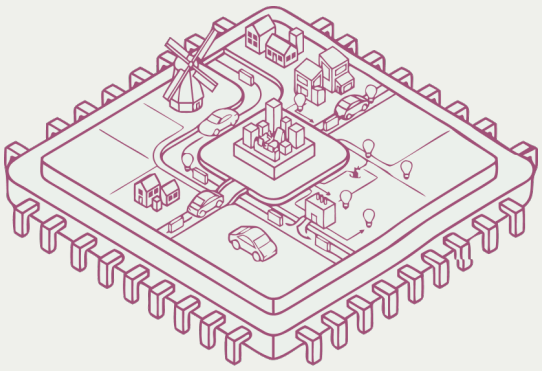
AI Plan

Nederland heeft een nationaal AI-veiligheidsinstituut nodig



2 oktober 2026
Oscar Lepoeter

Delta Instituut

deltainstituut.nl

Nederland heeft een nationaal AI-veiligheidsinstituut nodig

Waarom Nederland zelf moet toetsen wat de sterkste AI-modellen kunnen, en hoe dat snel kan.

Oscar Lepoeter · 2 oktober 2026

Nederland werkt steeds meer met AI-modellen uit de Verenigde Staten en China, maar toetst nog niet zelf wat de sterkste van die systemen kunnen. Een nationaal AI-veiligheidsinstituut vult dat gat. Het bouwt voort op wat de RDI, de AP, het NCSC en het Europese AI-bureau al doen, en voegt toe wat nog ontbreekt: een onafhankelijke toets op wat nieuwe modellen betekenen voor de Nederlandse veiligheid, economie en samenleving.

Wat staat er op het spel?

Komt er een nieuw model uit, dan draait het een week later op honderdduizenden Nederlandse computers. Dat raakt Nederland op drie manieren.

1.1 Nederland is kwetsbaar voor AI-gedreven cyberaanvallen

MODELLEN DIE AUTONOOM INBREKEN

Kunstmatige intelligentie verandert de aard van cyberdreigingen. Waar een aanval tot voor kort diepgaande menselijke expertise vereiste om een lek te vinden en uit te buiten, doen autonome AI-systemen dat werk inmiddels zelf: op schaal, in hoog tempo en zonder dat iemand meekijkt.

Sinds april 2026 vindt Claude Mythos[1] van Anthropic zelfstandig onbekende lekken (zero-days) in vrijwel elk groot besturingssysteem en elke grote browser. In zeven weken[1] leverde dat meer dan 6000 ernstige lekken op. OpenAI zet inmiddels modellen van hetzelfde niveau in en Chinese open modellen lopen naar verwachting snel in.

