AI-veiligheid Instituut voor Nederland

Een nationaal instituut voor veilige AI gericht op economische groei, tijdige duiding van risico's en strategische onafhankelijkheid

Auteurs:
Bram Poppink
Bart Gijsen
Li-anne Tjin
Willem Jan de Voogd

Datum: 10 Juni 2026

Opdrachtgever:
Martijn Neef (Ministerie van Economische Zaken en Klimaat)

Rapport: TNO-2026-16238
Rubricering: TNO Publiek

TNO innovation for life

Management Samenvatting

Aanleiding
Nederland en Europa streven ernaar hun technologische en digitale concurrentiekracht te vergroten en strategische afhankelijkheden af te bouwen in een toenemend geopolitiek instabiele wereld. Artificiële intelligentie (AI) is hierin een sleuteltechnologie: enerzijds als motor voor economische groei en welvaart, anderzijds als bepalende factor in geopolitieke machtsverhoudingen, met directe impact op cyberveiligheid, defensie en hybride conflictvoering.

Parallel aan de AI (beleids-) ontwikkelingen is ook het beleid op de sleuteltechnologie Cybersecurity sterk in ontwikkeling. Dit gebeurt voornamelijk via regelgeving en stimuleringsinitiatieven om de veiligheid van producten en diensten te vergroten, en de digitale weerbaarheid van burgers, bedrijven en publieke organisaties te versterken.

Met de ontwikkelingen in zowel het AI als het Cybersecurity vakgebied ontstaat er steeds meer behoefte aan strategisch beleid op het raakvlak van beide velden. De geopolitieke dynamiek en de snelle AI ontwikkelingen verhogen de urgentie voor dit beleid, soeverein handelingsperspectief en de opbouw van duurzame innovatiekracht.

AI-veiligheid
Tegelijkertijd waarschuwen internationale experts dat AI als technologie aanzienlijke risico's met zich meebrengt voor maatschappij, veiligheid en stabiliteit. Met AI wordt mogelijkerwijs de drempel verlaagd voor de ontwikkeling van onder andere offensieve cyber capaciteiten, en de ontwikkeling van chemische, biologische, radiologische en nucleaire wapens. Tegenstanders in het geopolitieke domein kunnen AI gebruiken om democratische processen te verstoren door de publieke opinie te beïnvloeden. Verder zijn er zorgen over scenario's waarin menselijke controle over AI technologie afneemt.

Om de balans tussen de meerwaarde van AI en de voortvloeiende risico's effectief te beheersen krijgt AI-veiligheid onderzoek (inter)nationaal steeds meer aandacht. AI-veiligheid richt zich op onderzoek naar het beheersen van technische veiligheidsrisico's en –verbeteringen van AI, het monitoren van risico's en het adviseren van beleidsmakers.

De vertaling van AI-veiligheid onderzoek naar een concrete beleidsaanpak is echter complex en onderhevig aan het *evidence dilemma*: het ontwikkeltempo van (met name general-purpose) AI is sneller dan dat de risico's ervan kunnen worden ingeschat. Dit stelt beleidsmakers voor het dilemma om snel maatregelen te treffen die mogelijk niet effectief zijn, of pas maatregelen te treffen als het kwaad al geschied is. Om dit dilemma het hoofd te bieden ontstaat wereldwijd een trend: landen richten AI-veiligheid Instituten (AISI's) op die gericht zijn op een adaptieve beleidsaanpak. In plaats van het inrichten van tragere beleidsinstrumenten is deze aanpak gericht op een continue beheersing van verantwoorde en veilige ontwikkeling en gebruik van AI door een gezaghebbende samenwerking tussen private, publieke en kennisorganisaties. De vraag naar de geschiktheid van deze aanpak voor Nederland staat in dit onderzoek centraal.

