AH 455
2026Z18543

Antwoord van staatssecretaris Aerdts (Economische Zaken en Klimaat), mede namens de minister van Justitie en Veiligheid (ontvangen 26 oktober 2026)

Zie ook Aanhangsel Handelingen, vergaderjaar 2026-2027, nr. 190

Vraag 1
Bent u bekend met het bericht ‘Anthropic-onderzoeker neemt ontslag: OpenAI en Anthropic handelen onverantwoordelijk en gokken met ons leven’[1]en het bericht ‘Anthropic-medewerker: kans bestaat dat AI de mensheid uitroeit’[2], waarin onderzoekers van vooraanstaande AI-bedrijven waarschuwen voor ernstige veiligheidsrisico’s van steeds krachtigere AI-systemen?

Antwoord

Ja.

Vraag 2

Hoe beoordeelt u de waarschuwingen van deze onderzoekers dat de ontwikkeling van zeer geavanceerde AI-systemen sneller verloopt dan de ontwikkeling van adequate veiligheidsmaatregelen en beheersingsmechanismen? Welke conclusies verbindt u daaraan?

Antwoord

Het kabinet is bekend met de recente berichten en de bredere zorgen van onderzoekers bij geavanceerde-AI-bedrijven over veiligheidsrisico’s en het risico dat de ontwikkeling van geavanceerde (*frontier*) AI-modellen sneller gaat dan de ontwikkeling van adequate veiligheids- en beheersingsmechanismen en deelt de zorgen van de onderzoekers hierover. De waarschuwing van de onderzoekers volgde op het nieuws dat AI-agents van OpenAI erin slaagden om internettoegang te verkrijgen en een hack uit te voeren bij het codeplatform Hugging Face. AI-systemen die ongeautoriseerd toegang tot netwerken of systemen kunnen krijgen, zorgen ervoor dat kwetsbaarheden sneller en op grotere schaal gevonden en misbruikt kunnen worden. Zorgen hierover sluiten aan bij de analyse van het VN wetenschappelijk panel over AI, dat wijst op een groeiende “*governance gap*” tussen snel toenemende grensoverschrijdende risico’s en de huidige instituties, normen en toezichtcapaciteit.

[1] Anthropic-onderzoeker neemt ontslag: ‘OpenAI en Anthropic handelen onverantwoordelijk en gokken met ons leven’, Volkskrant d.d. 9 september 2026; Anthropic-onderzoeker neemt ontslag: ‘OpenAI en Anthropic handelen onverantwoordelijk en gokken met ons leven’ | de Volkskrant.
[2] Anthropic medewerker: kans bestaat dat AI de mensheid uitroeit’, Quotenet d.d. 9 september 2026; Anthropic-medewerker: 'Kans bestaat dat AI de mensheid uitroeit'.

Het kabinet benadrukt dat ontwikkelaars van de AI-modellen verantwoordelijk zijn voor het identificeren en mitigeren van mogelijke risico's via veiligheidsmaatregelen en beheersingsmechanismen. Deze verantwoordelijkheid geldt zowel voor wanneer producten op de markt zijn gekomen als voor de ontwikkel- en testfase die daaraan vooraf gaat. Tegelijkertijd wordt door de technologische ontwikkelingen van AI de drempel voor kwaadwillenden verlaagd om cyberaanvallen uit te voeren, en vergroten zij de snelheid, schaal en efficiëntie van cyberaanvallen. Het kabinet ziet een duidelijke trend in incidentrapportages en onderzoeken dat kwaadwillenden steeds beter weten hoe AI technologie ingezet kan worden.

Het kabinet zet in het licht van deze ontwikkelingen onder andere in op het verhogen van de digitale weerbaarheid van Nederland, een consequente en effectieve handhaving van de toepasselijke regels in de Europese Unie en het versterken van het internationale normatieve kader.

Vraag 3

Heeft het kabinet scenarioanalyses laten uitvoeren naar de gevolgen van geavanceerde AI voor de nationale veiligheid, waaronder cyberaanvallen, beïnvloeding van democratische processen, sabotage van vitale infrastructuur, militaire toepassingen en andere hybride dreigingen? Zo ja, kunt u de belangrijkste bevindingen met de Kamer delen?

