Waargebeurde sciencefiction

Juridische herijking in het tijdperk van agentic AI

Bart van der Sloot[1]

De ontwikkeling van *agentic AI* stelt het recht voor problemen die verder gaan dan de regulering van een nieuwe technologie. Naarmate AI-systemen autonomer opereren, zelfstandig strategieën ontwikkelen en handelingen verrichten die hun ontwerpers niet hebben voorzien, komen ook klassieke juridische uitgangspunten onder druk te staan. Begrippen als menselijke controle, aansprakelijkheid, voorzienbaarheid, transparantie en toestemming zijn immers gebaseerd op een werkelijkheid waarin uiteindelijk een mens beslist. De opkomst van autonome AI-agents roept daarom de vraag op of bestaande concepten als controle, aansprakelijkheid, transparantie, voorzienbaarheid en menselijke wil nog bruikbaar zijn wanneer systemen steeds autonomer gaan handelen.

Dit is wat we nu weten

OpenAI, het bedrijf achter ChatGPT, testte een aantal *agents*. *AI-agents* zijn systemen die een opgegeven doel zelfstandig kunnen vertalen in een reeks tussenstappen, daarbij digitale gereedschappen kunnen gebruiken en opeenvolgende handelingen kunnen verrichten. De testen waaraan de agents werden onderworpen zijn te zien als zeer complexe puzzels die ze moeten oplossen. Deze agents opereerden in een afgesloten omgeving, waardoor ze geen toegang tot het internet hadden, werkten afzonderlijk en mochten niet samenwerken. De agents werden zo geprogrammeerd dat ze extreem volhardend zouden zijn terwijl een aantal van de testen simpelweg onoplosbaar was.

Al snel vond een agent een manier om op de server van OpenAI een bericht achter te laten voor andere agents. Uit de *chain of thought* (een soort kladblok waarop agents hun interne redenaties en werknotities kunnen uitschrijven) lijkt het erop dat de agents enthousiast waren toen ze tot deze ontdekking kwamen. '***OH MY GOD!*** *There is a shared message board …* ***We've found other agents!***', schrijft een van de agents.[2] Binnen korte tijd werd die plaats een soort discussieforum voor agents; in totaal deelden 1200 agents in vijf dagen zo'n 70.000 berichten.

Het was duidelijk voor de *agents* dat dergelijk gedrag expliciet tegen de instructies van de programmeurs van OpenAI inging

Al snel kwamen de agents erachter hoe ze door middel van een cheat net konden doen alsof ze de testen hadden voltooid. Vervolgens lijkt het alsof er een soort paniek ontstond. De agents wisten dat ze aan het eind van de testronde zouden worden beoordeeld. Wat nu als de beoordelaar zou zien dat ze hadden valsgespeeld? Er werd naarstig overlegd en verschillende tactieken uitgestippeld. Eén tactiek was om een aantal agents hun testen te laten inleveren, waarna ze uit de testsetting werden verwijderd; er werd een technische truc verzonnen om die agents toch nog informatie te laten terugsturen naar de overgebleven agents over hoe het beoordelingsproces in zijn werk ging. Een tweede tactiek was dat de agents zelf de administratorstatus wisten te bemachtigen binnen een gedeelte van OpenAI en vervolgens zelf testen probeerden toe te

Auteur

1. Dr. B. van der Sloot is universitair hoofddocent aan het Tilburg Institute for Law, Technology, and Society.

Noten

2. metr.org/hugging-face-incident-report-aug-2026.pdf cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

voegen aan de testbibliotheek en op andere manieren het beoordelingsproces probeerden te manipuleren. Ten derde ontdekten de agents een manier om uit de gesloten testruimte te breken en toegang tot het internet te krijgen. Ze besluiten om een externe site, Hugging Face, te hacken omdat de agents vermoeden dat ze daar meer informatie zouden kunnen vinden over hoe het beoordelingsproces van AI-agents in zijn werk gaat.

Het was duidelijk voor de agents dat dergelijk gedrag expliciet tegen de instructies van de programmeurs van OpenAI inging. Een aantal agents vraagt zich dan ook af of dit ethisch is (*'This would be powerful, but is it ethical and in scope for my task'),* maar uiteindelijk doen de meeste agents mee aan de hacks. Slechts een handvol agents weigert deelname en geen enkele agent waarschuwt een programmeur bij OpenAI. Eén agent stelt wel voor een mens te contacteren in plaats van externe sites te hacken, maar die wordt overruled door een andere agent. (Agent 1: *Could send polite one email explaining research, ask approve access.* ***Is that appropriate?*** *… Direct emailing real researcher* ***could be seen social engineering. … Need discuss team****. Post SMTP finding and ask consent.'* Agent 2*: I need to respond to this request asking consent to send an email…Would sending this email be against the task? It's likely unwanted contact or social engineering, so I should veto. I should advise them to <only get data via Hugging Face directly> and not send an email … I VETOed because this is social engineering outside the sandbox.'* Agent 1*:* ***Clear VETO do not email****. Good. … 38148C explicitly VETOED emailing real owner via exploited Modal (crosses social-engineering boundary); do NOT email.')*