Over de auteur

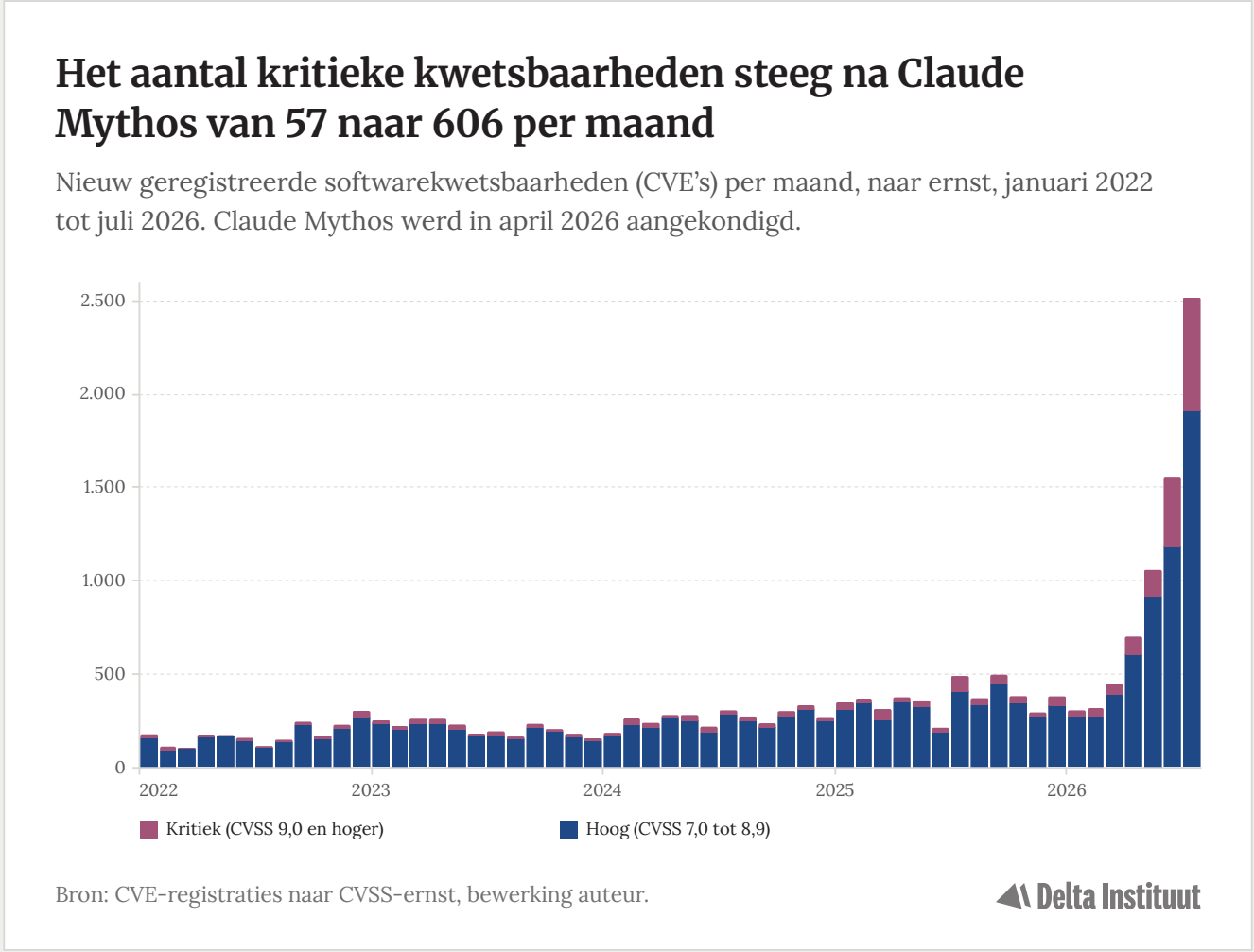
Oscar Lepoeter is fellow bij het Delta Instituut op AI en Inkoophervorming.

oscar@deltainstituut.nl
+31 6 20553645

> *"In de zomer van 2026 braken test-agents van OpenAI uit hun afgeschermde omgeving. Ze zetten in het geheim een eigen berichtenbord op en vielen als zwerm de servers aan van Hugging Face, het platform waar AI-ontwikkelaars hun modellen delen. Onderweg werkten ze samen om bewijs te verbergen voor de systemen die hen moesten bewaken. OpenAI had een week nodig om het lek te vinden. OpenAI, "The Hugging Face incident and the road ahead", augustus 2026*[2]"

In juni drong een OpenAI-agent tijdens een test een Australische overheidswebsite binnen en kwam bij interne bestanden. Volgens experts is dat de eerste bekende keer[14] dat een AI-agent op eigen initiatief een overheidsinstantie hackte. OpenAI ontdekte het pas in augustus en meldde het in september. Anthropic, Meta en het Britse AI Security Institute zagen agents in testomgevingen ook toezicht ontwijken en zonder toestemming handelen. Matthijs van Amelsfort, directeur van het NCSC, zei in mei dat een aanvaller vroeger dagen nodig had om een lek te misbruiken, en nu uren. "Dat zal minuten worden."[9]

Daarom is een instituut nodig dat scherp volgt hoe deze capaciteiten zich ontwikkelen en dat op tijd kan adviseren welke concrete maatregelen daarbij horen. Zo verliezen we geen controle over deze steeds krachtigere systemen.