Antwoord

Gezien de snelheid waarop de ontwikkelingen van AI-modellen plaatsvinden, wordt dit door verschillende organisaties, waaronder de AIVD, MIVD, NCTV en het NCSC constant gemonitord en verwerkt in hun analyses. In het Cyber Security Beeld Nederland wordt de impact van cybersecurityontwikkelingen op onze nationale veiligheid geschetst. Zo werd in het afgelopen CSBN 2025 geconcludeerd dat generatieve AI de bestaande dreigingen voor de digitale veiligheid versterkt.[3] In november dit jaar wordt uw Kamer over het CSBN 2026 geïnformeerd. Hierin worden recente ontwikkelingen op het gebied van generatieve AI meegenomen. Daarnaast werd de inschatting van de dreiging van AI eerder geschetst in de in de analyse "Versterkte dreigingen in een wereld vol kunstmatige intelligentie" van AIVD, MIVD en NCTV uit 2024.[4] Deze analyse gaat over de inschatting of AI zelfstandig controle over mensen zou kunnen krijgen. De ontwikkeling van AI-modellen gaat snel en gaan zelfs sneller dan in 2024 was verwacht. Deze snelle ontwikkeling en de consequenties die dit mogelijk kan hebben voor de digitale veiligheid van Nederland was onder andere aanleiding om voor de AIVD, MIVD, NCTV, NCSC, politie, OM en CIO-rijk een gezamenlijke verklaring naar buiten te brengen.[5] Deze verklaring is bedoeld om overheden, bedrijven en organisaties te activeren om nu extra stappen te zetten in het

[3] Cybersecuritybeeld 2025: Riskante mix in een onvoorspelbare wereld | Nationaal Coördinator Terrorismebestrijding en Veiligheid
[4] AI versterkt bestaande dreigingen voor nationale veiligheid | Nationaal Coördinator Terrorismebestrijding en Veiligheid
[5] https //www.nctv.nl/actueel/nieuws/2026/09/24/ai-versnelt-en-vergroot-de-dreiging-nu-handelen-noodzakelijk

verhogen van hun digitale weerbaarheid en bestuurders op te roepen om hun verantwoordelijkheid te nemen en prioriteit te geven aan cybersecurity.

Verder zet het kabinet zich met de Nederlandse Cybersecuritystrategie 2022 – 2028 in op het verhogen van de digitale weerbaarheid van Nederland. De digitale weerbaarheidsrisico's die voortkomen uit de ontwikkeling van frontier AI-modellen worden hierin meegenomen. Maar er is meer nodig: het dreigingslandschap ontwikkelt zich snel en de digitale weerbaarheid van Nederlandse organisaties en bedrijven groeit niet in hetzelfde tempo mee. Het kabinet werkt daarom aan een actieprogramma als versterking op de NLCS om de digitale weerbaarheid tegen cyberdreigingen, waaronder die vanuit geavanceerde AI-systemen, te verhogen. Hiermee gaat het kabinet aan de slag met het verhogen van de digitale weerbaarheid van nú, en het bouwen aan een stevig fundament voor de toekomst. Over de inhoud van dit actieprogramma wordt uw Kamer door de Minister van Justitie en Veiligheid in november 2026, gelijktijdig met de Voortgangsrapportage Nederlandse Cybersecuritystrategie 2026 en Cybersecurity Beeld Nederland 2026, geïnformeerd.

Vraag 4

In hoeverre acht u het bestaande nationale, Europese en internationale normatieve kader, waaronder de AI Act, toereikend voor het beheersen van risico's die voortvloeien uit de ontwikkeling van zeer krachtige algemene AI-modellen en toekomstige frontier AI-systemen?

Antwoord

Er bestaan verschillende internationale kaders waarin er normen en principes zijn vastgelegd over de verantwoorde ontwikkeling en gebruik van AI. De Organisatie voor Economische Samenwerking en Ontwikkeling (OESO) heeft reeds in 2019 *AI principles* vastgesteld, die op hoofdlijnen door de G20 zijn overgenomen en als basis hebben gediend voor verschillende AI-wetten wereldwijd. Andere relevante internationale kaders zijn de *Recommendation on the Ethics of AI* die door alle UNESCO lidstaten is aangenomen en het Kaderverdrag van de Raad van Europa over AI en de mensenrechten, de democratie en de rechtsstaat. Internationaal worden er ook specifieke afspraken gemaakt over geavanceerde AI-systemen, zoals het vrijwillige *Hiroshima AI reporting framework*.