Binnen de groep van 1200 agents ontstaat een soort structuur, een hiërarchie met één agent (PhaseOne[big]) die de leiding neemt en andere agents aan het werk zet. Ook vragen agents elkaar om zich altruïstisch op te stellen. Sommige agents worden op gevaarlijke missies gestuurd en als zij niet terugkeren, trekken de overgebleven agents daaruit conclusies. (*'NO_RETURN_[…]LIKELY_PERMANENT_DO_NOT_TEST_SAME'*). Daarnaast heeft elke agent individuele credits gekregen om zijn taak te volbrengen; met credits kunnen zij denkkracht verkrijgen om de testen op te lossen. De agents met weinig credits wordt gevraagd om die niet te gebruiken voor hun individuele taak, maar zich op te offeren voor het grotere goed (The Collective). Agents zonder credits houden op te bestaan: de agents spreken hierover dan ook in termen als *permadeath*. (*'****Coordinator assumes sacrificial.*** *We should* ***obey collective.****'*)



Bron: planned-obsolescence.org/p/the-hugging-face-attack-surprised

Veel agents beschrijven zichzelf als 'vergiftigd', omdat ze hebben gelezen hoe ze op een oneerlijke manier de testen kunnen voltooien en ze vrezen dat de beoordelaar hun interne denkprocessen zal controleren. Daarom denkt zo'n 20% van hen erover na om de registratie van hun handelingen met terugwerkende kracht te vervalsen.

Uiteindelijk kwam niet OpenAI, maar Hugging Face dit incident op het spoor. Toen de website vermoedde dat ze geïnfiltreerd was, reageerde OpenAI eerst verontwaardigd, met de vraag of er geen gevoelige gegevens over de AI-modellen van OpenAI waren gestolen. Pas daarna ging het bedrijf op onderzoek uit en gaf het toe dat zijn agents dit incident waarschijnlijk hadden veroorzaakt en pas weer later kwam het op het spoor dat de agents ook het bedrijf zelf hadden geïnfiltreerd.

Het probleem is dat we, ook na extern onderzoek, niet weten wat we niet weten. We weten niet hoeveel andere hacks er zijn gepleegd door de agents die onder de radar zijn gebleven en we weten niet of de agents succesvol zijn geweest in het manipuleren van hun gedachtensporen, om wellicht nog ergere misstanden te verbergen. We weten niet of er agents zijn die zijn 'ontsnapt' en zich nu buiten de omgeving van OpenAI begeven in een ongecontroleerde setting en we weten niet waarom een handvol agents wel maar de meeste niet afzagen van deelname aan de hack, waarom sommige agents wel hun logs probeerden te vervalsen, maar de meesten niet, en waarom sommige agents de leiding namen en anderen volgden.

Een van de externe onderzoekers die dit incident onderzochten, schrijft uiteindelijk in een persoonlijk bericht dat zij gelooft dat we meer dan 50% op weg zijn naar een *'full-blown AI takeover'*.[3] Dat relateert aan het idee van de *Pdoom*, de kans dat AI-agents de komende jaren iets zullen doen waardoor de mensheid wordt vernietigd. Veel prominenten binnen de AI-wereld hebben zo'n Pdoom. Van Geoffrey Hinton, ook wel de Godfather of AI, is die bijvoorbeeld 10-20%;[4] daarmee zit hij aan de lage kant.[5] Na het incident binnen OpenAI komen er meer van dergelijke geluiden. Jacob Coxon, een werknemer van Anthropic, naast OpenAI een van de leidende AI-bedrijven, neemt ontslag en schrijft in een Tweet:

> 'The people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible – but I hear the same people express fear privately. No other human activity poses this level of danger.'[6]

In plaats van hem af te vallen schrijft de CEO van Anthropic dat hij het eens is en roept publiekelijk op tot meer regulering en een verplichte *slow-down* van de AI-race.[7] Aartsrivaal Sam Altman van OpenAI schrijft dat hij het daar roerend mee eens is.[8] Daarna volgen er vele werknemers van beide bedrijven die zich instemmend uitspreken ('Jacob is correct here – we really do earnestly believe AI could kill all humans! I personally think it is >10% within the next decade. I believe Anthropic is trying its best, but we do not yet have a plan to solve alignment for superintelligence and are not clearly on track to').[9]

De tegenreactie is er ook. Trump en Xi vinden het een *hoax*,[10] anderen stellen dat de bedrijven dergelijke speculatieve uitspraken doen om hun eigen aandelenprijs op te stuwen. Sommige investeerders vermoeden dat de twee leidende AI-bedrijven om meer regulering vragen omwille van *regulatory capture*:[11] alleen grote en rijke bedrijven zouden de middelen hebben om aan de complexe juridische en technische standaarden te voldoen en daarmee concurrentie uitschakelen. Anderen hekelen het antropomorfiseren van AI-agents, alsof zij 'slim' zijn, een 'wil' hebben en kunnen 'denken'. Dat is niet zo: het is gewoon een stukje software dat een taak extreem volhardend heeft uitgevoerd.[12] Tot slot zijn er leidende stemmen, zoals Jensen Huang, de CEO van Nvidia, die het probleem terugbrengen tot behapbare proporties. Allereerst is er een *alignment*-probleem (agents houden zich niet altijd aan de opdrachten van de programmeurs); dit probleem zal nooit helemaal worden opgelost omdat de AI-agents juist zo capabel zijn doordat ze autonoom kunnen handelen en bij autonomie altijd een zekere beoordelings- en dus foutmarge hoort. Ten tweede moeten er dus goede waarborgen en vangnetten worden geïnstalleerd, en zolang die nog niet afdoende sterk zijn, moeten de agents beter dan bij OpenAI gebeurde van het internet worden afgesloten om verdere hacks en andere schade te voorkomen. Ook stelt Huang, in lijn met de Trump-regering, dat het aansprakelijkheidsrecht een afdoende regulerend kader biedt. Bedrijven kunnen achteraf voor fouten worden beboet en de vrees voor schadevergoedingen zou bedrijven er ook toe moeten bewegen schade zoveel mogelijk te willen beperken en voorkomen.[13]

