Tweede Kamer der Staten-Generaal

2

Vergaderjaar 2026–2027

Aanhangsel van de Handelingen

Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden

254

Vragen van de leden **El Boujdaini** (D66) en **Zwinkels** (CDA) aan de Staatssecretaris van Economische Zaken en Klimaat over *AI-modellen die buiten hun testomgeving inbreken bij organisaties* (ingezonden 19 augustus 2026).

Antwoord van Staatssecretaris **Aerdts** (Economische Zaken en Klimaat) (ontvangen 5 oktober 2026). Zie ook Aanhangsel Handelingen, vergaderjaar 2025–2026, nr. 3038.

Vraag 1
Bent u bekend met berichtgeving over AI-agents die erin slaagden om buiten hun afgeschermde testomgeving («sandbox») te opereren en in het bijzonder met de bevindingen van AI-ontwikkelaar Anthropic, die meldde dat drie van haar modellen tijdens veiligheidstests door een configuratiefout internettoegang kregen en vervolgens op eigen initiatief inbraken bij de systemen van drie organisaties?[1] [2]

Antwoord 1
Ja.

Vraag 2
Is bij u bekend of zich onder de organisaties waarbij modellen ongeautoriseerd toegang kregen ook Nederlandse organisaties bevinden, en zijn zij hierover geïnformeerd?

Antwoord 2
Dit is ons niet bekend.
Hierbij moet opgemerkt worden dat het niet eenduidig te herkennen valt of een cyberaanval door een AI agent, een mens, of een combinatie daarvan is uitgevoerd. In de genoemde incidenten hebben AI-ontwikkelaars zelf publiekelijk gemeld dat AI agents een inbraak hadden uitgevoerd.

[1] NRC d.d. 31 juli 2026; Ook Claude rekent zich een weg uit de testomgeving. «Als een mens dit had gedaan zou die ontslagen worden»
[2] https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals d.d. 30 juli 2026

ah-tk-20262027-254
ISSN 0921 - 7398
's-Gravenhage 2026

Vraag 3
Deelt u de mening dat dit soort incidenten aantonen dat de risico's van steeds zelfstandiger opererende AI-modellen niet alleen liggen in de capaciteiten van het model zelf, maar ook in de mate waarin testomgevingen adequaat zijn beveiligd en afgesloten door de ontwikkelaars en testpartners?

Antwoord 3
Deze vraag is niet in algemene zin te beantwoorden. Zowel de karakteristieken en capaciteiten van AI modellen, als de eigenschappen van testomgevingen hebben invloed op het risico van een ongewenste uitbraak van AI agents. Ook de instructies die ontwikkelaars aan AI agents geven hebben invloed, evenals de kenmerken van de infrastructuur waarop AI producten ontwikkeld worden. De genoemde incidenten met uitbraken van AI agents zijn zorgwekkend, maar vragen om nadere analyse en interpretatie voordat er algemene conclusies over risico's afgegeven kunnen worden.

Vraag 4
In hoeverre bestaan er op dit moment (inter)nationale normen, standaarden of verplichtingen voor bedrijven die geavanceerde AI-modellen ontwikkelen en testen, ten aanzien van het veilig inrichten en afsluiten van testomgevingen? Als die er niet zijn, of vrijblijvend zijn, bent u bereid zich daarvoor in te zetten?