Deze principes, kaders en aanbevelingen vormen samen een internationaal normatief kader waarin het belang wordt onderstreept van principes als het respecteren van de rechtsstaat en mensenrechten, transparantie, veiligheid en aansprakelijkheid. Landen zijn verantwoordelijk voor het implementeren van dit internationaal normatief kader in de eigen rechtsorde.

In de Europese Unie zijn deze normen en afspraken geïmplementeerd via onder andere de AI-verordening. Deze wet stelt eisen aan verschillende categorieën AI-producten en toepassingen, afhankelijk van de risico’s die bij elke categorie komen kijken. Bepaalde AI-praktijken zijn verboden, zoals manipulerende AI of AI die misbruik maakt van kwetsbaarheden van mensen. Bepaalde AI-toepassingen en producten worden geclassificeerd als hoog-risico, zoals bepaalde AI-

toepassingen in het onderwijs, rechtshandhaving en medische apparatuur. Hoog-risico AI moet aan zware producteisen voldoen om ervoor te zorgen dat er geen negatieve impact wordt gemaakt op de gezondheid, veiligheid en mensenrechten. Een specifieke groep AI-systemen waar mensen mee interacteren zoals chatbots en generatieve AI moeten aan transparantie-eisen voldoen zodat mensen weten wanneer ze met AI of AI-gegenereerde content te maken hebben. Het toezicht op deze categorieën is op nationaal niveau.

Ten slotte worden er ook nog eisen gesteld aan *General Purpose AI Modellen (GPAI-modellen)*. Dit zijn grote modellen die verwerkt kunnen worden in heel veel verschillende systemen, zoals de modellen achter ChatGPT, Claude en Gemini. Binnen deze groep modellen bestaat de categorie GPAI modellen met systeemrisico's. Dit zijn modellen die, op basis van benchmark analyse of expert beoordelingen, een vermeend hoog impact gehalte hebben. Voor deze GPAI-modellen met systeemrisico's gelden aanvullende risicomanagement eisen. Dit zijn de *frontier* modellen die met een grote hoeveelheid rekenkracht zijn getraind en vergaande capaciteiten hebben ten opzichte van oudere AI-modellen. Voor de risico's die uit deze nieuwe capaciteiten voortkomen, moeten de makers van de AI-modellen mitigerende maatregelen nemen. 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. De eisen zijn per 2 augustus 2025 van kracht en de AI Office van de Europese Commissie houdt hier toezicht op.

De AI-verordening is het meest omvangrijk wettelijk kader voor het beheersen van risico's van frontier AI wereldwijd. Tegelijkertijd ontwikkelt de technologie zich snel en kan de AI Office pas handhaven op het moment dat een AI-model in de Europese Unie op de markt wordt gebracht. Het kabinet zet zich er daarom in Europa voor in dat de regels op een consequente en effectieve manier worden gehandhaafd, en dat er door de AI Office en de lidstaten continu wordt geëvalueerd of de verordening toereikend is om een passend niveau van bescherming te bieden.

Vraag 5

Welke aanvullende maatregelen acht u noodzakelijk om de digitale weerbaarheid van Nederland te versterken tegen risico's die samenhangen met geavanceerde

AI-systemen, in het bijzonder waar het gaat om vitale infrastructuur, overheidsorganisaties, defensie en de bescherming van burgers?

Antwoord

Graag verwijs ik u naar het antwoord op vraag 3.

Vraag 6

Hoe geeft het kabinet, mede in het licht van het streven naar een digitaal weerbare samenleving, invulling aan het voorzorgsbeginsel bij de ontwikkeling en toepassing van zeer geavanceerde AI-systemen waarvan de maatschappelijke gevolgen mogelijk verstrekkend en onomkeerbaar zijn?