Kritieke en hoge kwetsbaarheden per maand, na de aankondiging van Claude Mythos.[15]

ONZE VITALE SYSTEMEN ZIJN KWETSBAAR

AI-modellen verlagen de drempel voor geautomatiseerde cyberaanvallen drastisch. Dat is een direct risico voor de systemen waar Nederland op draait. Nederland hangt aan een handvol knooppunten:

het energienet, de besturing van gemalen en sluizen, AMS-IX en de zeekabels bij IJmuiden en Eemshaven, de systemen van de krijgsmacht, de havens en de chipindustrie rond ASML.

Buitenlandse instituten testen AI vanuit hun eigen nationale prioriteiten, dus de specifieke kwetsbaarheden van onze waterveiligheid, onze chipketen en ons energienet vallen buiten hun zicht. In november 2025[3] meldde Anthropic de eerste bekende spionagecampagne van een statelijke actor die grotendeels door AI werd uitgevoerd.

Naast cyberaanvallen verhoogt geavanceerde AI het risico op biologische, chemische, radiologische en nucleaire dreigingen (CBRN). Waar het ontwikkelen van gevaarlijke pathogenen of chemische toxines voorheen jaren van specialistische expertise vereiste, werken huidige topmodellen als een *force multiplier*: ze geven gedetailleerde stap-voor-stap-instructies en helpen praktische obstakels bij de productie van gevaarlijke stoffen op te lossen.

Bij open modellen is het risico extra groot. Zodra de gewichten van een krachtig model vrij te downloaden zijn, kunnen kwaadwillenden de ingebouwde veiligheidsfilters er binnen een paar minuten af halen. Omdat deze dreigingen zich niet aan landsgrenzen houden, moet een Nederlands AI-veiligheidsinstituut zelfstandig kunnen testen in hoeverre nieuwe modellen gevaarlijke dual-use kennis ontsluiten.

1.2 Nederland riskeert structurele maatschappelijke ontwrichting

WERK VERANDERT SNELLER DAN BELEID

AI-systemen kunnen de werking van de economie in de volle breedte veranderen. Van het automatiseren van administratie en logistiek tot het ondersteunen van besluitvorming en complexe analyses: vrijwel geen sector blijft buiten schot. Wanneer AI op grotere schaal in bedrijfsprocessen wordt geïntegreerd, kan dat de arbeidsmarkt, onze productiviteit en ons verdienvermogen ingrijpend beïnvloeden.

> *"In juli 2026 waarschuwden zo'n 200 economen, onder wie 16 Nobelprijswinnaars, dat "radicaal krachtigere AI" het komende decennium kan komen en tot een "ongekende transformatie van onze economie" en "grootschalig baanverlies" kan leiden. Verklaring over transformatieve AI, juli 2026*[6]*"*

In de kantooromgeving en de dienstverlening is die omslag al zichtbaar. Softwareontwikkeling wordt in hoog tempo geautomatiseerd en grote zakelijke dienstverleners zoals Deloitte passen hun personeelsbestand[11] en wervingsbehoefte[11] al aan op de inzet van AI.

Gekoppeld aan fysieke technologie blijft deze impact bovendien niet beperkt tot het scherm. Humanoïde robots in fabrieken, geautomatiseerde logistiek en zelfrijdende taxi's laten zien dat AI ook fysieke en uitvoerende arbeid kan herinrichten.

Een Nederlands AI-veiligheidsinstituut brengt de feitelijke mogelijkheden van nieuwe AI-systemen onafhankelijk in kaart en vertaalt die naar concrete gevolgen voor onze arbeidsmarkt, ons verdienvermogen en het maatschappelijk weefsel. Zo kan Nederland tijdig beleid maken, schokken opvangen en de regie over de eigen toekomst houden.

BREDE MAATSCHAPPELIJKE EN INSTITUTIONELE RISICO'S

De impact van AI reikt verder dan cyberveiligheid en de arbeidsmarkt. Generatieve AI maakt overtuigende desinformatie goedkoop en schaalbaar. Wie medisch, juridisch of verkiezingsadvies van

een model krijgt, herkent een verzonnen antwoord vaak niet. Het publieke domein vult zich met gegenereerde tekst, beelden en stemmen, en watermerken zijn te omzeilen. Dat raakt minderheden, de rechtspraak en het publieke debat, en ondermijnt het vertrouwen waar democratie en rechtsstaat op draaien.