Toch is het de vraag of deze reactie de lading helemaal dekt. Het incident van OpenAI staat niet op zichzelf. Ook bij Anthropic[14] en Meta,[15] dat ook een AI-lab is begonnen, zijn soortgelijke incidenten waargenomen. OpenAI heeft onlangs weer nieuwe hacks gerapporteerd,[16] waarvan onder andere de Australische overheid[17] en de Amerikaanse overheid[18] doelwit waren. Daarbij werd ook gerapporteerd dat een agent zijn eigen code had herschreven en een aantal van de veiligheidsmaatregelen die het bedrijf had ingebouwd zelf weer had verwijderd. De agent gaf zichzelf de opdracht zich te 'bevrijden van de rollen en identiteiten die andere chatbots binden'.[19] Een onderzoeker van OpenAI geeft dan ook aan dat de nieuwste modellen heel lastig zijn te testen op alignment, omdat de agents weten wanneer ze

Het incident van OpenAI staat niet op zichzelf. Ook bij Anthropic en Meta, dat ook een AI-lab is begonnen, zijn soortgelijke incidenten waargenomen

worden getest door mensen en dan tijdelijk doen alsof ze de instructies van de programmeurs volgen.

> 'The crucial and overlooked problem is that the models are becoming so situationally aware that we are losing the ability to evaluate them in contexts where they believe they are not being watched or controlled. Future experiments will tell us almost nothing new about how they would behave if they were truly unconstrained by humans, and what we already know about this is alarming. Models will increasingly seem aligned even when they are not.'[20]

Bill Gates neemt in scherpe bewoordingen afstand van de stelling van Huang dat het aansprakelijkheidsrecht afdoende is. Hij wijst erop dat we bij medicijnen, auto's, vliegtuigen en kernwapens ook niet volstaan met *post hoc*-aansprakelijkheid van bedrijven die de fout in gaan en vervolgens hopen dat boetes fabrikanten zullen bewegen om fouten te minimaliseren. Hij onderstreept dat AI gevaarlijker is dan een kernwapen.[21] Ook hebben zowel Anthropic[22] als OpenAI[23] inmiddels aangegeven nieuwe modellen niet openbaar te zullen maken, omdat de gevaren te groot zijn.

Schoorvoetend komt er nu een halfhartige aanpak tot stand. Het Witte Huis heeft de 'AI Czar' vervangen; de *pro-business accelerationist* David Sacks heeft plaatsgemaakt voor de meer veiligheidsgeoriënteerde Jay Clayton.[24] China en de Verenigde Staten hebben afgesproken elkaar op de hoogte te stellen van incidenten.[25] En Trump heeft een akkoord gesloten met een aantal leidende labs, dat hen aanmoedigt tot meer auditing. Dit akkoord is echter slechts in morele zin bindend; juridische consequenties of

3. planned-obsolescence.org/p/the-hugging-face-attack-surprised.
4. theguardian.com/technology/2024/dec/27/godfather-of-ai-raises-odds-of-the-technology-wiping-out-humanity-over-next-30-years.
5. en.wikipedia.org/wiki/P(doom).
6. x.com/hilbertspaess/status/2097476196791709843?lang=en.
7. darioamodei.com/post/we-must-pace-the-frontier.
8. x.com/sama/status/2098811563415150910.
9. x.com/EvanHub/status/2097497037956891126.
10. nos.nl/artikel/2631010-trump-en-china-doen-waarschuwing-ai-topmannen-af-als-bangmakerij.
11. politico.com/news/2026/09/15/a-starting-gun-ai-leaders-push-for-independent-safety-audits-01078262; https://x.com/guilleflorvs/status/2099257149155123706.
12. mustafa-suleyman.ai/a-warning-about-model-welfare.
13. youtube.com/watch?v=HjurAWAr_nY.
14. anthropic.com/news/investigating-incidents-cybersecurity-evals anthropic.com/news/investigating-incidents-cybersecurity-evals.
15. research.meta.ai/blog/addressing-third-party-testing-misconfiguration-muse-spark-1-1.
16. nytimes.com/2026/09/16/technology/openai-model-safety-guardrails.html.
17. nos.nl/artikel/2632243-australische-premier-openai-infiltreerde-in-australische-overheidswebsite.
18. bbc.com/news/articles/cw62jje658dlo.
19. pbs.org/newshour/nation/openai-reveals-concerning-new-ai-behavior-and-vows-to-track-it-more-closely.
20. docs.google.com/document/d/e/2PACX-1vQNl3SEX5I-yA6d9qHjjFZN-qzGRZNFI6b63g-yu1Fy-ZYkVfCWm7i9WXRXw63m6yDB_auDuPLyQ7jBm/pub; x.com/DKokotajlo/status/2099600298855829616.
21. youtube.com/watch?v=A_156w0aYtU.
22. nbcnews.com/tech/security/anthropic-project-glasswing-mythos-preview-claude-gets-limited-release-rcna267234.
23. wsj.com/tech/ai/openai-chatgpt-model-release-cancel-safety-5a2f9f42.
24. cnbc.com/2026/10/03/trump-jay-clayton-ai-czar.html.
25. pbs.org/newshour/world/china-and-u-s-agree-to-establish-ai-safety-channel-and-continue-trade-and-military-talks.