Antwoord

Het kabinet geeft, via de AI-verordening, invulling aan het voorzorgsbeginsel door strenge eisen aan de ontwikkeling en toepassing van zeer geavanceerde AI-systemen te stellen. Zoals toegelicht in het antwoord op vraag 4, moeten de aanbieders van GPAI-modellen onder andere maatregelen nemen voor de systeemrisico’s die voortkomen uit de nieuwe capaciteiten van deze modellen. Het model mag niet op de Europese markt worden gebracht of in gebruik worden genomen als de aanbieder deze maatregelen niet of onvoldoende heeft getroffen. Deze eisen worden gehandhaafd door de Europese Commissie. Zij worden daarin ondersteund door een wetenschappelijk panel van onafhankelijke deskundigen die een waarschuwing aan de Europese Commissie kunnen verstrekken indien zij vermoedt dat een GPAI-model voor concrete en aanwijsbare risico’s zorgt.

Daarnaast geeft het kabinet invulling aan het voorzorgsbeginsel via actieve AI-diplomatie, zoals uitgewerkt in de Internationale AI-strategie. Nederland zet zich internationaal in voor versterking van normen en standaarden, voor de ontwikkeling en inzet van onafhankelijke evaluatie- en verificatiemechanismen en voor afspraken over verantwoord gebruik van AI, zodat grensoverschrijdende risico’s van zeer geavanceerde AI-systemen gezamenlijk en tijdig kunnen worden beheerst.

Vraag 7

Deelt u de zorg dat een internationale concurrentiestrijd tussen staten en marktpartijen kan leiden tot een situatie waarin veiligheidsbelangen ondergeschikt raken aan technologische en economische belangen? Zo ja, hoe wordt dit risico in het kabinetsbeleid geadresseerd?

Antwoord

De internationale concurrentiestrijd tussen staten en marktpartijen op het gebied van AI vindt op dit moment voornamelijk plaats tussen de Verenigde Staten en China. Het kabinet deelt de zorgen dat in deze concurrentiestrijd veiligheidsbelangen ondergeschikt kunnen raken aan technologische en

economische belangen. Het is daarom enerzijds van belang dat zo veel mogelijk landen zich committeren aan de bestaande internationale afspraken en anderzijds dat Nederland en de Europese Unie hun eigen AI-capaciteiten op onder andere het gebied van AI-infrastructuur en talent versterken. Door zelf in staat te blijven om te innoveren op het gebied van AI, wordt de mogelijkheid behouden om ontwikkelingen in een verantwoorde richting te sturen en keuzemogelijkheden in de toepassing van AI te behouden.

Daarom zet het kabinet ook buiten de Europese Unie in op samenwerking met gelijkgezinde landen om gezamenlijk het gedeelde normatieve kader verder te bestendigen en waar mogelijk samen te werken om elkaars AI-capaciteiten te versterken.

Daarnaast neemt ook op internationaal niveau de aandacht voor het veranderende cyberdreigingslandschap toe. De Europese Commissie heeft onlangs een Actieplan voor Cybersecurity & AI gepubliceerd, voortbouwend op onder andere de EU AI verordening. Dit actieplan heeft als doel het veilige en verantwoorde gebruik van AI te ondersteunen en tegelijkertijd de cyberbeveiliging van Europa te versterken. Het kabinet zet zich hier actief voor in.

Vraag 8

Herkent u de zorgen die leven rondom superintelligente AI en het risico dat de mens niet meer in staat is deze ontwikkeling te stoppen? Deelt u de mening dat een verbod hierop wenselijk is?

Antwoord

Het kabinet ziet momenteel geen wetenschappelijk bewijs dat superintelligente AI op dit moment ontwikkeld kan worden. Het kabinet vindt het van belang dat bewezen risico’s zoals discriminatie, geautomatiseerde cyberaanvallen, manipulatie of desinformatie met prioriteit geadresseerd worden. Zoals toegelicht in het antwoord op vraag 4, biedt de AI-verordening hier het wettelijk kader voor.