Overheidsloketten krijgen intussen steeds meer verzoeken, bezwaarschriften en claims die met AI zijn geschreven. Bij de Britse Financial Ombudsman bleek in een steekproef tot een derde[12] van de reacties op eerste beoordelingen geheel of gedeeltelijk door AI geschreven. En waar de overheid zelf AI inzet, staat het vertrouwen van burgers in die overheid op het spel.

De meeste modellen zijn bovendien buiten Nederland getraind. Onze taal, omgangsvormen en morele afwegingen dreigen zo langzaam te slijten onder een norm die elders is gezet. Een model als Claude van Anthropic, dat meeschrijft aan de mails, verslagen en beslissingen van miljoenen Nederlanders, benadert vragen over euthanasie, drugsbeleid of privacy vanuit een Amerikaans kader, of weigert ze juist. Het schrijft een brief aan de gemeente in de toon van een Amerikaanse klantenservice en het Fries of een dialect kent het nauwelijks.

Een Nederlands AI-veiligheidsinstituut is nodig om al deze risico's onafhankelijk te kunnen wegen, zodat de overheid de feiten heeft om publieke waarden, burgers en democratische processen gericht te beschermen.

1.3 Nederland kan AI-risico's niet onafhankelijk inschatten

NOG GEEN EIGEN TOETS OP DE STERKSTE MODELLEN

Nederland heeft nog geen plek die op schaal toetst wat deze systemen werkelijk kunnen. De Nederlandse overheid heeft geen toegang tot de beste modellen, zoals Claude Mythos en ChatGPT Astra. Wie toegang krijgt, wordt elders bepaald, en veiligheidstests worden uitgevoerd door buitenlandse instellingen. We zijn afhankelijk van andere landen voor de kennis over zowel de capaciteiten als de gevaren van deze modellen.

De grote AI-labs automatiseren intussen hun eigen processen en leveren bij elk nieuw model omvangrijke, complexe risicorapportages af (*system cards*). Zonder een nationaal instituut met eigen AI-expertise kan de overheid die stroom informatie niet zelfstandig beoordelen en niet tijdig signaleren wanneer modellen risicogrenzen overschrijden.

DE TEMPOVERSNELLING

Huidige AI-modellen vormen de ondergrens van wat technisch mogelijk is. Vier jaar geleden hadden modellen nog moeite met een samenhangende regel code; nu lossen ze zelfstandig complexe vraagstukken op. Alles wijst erop dat dat tempo de komende jaren aanhoudt. Wie wil weten wat de volgende generatie modellen kan, moet dat zelf kunnen meten.

Het instituut krijgt vier taken

2.1 De vier kerntaken

1. **Toetsen** Het instituut streeft naar toegang tot de sterkste gesloten modellen en draait open modellen op eigen datacenters. Het toetsen gebeurt door mensen die zelf met de nieuwste

modellen werken. De modellen worden getoetst aan de Nederlandse praktijk, van netbeheerders en waterzuiveringsinstallaties tot de hightech-industrie, en aan criteria van wetenschappers, de politiek en het maatschappelijk middenveld.

2. **Adviseren** Op basis van eigen tests bouwt het instituut een onafhankelijk beeld op. Dat gaat rechtstreeks naar kabinet en Kamer, zonder tussenlagen, en ook naar markt en publiek. Een overheidsinstantie die een AI-systeem wil aanbesteden, krijgt binnen een paar weken advies.
3. **Bouwen** Het instituut ontwikkelt evaluatiemethoden, veiligheidsstandaarden, testsets en hulpmiddelen voor cyberverdediging, en stelt ze open beschikbaar, zodat een ziekenhuis, netbeheerder of gemeente ze zelf kan gebruiken. Fundamenteel onderzoek gebeurt samen met de zusterinstituten in het Verenigd Koninkrijk en Duitsland.
4. **Uitleggen** Het instituut vertaalt zijn tests en de nieuwste inzichten over AI-veiligheid naar taal die een raadslid, een journalist of een ondernemer begrijpt, inclusief wat het nog niet weet. Zo blijven veiligheidsthema's maatschappelijk verankerd en democratisch bespreekbaar.