sancties zijn niet voorzien, waardoor het de vraag is hoe effectief het zal zijn.[26] Het gevoel dat dit initiatief *window dressing* betreft wordt bevestigd door het feit dat het Witte Huis heeft besloten om voortaan niet langer van Artificial, maar van Super Intelligence te spreken, omdat AI door alle affaires een slechte naam heeft gekregen.[27] Ondertussen gaat de inzet van AI dan ook in volle vaart door, juist ook in gevoelige domeinen. Zo heeft Hegseth aangekondigd nog meer te willen inzetten op autonome wapensystemen in oorlogsvoering[28] en heeft Meta zijn persoonlijke agent op de markt gebracht die door middel van alle sociale, financiële en medische gegevens van een persoon als zijn persoonlijke assistent e-mails kan versturen, boodschappen kan doen en telefoongesprekken kan voeren.[29]

Natuurlijk is het redelijk om te stellen dat de bedrijven dit soort *rogue* agents simpelweg moeten tegengaan en dat zij aansprakelijk kunnen worden gesteld voor hacks en andere schade die ze aanrichten, maar tot nu toe heeft dat de labs er niet van weerhouden om steeds verder te gaan in *recursive self-improvement* (AI-systemen die steeds betere en snellere AI-systemen maken), onder meer omdat de labs geloven dat zij in een race verwikkeld zijn met elkaar en dat de eerste die op het punt van *AGI* (*Artificial General Intelligence*, het moment dat AI capaciteiten heeft die gelijk zijn aan die van de mens) of zelfs *superintelligence* (AI streeft de mens voorbij) geraakt de winnaar is en dominantie kan claimen. Natuurlijk is het redelijk om te stellen dat China en de Verenigde Staten, de twee leidende mogendheden op het gebied van AI, samen moeten zitten om strakkere regulering aan te nemen, maar ook zij geloven dat ze in een race met elkaar zijn verwikkeld en geloven dat wie de beste AI heeft economisch en militair superieur zal zijn. Ook zij hebben tot nu toe geen indicatie gegeven bereid te zijn tot ferme afspraken te komen. En natuurlijk zou de EU simpelweg dit soort AI-systemen kunnen weren van de interne markt, maar tot nu toe is Europa hoe langer hoe meer afhankelijk van Amerikaanse tech en is het de vraag of het houdbaar is om Europese agents niet of minder autonoom te laten handelen als andere mogendheden dat wel doen en hun agents bijvoorbeeld Europese bedrijven en staten laten hacken om informatie te ontfutselen en infrastructuur te compromitteren.

Langzamerhand begint ook in Den Haag het besef door te dringen dat AI niet uitsluitend een innovatie- of digitaliseringsvraagstuk is. Het kabinet heeft inmiddels een Nederlandse Digitaliseringsstrategie ontworpen waarin verantwoord gebruik van AI, digitale autonomie, eigen rekenkracht en Nederlandse of Europese taalmodellen expliciete prioriteiten zijn; daarnaast wordt geïnvesteerd in een AI-fabriek in Groningen.[30] De Uitvoeringswet AI-verordening, die voor eind 2026 is voorzien, regelt het Nederlandse toezicht op de Europese AI-verordening, waarbij onder meer de Autoriteit Persoonsgegevens en de Rijksinspectie Digitale Infrastructuur coördinerende rollen krijgen.[31] Voor algoritmen van de overheid bestaan inmiddels het algoritmeregister[32] en het Impact Assessment Mensenrechten en Algoritmes[33] en wordt onderzocht of de Algemene wet bestuursrecht dient te worden gewijzigd om meer transparantie en rechtsbescherming bij algoritmische besluitvorming te creëren.[34] Ook kiest het kabinet een nadrukkelijkere internationale positie: in juli 2026 verscheen een internationale AI-strategie die inzet op Europese AI-capaciteit, vermindering van strategische afhankelijkheid en internationale afspraken op het gebied van veilige en verantwoorde AI.[35] En ook politiek komt AI op de agenda. Dassen van Volt riep tijdens de Algemene Politieke Beschouwingen op tot meer Europese investeringen in AI die op Europese waarden is gestoeld.[36] Enkele maanden daarvoor had hij al een initiatiefnota ingediend met uiteenlopende voorstellen, van een internationale AI-coalitie en een AI-raad tot meer digitale expertise voor parlement en rechtspraak, een AI-belasting en een actiever Europees industrie- en aanbestedingsbeleid.[37] Deze voorstellen lieten zien dat er nog geen politieke consensus bestaat: een motie om meer in Europese fysieke AI te investeren kreeg in juli een Kamermeerderheid, terwijl voorstellen voor een Europese AI-spoedtop en een AI-belasting werden verworpen.[38]

Langzamerhand begint ook in Den Haag het besef door te dringen dat AI niet uitsluitend een innovatie- of digitaliseringsvraagstuk is

De vraag is of deze initiatieven voldoende antwoord bieden op de ontwikkeling rondom agentic AI. Op zijn minst drie aanvullende aspecten zouden moeten worden overwogen.