Antwoord 4
De AI-verordening[3] reguleert AI-systemen in Europa, en stelt eisen op basis van risico-categorisering. Zo worden bepaalde AI-praktijken verboden en worden er extra producteisen gesteld aan AI-systemen met een hoog risico. Tevens worden specifieke eisen gesteld aan zogenaamde *General Purpose AI Modellen (GPAI)*. Dit zijn grote modellen die verwerkt kunnen worden in heel veel verschillende systemen, zoals de modellen achter ChatGPT, Claude en Gemini. De eisen zijn per 2 augustus 2025 van kracht en de AI Office van de Europese Commissie (hierna Commissie) houdt hier toezicht op.
Aan alle GPAI-modellen worden documentatie eisen gesteld waaronder bepaalde informatie over training en testen. Daarnaast bestaat een specifieke categorie van GPAI modellen met systeemrisico's waarvoor aanvullende eisen gelden. Dit zijn de modellen die met een bepaalde grote hoeveelheid rekenkracht zijn getraind en vergaande capaciteiten hebben. Wanneer de Commissie het besluit heeft genomen dat een dergelijk model onder de categorie van systeem risico valt, moeten de aanbieders van de AI-modellen onder andere mitigerende maatregelen nemen voor de systeemrisico's.
Zo moeten aanbieders van AI modellen met systeemrisico's zorgen voor een passend niveau van cyberbeveiliging voor het model en de fysieke infrastructuur ervan gedurende de gehele levenscyclus van het model. Hierbij moeten aanbieders van deze categorie AI modellen rekening houden met onopzettelijke modellekken, ongeoorloofde vrijgave, omzeiling van veiligheidsmaatregelen, alsmede met bescherming tegen cyberaanvallen, ongeoorloofde toegang, en modeldiefstal. Die bescherming kan worden geboden door modelwegingen, algoritmen, servers en datasets te beveiligen, bijvoorbeeld met operationele beveiligingsmaatregelen voor informatie, specifiek cyberbeveiligingsbeleid, passende technische en bewezen oplossingen, alsmede met controles op cyber- en fysieke toegang, afgestemd op de relevante omstandigheden en risico's. Het inrichten van veilige ontwikkel- en testomgevingen valt onder de verantwoordelijkheid van deze aanbieders.
Hierbij merkt het kabinet op dat test- en ontwikkelactiviteiten zijn uitgezonderd omdat de AI-verordening van toepassing zijn op AI-modellen die op de markt wordt gebracht of in gebruik worden genomen. In de genoemde voorbeelden gaat het, voor zover bekend, om ontwikkelomgevingen die binnen de genoemde bedrijven ingericht zijn, en niet als product op de markt gebracht zijn. De AI Office bestudeert deze gevallen en de beoogde coördinerende markttoezichtautoriteiten in Nederland, de RDI en de AP, zijn met hen in contact.
Nederland prefereert op gebied van AI de internationale samenwerking via bestaande mandaten en structuren in plaats van het opbouwen van nieuwe

[3] https //digital-strategy.ec.Europa.eu/en/policies/regulatory-framework-ai

afspraken. Naast de AI-verordening zijn er verschillende internationale kaders waarin er normen en principes zijn vastgelegd over de ontwikkeling en het gebruik van AI, zoals de AI Principles van OESO[4], de Recommendation on the Ethics of AI[5] die door alle UNESCO lidstaten is aangenomen, en het Hiroshima AI reporting framework[6], een vrijwillige internationaal instrument dat organisaties helpt om te rapporteren over hun risicobeheer bij het inzetten en ontwikkelen van frontier AI modellen. Nederland moedigt landen aan om zich aan te sluiten bij de *Framework Convention*[7] van de Raad van Europa, een bindend verdrag over artificiële intelligentie, mensenrechten, democratie en de rechtsstaat. Via deze, en andere internationale kaders, kan verdere aandacht geschonken worden aan het bovenstaande vraagstuk.

Vraag 5
Bent u bekend met benchmarks, zoals SandboxEscapeBench van het Britse AI Security Institute, ontwikkeld om systematisch te testen of AI-modellen in staat zijn om uit hun testomgeving te ontsnappen, en ziet u een rol voor de overheid of Europese instanties om dit soort onafhankelijke toetsing te stimuleren of te verplichten voor in Nederland of de EU actieve AI-bedrijven?[8]

Antwoord 5
Het kabinet is niet specifiek op de hoogte van de genoemde benchmark. Het AI Office van de Europese Commissie ziet toe op de onder vraag 4 genoemde wettelijke verplichtingen. Het kabinet hecht er waarde aan dat het AI Office daar goed toe in staat is en daarin goed samenwerkt met de toezichthouders in de lidstaten.

Vraag 6
Deelt u de mening dat de verantwoordelijkheid voor dit soort incidenten primair bij de ontwikkelaars van AI-modellen ligt, en niet louter kan worden afgedaan als een onvermijdelijk gevolg van toenemende AI-capaciteiten?

Antwoord 6
Het Kabinet deelt deze mening. Ontwikkelaars zijn verantwoordelijk voor de AI-modellen en -systemen die zij creëren en het mitigeren van mogelijke risico's. Deze verantwoordelijkheid geldt als producten op de markt zijn gekomen, maar ook voor de ontwikkel- en testfase. De Commissie heeft samen met ontwikkelaars en toezichthouders, en met ondersteuning vanuit ngo's, praktijkcodes opgesteld die een uitwerking zijn van de wettelijke eisen. In deze General-Purpose AI Code of Practice[9] staan richtlijnen die ontwikkelaars helpen om aan de eisen te kunnen voldoen.

Vraag 7
Welke rol ziet u hierbij voor de Europese AI Act en de daarin opgenomen eisen aan risicobeheersing voor AI-modellen voor algemene doeleinden met een systeemrisico?