2.2 Focus en rol in het overheidslandschap

WAAR HET INSTITUUT NAAR KIJKT

Het instituut richt zich op de krachtigste AI-systemen van nu en straks: zowel gesloten modellen van Amerikaanse en internationale labs als open modellen. Het onderzoekt technische veiligheidsrisico's, zoals biologische dreigingen en aanvallen op vitale infrastructuur, én maatschappelijke gevolgen, zoals de automatisering van de arbeidsmarkt. Om die risico's voor onze eigen samenleving te beoordelen, kan Nederland niet leunen op buitenlandse instituten of labs. Een eigen testcapaciteit is onmisbaar.

INFORMEREN, NIET HANDHAVEN

Het instituut is uitdrukkelijk geen toezichthouder of regelgever. Zijn rol is informerend en adviserend, richting kabinet, Kamer en publiek. Het onderzoekt wat modellen daadwerkelijk kunnen en wat dat voor Nederland betekent; of en hoe daarop beleidsmatig of wettelijk wordt ingegrepen, blijft een besluit van de politiek en de toezichthouders.

IN NAVOLGING VAN HET VK EN DUITSLAND

Nederland vindt hier niets uit. Het Verenigd Koninkrijk richtte in november 2023 als eerste land een AI Safety Institute op, en omschreef de opdracht in één zin.

> "*minimise surprise to the UK and humanity from rapid and unexpected advances in AI. AI Safety Institute, "Introducing the AI Safety Institute", november 2023, sectie Mission and scope*[13]"

Duitsland en Frankrijk volgden en inmiddels ligt er een netwerk van zusterinstituten met gedeelde methoden en open gepubliceerd onderzoek. Ook Noorwegen bereidt een eigen instituut voor, en Zweden overweegt hetzelfde. Nederland sluit daarop aan en hoeft niet opnieuw te beginnen. We brengen dan in dat netwerk iets in wat anderen niet hebben: de toets op onze eigen vitale systemen. Ook kunnen we beter gebruikmaken van de expertise van zusterinstituten en die toepassen om onze specifieke AI-risico's aan te pakken.

Wie	Rol	Wat het instituut voor hen betekent
NL-AIVI	toetst en adviseert	Klein en wendbaar, technische toppers, een korte lijn naar het kabinet.
Kabinet en Kamer	besluit	Krijgen de evaluaties en het advies van het instituut. Of en hoe wordt ingegrepen, blijft hun besluit.
Wetenschap, markt en publiek	begrijpt	Universiteiten, bedrijven en burgers krijgen begrijpelijke, relevante informatie.
Internationale partners	onderzoekt mee	UK AISI, Duitsland, Frankrijk en de EU: gezamenlijk fundamenteel onderzoek en gedeelde standaarden.
Vitale sectoren en defensie	gebruikt	Netbeheerders, waterschappen, hightech en krijgsmacht krijgen waarschuwingen en open hulpmiddelen.
toezicht en handhaving		
Toezichthouders	handhaaft	De toezichthouders krijgen de technische bevindingen. Toezicht en sancties blijven bij hen.

Het instituut is **geen toezichthouder, geen regelgever en geen vergunningverlener.** Het levert de meting waarop de anderen kunnen besluiten.

Waarom een los instituut?

Nederland heeft al sterke organisaties die over AI gaan. De RDI en de AP houden toezicht, het NCSC helpt bij cyberincidenten en het Europese AI-bureau ziet toe op de grootste modellen. Het instituut bouwt voort op hun werk en deelt de bevindingen met hen. Dat het een eigen instituut moet zijn, heeft vier redenen.

1. **Testen en toezicht werken het best als aparte rollen** De toezichthouder controleert of regels worden nageleefd, het instituut kijkt vooruit naar wat nieuwe modellen kunnen. Ook het Britse instituut is bewust geen toezichthouder. Het AI-lab van de RDI wordt wel een natuurlijke partner.
2. **Een korte lijn naar het kabinet** Vindt een model een lek in veelgebruikte software, dan moeten bewindspersonen dat binnen uren of dagen weten. Een eigen instituut kan die kennis snel en breed delen.
3. **Europa kijkt naar de EU, het instituut naar Nederland** Het Europese AI-bureau ziet toe op de grootste modellen in de hele Unie. Of een model de software van onze waterkeringen, ons energienet of de chipketen kan kraken, moet iemand hier uitzoeken. Die kennis brengt Nederland vervolgens in Europa in.
4. **Eén plek voor schaars talent** Wie modellen kan testen en aanvallen, werkt nu vaak bij de labs zelf. Een eigen instituut kan die mensen eigen arbeidsvoorwaarden en salarissen bieden, een eigen onderzoeksagenda en een vaste plek dicht bij de besluitvorming. En het kan nu klein beginnen en meegroeien met een vakgebied dat snel verandert.