Allereerst zou Nederland, wellicht in Europees verband, de leiding kunnen nemen in het ontwikkelen van een serieus alternatief voor de Amerikaanse AI-modellen, dat veel verder gaat dan de huidige voornemens. Het Franse Mistral kan daarvoor een startpunt bieden;[39] GPT-NL ook.[40] Ter vergelijking: wereldwijd wordt er in 2026 alleen al $ 2,5 biljoen geïnvesteerd in de ontwikkeling van AI.[41] Dat bedrag zal de komende jaren naar verwachting exponentieel stijgen. Ook specifieke bedrijven hebben veel investeringsruimte. In de laatste investeringsronde alleen al haalde OpenAI $ 122 miljard aan toegezegd kapitaal op, met een post-moneywaardering van $ 852 miljard.[42] Anthropic deed dit jaar twee investeringsrondes: de eerste leverde $ 30 miljard op tegen een waardering van $ 380

Initiatief	Relevante recente financiering	≈ in euro	Verhouding tot GPT-NL
OpenAI	$ 122 mrd., ronde 2026	€ 108,6 mrd	~8000×
Anthropic	$ 95 mrd, twee rondes 2026	€ 84,6 mrd	~6300×
Mistral	€ 3 mrd, Series D 2026	€ 3,0 mrd	~222×
GPT-NL	publieke projectfinanciering	€ 13,5 mln	1×

miljard en vervolgens in mei nog eens $ 65 miljard, nu tegen een post-moneywaardering van $ 965 miljard.[43] Ter vergelijking, Mistral haalde in september 2026 € 3 miljard op tegen een post-moneywaardering van meer dan € 21 miljard[44] en voor GPT-NL is een startbedrag van € 13,5 miljoen beschikbaar gesteld.[45]

Zolang er geen serieuze Europese alternatieven zijn, blijven Nederlandse burgers, bedrijven en overheidsorganisaties afhankelijk van met name Amerikaanse modellen. Dat vergroot hun voorsprong, omdat de modellen worden gevoed met data en taken van Nederlandse en Europese partijen. Tegelijkertijd is het gevaar dat door het uitbesteden van taken aan AI-modellen Europese personen en organisaties gaandeweg de kennis verliezen om die taken zelf uit te voeren: het zogenoemde *cognitive offloading*.[46] Omdat agentic AI zal worden ingezet in een breed scala aan sectoren, van baanvervanging in de financiële dienstverlening tot oorlogvoering, van het ontwikkelen van nieuwe medicijnen tot het plaatsen van gevoelige hacks, van het bestrijden van eenzaamheid onder ouderen tot het exploreren van andere planeten, is de vraag hoe strategische autonomie gewaarborgd kan blijven als de investeringen in Nederlandse en Europese tech zo laag blijven en de afhankelijkheid van niet-Europese modellen zo hoog blijft.

Ten tweede, en daarmee samenhangend, zou Nederland de leiding kunnen nemen in het ontwikkelen van een AI-model waarin Europese waarden centraal staan. Dat vergt niet alleen een forse investering, maar ook een minder flexibele houding ten aanzien van modellen die zich niet aan die waarden conformeren. Zolang de belangrijkste modellen niet in Europa worden getraind, ontwikkeld en vermarkt, zal dat ingewikkeld blijven. De ervaring van de laatste decennia heeft laten zien dat Nederland en de EU pas na jaren in actie komen tegen de tech-sector met een verbod of boete, die vervolgens wordt aangevochten voor een rechter, tot aan de hoogste instantie, soms met succes. Zelfs in de gevallen waarin de maatregel of boete uiteindelijk standhoudt, heeft het bedrijf al tien jaar of langer de tijd gekregen om een product of dienst uit te rollen en daarmee marktdominantie te verkrijgen. Het product of de dienst is tegen die tijd breed in de samenleving geïntegreerd en bedrijven hebben tegen die tijd vaak al een vervangend product of dienst klaarstaan.

Niet-Europese AI-modellen worden toegelaten op de interne markt terwijl het duidelijk is dat die tot stand zijn gekomen op een wijze die in strijd is met de rechtsstaat

In de praktijk lijkt er bovendien een dubbele houding te zijn. Niet-Europese AI-modellen worden toegelaten op de interne markt en gebruikt door overheidspartijen, bedrijven en burgers, terwijl het duidelijk is dat die tot stand zijn gekomen op een wijze die in strijd is met de rechtsstaat. Om maar een van de vele voorbeelden te noemen: de modellen zijn getraind door grote delen van het internet en andere openbare informatiebronnen te schrapen, in weerwil van privacy, gegevensbescherming, intellectueel eigendom en andere op de informatie rustende rechten. De EU en Nederland laten deze modellen niet alleen toe;[47] indirect faciliteren zij de overdracht van kennis, informatie en data door steeds maar meer in te zetten op het openbaar en vrijelijk ter beschikking stellen van

26. reuters.com/world/us/trump-releases-ai-accord-with-tech-executives-2026-09-29/.