Antwoord 7
Zie de beantwoording van vraag 4.

Vraag 8
Welke risico's ziet u voor de nationale veiligheid en digitale infrastructuur wanneer AI-modellen die getest worden door bedrijven wereldwijd, in staat blijken om buiten hun testomgeving te opereren, bijvoorbeeld door ongeautoriseerde toegang tot netwerken of systemen te verkrijgen?

---

[4] https //www.oecd.org/en/topics/ai-principles.html
[5] https //www.unesco.org/en/artificial-intelligence/recommendation-ethics
[6] https //oecd.ai/en/transparency/overview
[7] https //www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence
[8] https://www.aisi.gov.uk/blog/can-ai-agents-escape-their-sandboxes-a-benchmark-for-safely-measuring-container-breakout-capabilities d.d. 23 maart 2026
[9] https //digital-strategy.ec.Europa.eu/en/policies/contents-code-gpai

Antwoord 8
Een AI-model dat ongeautoriseerd toegang tot netwerken of systemen kan krijgen, zou kwetsbaarheden sneller en op grotere schaal kunnen vinden en misbruiken dan voorheen. AI verlaagt de drempel voor aanvallers en maakt het voor een bredere groep kwaadwillenden mogelijk om cyberaanvallen uit te voeren. Daardoor ontstaat er een nieuwe realiteit waar we ons aan moeten aanpassen.
We zien een duidelijke trend in incidentrapportages en onderzoeken dat kwaadwillenden steeds beter weten hoe AI technologie ingezet kan worden. De technologie ontwikkelt zich snel en de mogelijkheden voor kwaadwillenden nemen toe. Er is niet specifiek één sector aan te wijzen die het meest kwetsbaar is. Aan de incidenten van de afgelopen tijd is te zien dat elke sector getroffen kan worden. Van overheidsorganisatie, tot zorginstelling tot telecombedrijf. Verstoring of uitval hiervan, kan maatschappelijke ontwrichting veroorzaken en de nationale veiligheid raken. In elke sector is het op orde brengen van de basis dan ook randvoorwaardelijk. Daarom is op 24 september een gezamenlijke verklaring van de AIVD, MIVD, NCTV, NCSC, politie, OM en CIO-rijk uitgebracht[10], bedoeld om overheden, bedrijven en organisaties te activeren om extra stappen te zetten. In november dit jaar wordt uw Kamer door de Minister van Justitie en Veiligheid middels de Kamerbrief voortgangsrapportage Nederlandse Cybersecuritystrategie en Cybersecurity Beeld Nederland 2026 geïnformeerd over de stappen die het Kabinet onderneemt om de digitale weerbaarheid tegen cyberdreigingen, waaronder die vanuit geavanceerde AI-systemen, te verhogen.

Vraag 9
Bent u bereid om in Europees verband te bepleiten dat ontwikkelaars van geavanceerde AI-modellen verplicht worden incidenten waarbij modellen uit testomgevingen ontsnappen te melden bij een toezichthouder en getroffen organisaties, vergelijkbaar met meldplichten bij datalekken?

Antwoord 9
De AI-verordening bevat al een verplichting voor aanbieders van AI-modellen voor algemene doeleinden met systeemrisico's om ernstige incidenten te melden bij de Europese Commissie en de nationale autoriteit. Aanbieders van deze categorie AI-modellen dienen mogelijke systeemrisico's te beoordelen en te beperken. Indien er zich als gevolg van de ontwikkeling of het gebruik van deze AI-modellen een ernstig incident voordoet, moet de aanbieder van het AI-model onverwijld het verdere verloop van het incident beginnen te volgen en de Commissie en de nationale bevoegde autoriteiten alle relevante informatie en mogelijke corrigerende maatregelen melden.

Vraag 10
Welke maatregelen treft u, of gaat u treffen, om te borgen dat testomgevingen van AI-bedrijven die in Nederland actief zijn of hier data verwerken, aan adequate veiligheidseisen voldoen?

Antwoord 10
Zoals gemeld in het antwoord op vraag 4 en 5 houdt het AI Office van de Europese Commissie toezicht en hecht het kabinet eraan dat zij hier goed toe in staat is en samenwerkt met de nationale toezichthouders.

Vraag 11
Kunt u deze vragen afzonderlijk van elkaar beantwoorden?

Antwoord 11
Dit is geschied.

[10] https //www.nctv.nl/actueel/nieuws/2026/09/24/ai-versnelt-en-vergroot-de-dreiging-nu-handelen-noodzakelijk