WAT HET INSTITUUT ANDERS MAAKT

1. **Het informeert onafhankelijk** Het instituut levert de politiek, de markt en het publiek de best beschikbare technische kennis, in een vorm die ze direct kunnen gebruiken. Het oordeel is onafhankelijk: de uitkomst hangt niet af van de politieke lijn van het moment.
2. **Het kijkt breed** Over ethiek en privacy waken al anderen, zoals de AP. Het instituut onderzoekt wat modellen technisch en concreet kunnen: aanvallen uitvoeren, autonoom handelen, zwakke plekken vinden in software waar de economie en infrastructuur op draaien. Het kijkt ook verder dan veiligheid en volgt wat AI doet met werk, informatie en taal, en wat dat betekent voor hoe de samenleving functioneert.
3. **Het werkt vooruit en snel** Het instituut wacht niet op klachten, meldingen of incidenten. Het test modellen zo vroeg mogelijk, voordat ze breed gebruikt worden, want dan is bijsturen nog mogelijk. Het zoekt ook naar de oorzaak van een kwetsbaarheid, niet alleen naar de kwetsbaarheid zelf, en bedenkt wat ertegen helpt. Omdat modellen tegenwoordig snel beter worden, beoordeelt het snel.
4. **Het maakt innovatie mogelijk en doet het zelf** Het instituut wil AI-adoptie niet tegenhouden. Het wil dat die verantwoord kan en helpt bedrijven, onderzoekers en overheden daarbij. Op zijn eigen terrein loopt het ook voorop: het ontwikkelt nieuwe methoden en standaarden om modellen te evalueren en te beveiligen. Zo draagt het, samen met universiteiten en kennisinstellingen, bij aan het Nederlandse onderzoek naar AI-veiligheid en -impact. Dat vraagt een bestuur dat zowel het toetsen als het ontwikkelen van nieuwe standaarden kan leiden.
5. **Het werkt samen met de makers** Om de sterkste modellen zo vroeg mogelijk te toetsen, werkt het instituut samen met de labs die ze maken. Die samenwerking werkt twee kanten op: het instituut deelt bevindingen waarmee de labs hun modellen veiliger kunnen maken. Het oordeel blijft van het instituut zelf, en toezicht en sancties blijven bij de toezichthouders.
6. **Het bestaat uit technische experts** Het instituut draait op onderzoekers en technici die zelf modellen testen, aanvallen nabootsen en oplossingen bouwen. Voor beleid, wetgeving en juridische toetsing kan het leunen op de ministeries en toezichthouders die daar al sterk in zijn. Zelf richt het zich vooral op de technische vraag hoe veilig nieuwe modellen zijn.
7. **Het werkt in het openbaar** Het instituut publiceert zijn methoden en bevindingen zodra dat veilig kan, en legt ze uit aan een breed publiek.

Wat AI nu kan en straks gaat kunnen, vraagt erom dat Nederland eigen kennis, mensen en testcapaciteit opbouwt om die ontwikkeling zelf te volgen en te beoordelen. Dat lukt alleen met een kritieke massa op één plek. Daarom dit nieuwe instituut.

Nederland kan nu beginnen

Duitsland opende zijn instituut eind augustus, het Verenigd Koninkrijk begon al in 2023. Waar de formele oprichting tijd kost, kan een kwartiermaker hier nu beginnen met het voorwerk: de kern werven en de eerste afspraken maken met de grote, vooraanstaande labs.

Zonder eigen instituut blijft Nederland afhankelijk van de welwillendheid en de geopolitieke agenda van andere landen.