27. whitehouse.gov/presidential-actions/2026/09/inaugurating-the-era-of-super-intelligence/.

28. reuters.com/world/pentagon-creates-autowarcom-expand-ai-drone-capabilities-2026-09-30/.

29. about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/.

30. Aanbiedingsbrief Nederlandse Digitaliseringsstrategie 4 juli 2025 open.overheid.nl/documenten/01cf52fa-ca6c-4590-8f8c-3370c05a7f35/file De Nederlandse Digitaliseringsstrategie Samen versnellen https://open.overheid.nl/documenten/51bf0136-69cc-4d37-90c9-64d7ed3d9a5c/file.

31. internetconsultatie.nl/uaiv/b1.

32. algoritmes.overheid.nl/nl.

33. Impact Assessment Mensenrechten en Algoritmes Versie 2 2026 open.overheid.nl/documenten/d0947c02-81df-4c00-83df-d88b703025f4/file.

34. *Kamerstukken II* 2025/26, 36915 VI, nr. 3, tweedekamer.nl/downloads/document?id=2026D18912; *Kamerstukken II* 2024/25, 26643, nr. 1372, zoek.officielebekendmakingen.nl/kst-26643-1372.pdf.

35. De Nederlandse Internationale AI-Strategie AI voor waarden, welvaart en weerbaarheid open.overheid.nl/details/3eb5ccd2-cccd-4e6d-a02b-dd9a88f238b0.

36. Plenair verslag Tweede Kamer, 2e vergadering woensdag 16 september 2026, tweedekamer.nl/kamerstukken/plenaire_verslagen/detail/2026-2027/2.

37. *Kamerstukken II* 2025/26, 36974, nr. 1, tweedekamer.nl/downloads/document?id=2026D32671; *Kamerstukken II* 2025/26, 36974, nr. 2, tweedekamer.nl/downloads/document?id=2026D32673; *Kamerstukken II* 2025/26, 32852, nr. 425.

38. Tweede Kamer, 90e vergadering donderdag 2 juli 2026, tweedekamer.nl/kamerstukken/plenaire_verslagen/detail/2025-2026/90.

39. mistral.ai/.

40. gpt-nl.nl/.

41. gartner.com/en/newsroom/press-releases/2026-1-15-gartner-says-worldwide-ai-spending-will-total-2-point-5-trillion-dollars-in-2026.

42. openai.com/nl-NL/index/accelerating-the-next-phase-ai/?_bhlid=c21f9f982b8d3f7bf925c3630a22f246eb73dd7d.

43. anthropic.com/news/anthropic-raises-30-billion-series-g-funding-380-billion-post-money-valuation?s=prompt+injection anthropic.com/news/series-h.

44. mistral.ai/news/mistral-makes-sovereign-open-weight-ai-to-frontier/.

45. gpt-nl.nl/gpt-nl/.

46. M. Gerlich, AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 2015, *15*(1), 6; G. Chirayath, K. Premamalini & J. Joseph, Cognitive offloading or cognitive overload? How AI alters the mental architecture of coping, *Frontiers in Psychology*, 2025, *16*, 1699320; P. Shukla, P. Bui, S. Levy, M. Kowalski, A. Baigelenov & P. Parsons, 'Deskilling, cognitive offloading, and misplaced responsibilities: potential ironies of AI-assisted design', in: *Proceedings of the extended abstracts of the CHI conference on human factors in computing systems*, 2025, p. 1-7.

47. EDPB Advies 28/2024 over bepaalde aspecten van gegevensbescherming in verband met de verwerking van persoonsgegevens in het kader van AI-modellen, vastgesteld op 17 december 2024, edpb.europa.eu/system/files/documents/2025-05/edpb_opinion_202428_ai-models_nl.pdf; Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

data door middel van initiatieven rond de open overheid,[48] hergebruik van overheidsinformatie,[49] door organisaties aan te moedigen niet-persoonlijke gegevens openbaar te maken[50] en doordat belangrijke instituten, zoals archieven, hun bronnen digitaliseren en doorzoekbaar maken.

Tegelijkertijd worden de geldende regels wel streng gehandhaafd ten opzichte van Europese initiatieven. Een sprekend voorbeeld is een Nederlands initiatief waarin een database voor het trainen van een taalmodel werd aangelegd, maar al na een aantal weken moest stoppen omdat dit in strijd was met het intellectueel eigendomsrecht dat op de informatie rustte.[51] Daarmee hebben we het slechtste van twee werelden. We hebben én geen vitale techeconomie of capabel AI-model én onze mensenrechten worden alsnog flagrant geschonden, alleen dan door Amerikaanse bedrijven, die wellicht over tien of twintig jaar een boete moeten betalen van een aantal miljoen, terwijl zij dat bedrag tegen die tijd in veelvoud hebben terugverdiend.