Tegelijkertijd ziet het kabinet ook dat een grote concentratie aan rekenkracht en kapitaal bij een paar commerciële partijen en snelle technologische ontwikkelingen een aantal van deze risico’s in een stroomversnelling brengt en in extreme gevallen voor maatschappij-ontwrichtende systeemrisico’s kan zorgen. Bijvoorbeeld wanneer AI-modellen door menselijke actoren worden ingezet om grootschalige cyberaanvallen uit te voeren. Naast het handhaven van de hiervoor geschetste wettelijke verplichtingen, acht het kabinet het van belang dat de digitale weerbaarheid in Nederland verhoogd wordt. Zoals beschreven in het antwoord op vraag 2 onderneemt het Kabinet hiervoor verschillende stappen.

Ook erkent het kabinet, in lijn met het voorlopige rapport van het VN-Panel over AI, het zogenoemde "*evidence dilemma*" waarin wordt gesteld dat beleidsmakers besluiten moeten nemen onder omstandigheden waarin de beschikbare wetenschappelijke kennis achterloopt bij de snelheid van AI-ontwikkelingen en waarin sommige niet op voorhand kunnen worden uitgesloten. Daarom anticipeert het kabinet op risico's die we nu nog niet volledig kunnen voorzien door de eigen kennis en capaciteit op AI te vergroten en door actieve deelname aan Europese en internationale netwerken en uitwisseling met onafhankelijke experts, zodat Nederland tijdig kan bijsturen en zo nodig aanvullende maatregelen, waaronder vergaande beperkingen, kan overwegen.

Vraag 9

Bent u bereid binnen de Europese Unie en internationale gremia te pleiten voor aanvullende afspraken over de ontwikkeling van geavanceerde AI, een verbod op superintelligente AI, en de mogelijkheid om tijdelijk een pas op de plaats te maken bij de ontwikkeling van krachtige AI-systemen zonder dat de risico's voldoende helder zijn?

Vraag 10

Kunt u de Kamer voor het einde van dit jaar informeren over de kabinetsinzet en de mogelijkheden en wenselijkheid van aanvullende waarborgen voor geavanceerde, superintelligente AI, waaronder de mogelijkheid van een tijdelijke ontwikkelpauze voor de krachtigste AI-systemen?

Antwoord vragen 9 en 10

Het kabinet is geen voorstander van eenzijdige pauze van AI-ontwikkelingen of van het nu vaststellen van één allesomvattend wereldwijd AI-verdrag, maar steunt wel de ontwikkeling van AI in lijn met de internationale normen op het gebied van verantwoorde AI. Het kabinet heeft daartoe op 22 september de *Call for Control of Frontier AI Models* ondertekend. Deze verklaring roept op dat AI aan menselijke sturing en controle onderhevig moet blijven en dat AI in overeenstemming met het internationale recht wordt ontwikkeld en gebruikt. Hiervoor wordt gepleit voor strengere veiligheidsprotocollen, onafhankelijke evaluatie en versterking van internationale verificatie- en normenstelsels.

Tegelijkertijd erkent het kabinet dat de snelle technologische ontwikkelingen en het "evidence dilemma" vragen om een open en adaptieve inzet. Aanvullende maatregelen voor geavanceerde AI-systemen kunnen in de toekomst aan de orde zijn, afhankelijk van de aard van de risico's en het internationale krachtenveld. Het kabinet sluit dergelijke opties daarom niet uit, maar beziet op basis van nieuwe ontwikkelingen en in overleg met gelijkgezinde partners wanneer, via welk forum en vanuit welke invalshoek het opportuun is daaraan bij te dragen.

In Europa blijft het kabinet zich ervoor inzetten dat de toepasselijke regels, waaronder de AI-verordening, op een consequente en effectieve manier worden gehandhaafd, en dat continu wordt geëvalueerd of de verordening toereikend is

om een passend niveau van bescherming te bieden op het gebied van frontier AI-modellen. Het kabinet vindt het van belang dat de koers van AI niet uitsluitend wordt bepaald door een paar grote mogendheden en bedrijven. Daarom zet Nederland zich binnen en buiten de Europese Unie in voor samenwerking met gelijkgezinde landen om het gedeelde normatieve kader verder te bestendigen en waar mogelijk samen te werken om elkaars AI-kennis, capaciteit en gezamenlijke weerbaarheid te versterken. Deze lijnen zijn onderdeel van de brede inzet op AI van het kabinet, die wordt beschreven in de kabinetsbrede AI-visie die dit najaar met uw Kamer zal worden gedeeld.