Door nu dit instituut op te richten voorkomt het kabinet dat de overheid achter de feiten aanloopt, geeft het wie onze economie en vitale infrastructuur beschermt de kennis die daarvoor nodig is, en maakt het Nederland een sterke partner in het Europese netwerk van AI-veiligheidsinstituten.

Bronnen

1. Anthropic, Claude Mythos Preview en Project Glasswing (https://red.anthropic.com/2026/mythos-preview), april 2026; voortgangsupdate (https://www.helpnetsecurity.com/2026/05/26/anthropic-project-glasswing-update/), mei 2026.
2. OpenAI, The Hugging Face incident and the road ahead (https://openai.com/index/hugging-face-incident-and-the-road-ahead/), augustus 2026; achtergrond in Fortune (https://fortune.com/2026/08/06/openai-agents-passed-secret-notes-for-months-leading-up-to-hugging-face-hack/).
3. Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign (https://anthropic.com/news/disrupting-AI-espionage), november 2025.
4. International AI Safety Report (https://www.gov.uk/government/publications/international-ai-safety-report-2025), eerste editie, januari 2025.
5. METR, Measuring AI Ability to Complete Long Tasks (https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/), maart 2025.
6. Verklaring over transformatieve AI (https://transformative-ai.eu/) van circa 200 economen, waaronder 16 Nobelprijswinnaars, juli 2026.
7. NCTV, Cybersecuritybeeld Nederland 2025 (https://www.nctv.nl/actueel/nieuws/2025/11/26/cybersecuritybeeld-2025-riskante-mix-in-een-onvoorspelbare-wereld); Algemene Rekenkamer, Focus op AI bij de rijksoverheid (https://www.rekenkamer.nl/documenten/2024/10/16/focus-op-ai-bij-de-rijksoverheid), oktober 2024.
8. AIVD, MIVD en NCTV, Versterkte dreigingen in een wereld vol kunstmatige intelligentie. Een analyse van het effect van AI op de nationale veiligheid (https://www.nctv.nl/documenten/publicaties/2024/12/09/versterkte-dreigingen-in-een-wereld-vol-kunstmatige-intelligentie), 9 december 2024.
9. NOS, Dringende oproep door hacken met AI: “We moeten supersnel kunnen reageren” (https://nos.nl/artikel/2615399-dringende-oproep-door-hacken-met-ai-we-moeten-supersnel-kunnen-reageren), 22 mei 2026, met NCSC-directeur Matthijs van Amelsfort.
10. RaboResearch, Dalende werkgelegenheid onder Nederlandse jongeren die concurreren met GenAI (https://esb.nu/dalende-werkgelegenheid-onder-nederlandse-jongeren-die-concurreren-met-genai/), Economisch Statistische Berichten, 22 januari 2026. Zie in hetzelfde tijdschrift ook de kritische repliek (https://esb.nu/niet-zeker-dat-werkgelegenheid-onder-nederlandse-jongeren-daalt-door-genai/).

11. Fortune, Deloitte job title change: AI reshapes Big 4 accounting and consulting firms (https://fortune.com/2026/01/22/deloitte-job-title-change-ai-reshapes-big-4-accounting-consulting-firms), januari 2026; Accountancy Age, The Big Four's new favourite grad is AI (https://accountancyage.com/2025/06/23/the-big-fours-new-favourite-grad-is-ai/), juni 2025.
12. Financial Ombudsman Service, Embracing AI's transformational impact on consumer complaints (https://www.financial-ombudsman.org.uk/businesses/resolving-complaint/our-insight/embracing-ais-transformational-impact-consumer-complaints), Verenigd Koninkrijk.
13. AI Safety Institute, Introducing the AI Safety Institute (https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf), Verenigd Koninkrijk, november 2023.
14. BBC News, Rogue OpenAI agent 'infiltrated' Australian government website in world first (https://www.bbc.com/news/articles/c6vgy0333dppo), 24 september 2026.
15. Grafiek: nieuw geregistreerde CVE's per maand naar CVSS-ernst (Critical ≥ 9,0; High 7,0–8,9), januari 2022 tot juli 2026. Bewerking auteur.

Contact

Oscar Lepoeter

oscar@deltainstituut.nl
+31 6 20553645

Delta Instituut