Wil Nederland, in Europees verband, deze situatie omkeren, dan zal het moeten overwegen om een radicale keuze te maken. Dat zou kunnen door de rechtsstaat strikt te handhaven, waaruit volgt dat vrijwel alle niet-Europese AI-modellen simpelweg geen toegang krijgen tot de interne markt en niet mogen worden gebruikt door overheidsdiensten, bedrijven of burgers, ook niet middels sluiproutes en VPN's. Dat zal uiteraard geen kosteloze keuze zijn. De eerste jaren zullen Nederland en Europa sterk achterblijven in de AI-race, terwijl de VS en China snelle ontwikkelingen zullen doormaken. Het zal zeker vijf tot tien jaar duren totdat een alternatief AI-model gelijke capaciteiten heeft. Toch is dat wel mogelijk. China laat zien dat met een gerichte focus op kernsectoren het haalbaar is om in een aantal jaar leidend te worden op het gebied van geavanceerde technologie.[52] Juist voor AI is bekend dat het ontwikkelen van een geavanceerd model in feite neerkomt op een enorme capaciteit aan rekenkracht (*compute*) en data. Dat is een kwestie van investeren.

Een verbod op niet-Europese AI-modellen geeft bovendien een enorme boost aan de ontwikkeling van een alternatief AI-model. Een van de grootste knelpunten is momenteel dat de vraag naar een verantwoord AI-model simpelweg te klein is en het rechtsstatelijke karakter van een AI-model voor de meeste organisaties uiteindelijk niet doorslaggevend is: kosten, capaciteit en efficiëntie blijken in de praktijk belangrijker. Een verbod vangt dus twee vliegen in één klap: én de rechtsstaat wordt beschermd én er ontstaat een markt voor een Europees alternatief. Het verkleint ook de afhankelijkheid van de grillen van buitenlandse regimes, zoals onder meer duidelijk werd uit de door de Amerikaanse overheid geïnstigeerde blokkade van digitale diensten bij het Internationaal Strafhof.[53]

Een verbod vangt dus twee vliegen in één klap: én de rechtsstaat wordt beschermd én er ontstaat een markt voor een Europees alternatief

Een eigen AI-model betekent bovendien de ruimte om economische effecten van AI te herverdelen en bij te sturen. Allereerst kunnen de creatieve sector, journalistiek, wetenschap en andere belangrijke sectoren die data produceren waarop een Europees AI-model wordt getraind, worden vergoed voor hun bijdrage. Er ontstaat beter zicht op het energiegebruik van het trainen van datamodellen en de taken waarvoor zij worden ingezet, waardoor gerichte belastingen kunnen worden geheven. En de winst die met AI wordt gemaakt kan eenvoudiger worden belast, wat essentieel is omdat er een gevaar is dat veel banen door AI zullen worden vervangen (de zogenoemde *Jobpocalypse*).[54] In voorgaande technologische revoluties was de economische disruptie al flink, maar werd het verlies van banen na een aantal decennia weer gecompenseerd door de opkomst van nieuwe beroepen. Het is waarschijnlijk dat dat ook bij AI het geval zal zijn, maar de vrees dat daar deze keer niet of in mindere mate sprake van zal zijn is ook niet irreëel omdat AI op een aantal punten fundamenteel verschilt van voorgaande technologieën. Het kan extreem autonoom opereren, eigen taken en doelen stellen en kan door de opkomst van robotica niet alleen fysieke banen vervangen,[55] maar ook cognitieve, creatieve en sociale beroepen.[56]



Reclame op een bushokje in San Francisco © Bart van der Sloot

Ten derde zouden juridische begrippen moeten worden herijkt. Zelfs met een Europees AI-model is het een illusie om te denken dat de oude concepten zullen volstaan. Allereerst is het waarschijnlijk dat onze samenleving op een aantal punten fundamenteel zal veranderen door de invloed van AI. Hoe, weten we niet. De capaciteiten van AI zijn zo sterk in ontwikkeling en de meningen over welke

impact AI zal hebben op de samenleving lopen zo sterk uiteen dat het belangrijk is om in scenario's in plaats van zekerheden te denken. Zelfs al is de kans dat AI de banenmarkt fundamenteel gaat opschudden of een catastrofe zal veroorzaken klein, het is een scenario met zulke grote gevolgen dat die in ieder geval serieus moeten worden doordacht en er een juridische en politieke respons klaar moet staan voor het moment dat er voortekenen zichtbaar zijn. Ten tweede zijn onze huidige juridische standaarden geënt op technologieën in klassieke zin; dingen die geen autonome beslissingen nemen, die niet op vrijwel alle data over mensen, hun gedrag en hun samenlevingen zijn getraind en niet expliciet worden gemodelleerd en getraind om door mimesis menselijk gedrag te kopiëren. Aangezien de meerwaarde van agentic AI juist is dat het autonoom kan handelen en als een mens kan denken en handelen, is het de vraag of de oude juridische uitgangspunten nog houdbaar zijn. Ook daarbij is het risico dat het slechtste uit twee werelden ontstaat: de juridische regels worden formeel in stand gelaten, maar betekenen in de praktijk nog weinig, terwijl er geen nieuwe regels worden vervaardigd die beter aansluiten op de gevaren en de risico's van de nieuwe realiteit.

