Tweede Kamer der Staten-Generaal

2

Vergaderjaar 2025–2026

Vragen gesteld door de leden der Kamer

2026Z17015

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).

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]

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?

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?

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?

[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

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?[3]

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?

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?

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?

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?

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?

Vraag 11
Kunt u deze vragen afzonderlijk van elkaar beantwoorden?

[3] 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