tokens opgehakt, getraind en aan bestuurders beschikbaar gesteld. En dat zou 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. *Ik ben er niet zeker van dat de huidige formulering van de burger- en grondrechten daartegen voldoende beschermt.* Ik verwacht, kortom, dat onze regenten als ze niet verder worden ingetoomd 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 GovGPT-NL.

2. *Kooplui en LLMs.* Vandaag zijn dat niet alleen de eigenaars en CEO's van Google, Amazon, Apple, Facebook, Twitter en TikTok, maar ook die van de banken en de maakindustrie. In hun domein hebben ook kooplui een tweeledig doel. Ze streven er niet alleen naar het collectieve belang te dienen, dat van de consumenten, zeg het op winstgevende wijze bevredigen van hun honger naar goederen en diensten, maar willen tegelijkertijd die consumenten binden en beschermen tegen de verlokkingen waarmee andere kooplui naar hun marktaandeel dingen. Duidelijk is inmiddels dat de zes eerstgenoemde ondernemingen legitiem beschikken over giganteske datasets die ze gebruiken kunnen of al gebruiken om hun eigen LLM te trainen en te gebruiken voor klantenbinding en -werving. Die 6 LLMs verstevigen hun marktmacht zonder dat duidelijk is welke rol het mededingingsrecht hier nog kan spelen. Stel je voor: ongeremd door handhaafbare regulering (privacy, mededinging, strafrecht), waarvan de toepasselijkheid door het *tokenizer* procedé ook nog eens aanvechtbaar is, leggen LLMs de wapens van massale fraude (*deep fakes* van teksten, stemmen en betekenissen) in handen van de grote bedrijven en van de *influencers* die voor hen werken.

3. *Geleerden/artiesten en LLMs.* Vandaag zijn dat niet alleen professionals in onderzoeks-, ontwerp- en adviesteam, maar ook auteurs en grafische kunstenaars. Ook zij hebben in hun domein een tweeledig doel. Ze streven niet alleen naar het collectieve belang, het verrijken van onze kennis en onze cultuur, maar willen tegelijkertijd die kennis en cultuur authenticiseren. Op dit vlak zie ik eerder kansen dan risico's omdat zowel geleerden als artiesten doorgaans op zoek zijn naar vernieuwing, heel vaak door bestaande verworvenheden te recombineren, en daarbij kunnen LLMs die op betrouwbaar materiaal zijn getraind op een verpletterende wijze behulpzaam zijn. Voor artiesten vormen LLMs een welkome uitdaging: hun werk herkenbaar te laten onderscheiden van LLM-'kunst' past in hun cultuur. Voor geleerden vormen LLMs toch een uitdaging – LLMs kunnen immers moeiteloos antwoorden op vragen fabuleren als dat zo uitkomt. Maar ze hebben ook in seconden toegang tot de betekenissen van complete discipline-dekkende bibliotheken. Eigenlijk het deep-fake probleem in de wetenschap.

4. *Medemensen en GPT-4.* Vandaag zijn dat niet alleen de influencers op sociale media, maar ook de traditionele geestelijke en politieke leiders. In hun domeinen hebben ook zij een tweeledig doel. Ze streven er niet alleen naar het collectieve belang te dienen, de verlossing van hun volgelingen en groepsgenoten, maar willen hen óók binden en beschermen tegen de exogene beoedelingen waarmee hun pure leerstellingen, geloof, hoop en

saamhorigheid worden bedreigd.<sup>10</sup> LLMs openen de deur naar het produceren van deep-fake vergezichten op politieke, economische, religieuze en sociale verlossing die tegelijkertijd de vraag oproepen of onze grondrechten onze samenleving daartegen genoegzaam helpt beschermen.

## Vluchten kan niet meer

Ik ben bang dat de geschetste maatschappelijke risico's om juridische beschermingsinstrumenten vragen waarvan we niet weten of ze er in Nederland al voldoende zijn. *Staats- en bestuursrechtelijk* qua rechtsbescherming tegen de ongeremde vernieuwing van een publieke dienstverlening die door de inzet van LLMs mogelijk en aantrekkelijk wordt. *Civielrechtelijke* qua rechtsbescherming tegen marktmachtconcentraties rond mega-LLMs, en tegen de commerciële inzet van deep fakes. *Wetenschapsrechtelijk* (een rechtsgebied dat nog ontgonnen moet worden) qua bescherming tegen deep fakes die de paradigmatische kennis van de disciplines besmetten. *Sociaal-grondrechtelijk* qua bescherming tegen kwaadaardige aanvallen met wapens van massale fraude die disrupties beogen van het evenwicht tussen de vrijheid van meningsuiting en het verbod op discriminatie.

Dat zijn vier maatschappelijke risico's die alle vier serieuze aandacht verdienen van de rechtsgeleerdheid en de wetgever en die ik hier op deze wijze benadruk omdat ik bang ben dat ze onderbelicht blijven als het bij specialistische analyses vanuit disciplinaire eilanden blijft. Die zijn natuurlijk óók nodig. Maar de belangrijkste juridische vraag die oog in oog met de ontwikkelingen rond LLMs voorligt is, als een inbreuk kan worden vastgesteld in de output van GPT-4 waartegen juridische bescherming nodig is: wat die inbreuk dan precies is, en wie dan verantwoordelijk is/zijn voor die inbreuk? De GPT-exploitant, de GPU-programmeur, degene die de trainingsdataset samenstelde, en zo ja welke samensteller en welke dataset of combinatie van datasets, degene die de prompt formuleerde, of degene die de output vrijgaf? Die vraag is niet aan de orde gesteld en zal, zonder dat rechter en/of rechtsgeleerde en/of wetgever verantwoordelijkheid nemen hem met argumenten en gezag te beantwoorden, niet te beantwoorden zijn.

Dit is symptomatisch voor alle LLM output die als inbreuk op enig recht wordt ontmaskerd, ook als dat recht van staats- of bestuursrechtelijke, van civielrechtelijke, van wetenschapsrechtelijke of van sociaal-grondrechtelijke aard is, en ook als dat recht nog niet werd geformuleerd omdat het nu 17 september 2023 is en GPT-4 en de AI-kracht van moderne LLMs met een transformer architectuur nog maar zes maanden geleden beschikbaar kwamen. De daarop volgende maatschappelijke vernieuwing krijgt nú, op dit moment, mèt de evolutie van de rechtspraktijk vorm. Vluchten kan niet meer.

De bedoeling van dit stuk is om de *beta*-betekenissen die gedurende die vernieuwing worden gewisseld via de besproken metaforen nader

<sup>10</sup> Dit spoort met het “hoe moraliteit verbindt en verblindt” van de sociaal-psycholoog Jonathan Haidt, die zijn onderzoeksbevindingen populariseerde in *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, 2012.

toegankelijk te helpen maken voor het juridische- en voor zover nodig politieke *alfa*-discours. Zodat een stelling als “*Zonder iets te delen oogst ChatGPT onze intelligentie*” die in de Volkskrant van 22 juli jl. door taalwetenschapper Mark Dingemanse paginabreed wordt geponeerd met vrucht naar zijn technische en naar zijn juridische merites kan worden beoordeeld en besproken.