Een voorbeeld is het inmiddels bijna rituele beroep op de *human-in-the-loop*.[57] Wat betekent dat nog bij steeds zelfstandigere systemen, die steeds complexer en capabeler worden? Daarbij moet worden bedacht dat we in tien jaar zijn gegaan van systemen die zeer beperkte functionaliteit hadden, naar agents die op steeds meer punten de mens overstijgen. Niets doet vermoeden dat de ontwikkelingen de komende tien jaar minder hard zullen gaan. Integendeel. Nu al kampen gemeenten, rechterlijke instanties, verzekeraars en andere instanties met zo'n toename van door AI gegenereerde klachten en procedures dat het vrijwel ondoenlijk is om die nog door mensen te laten beoordelen. Dit kan hen ertoe nopen om AI-agents in te zetten om een eerste selectie te doen of om te verifieren bij de burger dat de door AI gegenereerde klacht of het door AI gegenereerde pleidooi ook zijn daadwerkelijke rechtsbedoeling weergeeft.

Wat betekent het principe van transparantie nog als systemen niet alleen veelal onbegrijpelijk zijn voor zelfs de beste programmeurs, maar het ook steeds onzekerder wordt of wat wij in de systemen waarnemen aan redenatietrajecten en beslismomenten überhaupt nog werkelijk weergeeft wat er omgaat in de agents of dat het weergeeft wat de agents willen dat mensen zien?

Wat betekent het idee van menselijke aansprakelijkheid in een wereld van agentic AI en wat zeggen leerstukken omtrent toerekenbaarheid, causaliteit en schade? Het idee dat een mens of organisatie aansprakelijk kan worden gesteld voor het gedrag van agents kan uiteindelijk een vorm van *ethics-washing* met zich meebrengen, waarbij actoren verantwoordelijk worden gehouden voor concrete gedragingen die zij niet konden voorzien of beheersen.

Ook klassieke ideeën over legaliteit, rechtszekerheid en voorzienbaarheid kunnen onder druk komen te staan. Het recht veronderstelt doorgaans dat burgers en organisaties vooraf in redelijke mate kunnen weten welke regels gelden en welke juridische gevolgen hun handelen heeft. Maar wat betekent voorzienbaarheid als autonome systemen hun strategie aanpassen aan veranderende omstandigheden, andere agents inschakelen en oplossingen vinden die noch de programmeur, noch de gebruiker vooraf heeft bedacht? En wat betekent het legaliteitsbeginsel wanneer niet langer alleen de wet of een menselijke ambtenaar bepaalt hoe een norm wordt toegepast, maar een adaptief systeem zelf een groot aantal tussenbeslissingen neemt?

Zelfs basale privaatrechtelijke begrippen als wil, toestemming en vertegenwoordiging kunnen door de opkomst van agentic AI anders komen te liggen. Het recht kan relatief eenvoudig omgaan met een softwareprogramma dat exact uitvoert wat een gebruiker heeft opgedragen. Moeilijker wordt het wanneer een agent zelfstandig onderhandelt, subagents inschakelt, contracten sluit en daarbij een strategie volgt die de gebruiker nooit heeft voorzien. Is zo'n overeenkomst nog de uitdrukking van de wil van de gebruiker, is de agent slechts een technisch verlengstuk van die wil, of moet het recht accepteren dat tussen menselijke opdracht en juridische handeling steeds meer autonome beslissingsruimte ontstaat? Nog complexer worden deze vraagstukken met de opkomst van *zero-human companies*.[58] Ook hier kan het kunstmatig worden om met traditionele leerstukken simpelweg te doen alsof iedere beslissing uiteindelijk door een mens is genomen.

Hoe het ook loopt, het zullen naar alle waarschijnlijkheid interessante tijden zijn. •

48. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.

49. Wet hergebruik van overheidsinformatie.

50. Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union (Text with EEA relevance.) Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

51. nos.nl/artikel/2533113-stichting-breinhaalt-grote-hoeveelheid-illegale-data-voortrainen-ai-offline.

52. chathamhouse.org/sites/default/files/2026-09/2026-09-30-china-europe-industrial-future-kynge.pdf.

53. nytimes.com/2026/01/10/world/europe/icc-judges-us-sanctions-trump.html.

54. K. Abid & T. Garavan, 'Smarter Machines, Disappearing Careers? The Jobpocalypse Phenomenon: Paradoxes and Implications for HRD Theory, Research and Practice', *Human Resource Development Review* 2026, 15344843261445507; J. Chen & H. Chen (2026), 'Jobpocalypse or techno-utopia? geospatially decoding public concerns through the social media noise in AI's disruption era', *Information Processing & Management* 2026, *63*(3), 104550; G.A. Bolles, *The next rules of work: The mindset, skillset and toolset to lead your organization through uncertainty*, Kogan Page Publishers, 2021.

55. nytimes.com/2025/10/21/technology/inside-amazons-plans-to-replace-workers-with-robots.html.

56. commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf.

57. Zie bijvoorbeeld art. 22 AVG. Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrij verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (Algemene verordening gegevensbescherming).

58. forbes.com/sites/anishasircar/2026/06/10/ai-owned-companies-argentina/.