Aanpak Cybersecurity en AI
Dit rapport is opgesteld in opdracht van het Ministerie van Economische Zaken en Klimaat, vanuit de ambitie om de beleidsdomeinen van AI en Cybersecurity in Nederland beter te verbinden, zodat publieke belangen rond veiligheid, soevereiniteit en economische groei beter geborgd kunnen worden. Hierbij staan de volgende vragen centraal:

- Welke kansen en dreigingen in het raakvlak van cybersecurity en AI zijn het meest relevant t.a.v. verdienvermogen en de publieke belangen?
- Hoe kunnen cybersecurity en AI ontwikkelingen en hun invloed op verdienvermogen en publieke belangen worden gemonitord?
- Welke beleidsmatige interventies zijn meest geschikt voor het balanceren van elk van die belangen?

Dit onderzoek is uitgevoerd in 4 onderzoeksporen: een scenarioanalyse, duiding AI-veiligheid Instituten (AISIs), een marktconsultatie en een inventarisatie van nationale AI-veiligheid initiatieven.

Onderzoekspoor 1: Scenarioanalyse
Over AI en Cybersecurity gerelateerde kansen en bedreigingen wordt steeds meer gepubliceerd. In het voorgaande, verkennende onderzoek van TNO zijn deze kansen en bedreigingen op hoofdlijnen geschetst en geïllustreerd met voorbeelden. Het International AI Safety Report geeft een beeld o.b.v. recente wetenschappelijke inzichten (met nadruk op de risico's). Ondanks dit actuele beeld blijft onduidelijk welke concrete invloed die risico's hebben op Nederlandse private en publieke organisaties en burgers, en welke concrete kansen AI biedt voor cybersecurity in Nederland. Daarmee ontbreekt het zicht op concrete Cybersecurity & AI beleidsprioriteiten.

Om een meer concreet fundament te geven voor het opstellen van een doelgerichte aanpak voor Cybersecurity en AI, zijn in dit onderzoek scenario's opgesteld en getoetst bij private en publieke organisaties.

De geleerde lessen uit de scenarioanalyse bevestigen het evidence dilemma voor beleidsmakers. AI-veiligheid is gebaat bij een gecoördineerde, sector-overstijgende aanpak en deze ontstaat niet vanzelf. Verder vergt de dynamiek van AI-veiligheid continue monitoring en adaptieve (beleids-) maatregelen, want wetgeving is belangrijk maar simpelweg te traag.

Onderzoekspoor 2: AI-veiligheid Instituten (AISIs)
Ontwikkelingen in Cybersecurity en AI gaan in hoog tempo. In eerdere onderzoeken wordt daarom ook aangegeven dat de inschatting van kansen en bedreigingen geen éénmalige opgave is, maar continue monitoring vergt.

De beleidsuitdaging van snel veranderende AI-risico's en beperkte informatie om beleid op te baseren, wordt internationaal herkend.

Daarom kiezen koplopers voor het opbouwen van permanente capaciteit die risico's continu kunnen beoordelen en duiden. Het AISI-model is daarmee in de praktijk een antwoord op het evidence dilemma. AISI's zijn ontworpen om twee werelden te verbinden. Enerzijds leveren ze technisch diepgaand inzicht: onderzoek, testen en evalueren van AI. Anderzijds maken ze die inzichten bruikbaar door handelingsperspectief te geven aan overheid en de private sector.

Op deze wijze zijn AISI's een middel om veilige AI innovatie te stimuleren. Organisatorisch laat het internationale beeld zien dat AISI's bestaan uit een compact maar hoogkwalitatief kernteam, dat van voldoende middelen moet worden voorzien. De effectiviteit hangt mede af van toegang tot relevante testmogelijkheden en van werkbare samenwerking met AI ontwikkelaars (i.e. Frontier-AI Labs) om vroegtijdige evaluaties mogelijk te maken. Het internationale netwerk versterkt dit door kennis en methoden te delen en door toe te werken naar meer standaardisatie van evaluatiepraktijken.

Onderzoekspoor 3: Marktconsultatie
Om op te halen of er ook binnen Nederland behoefte is aan een nationale AISI, en welke taakstelling en kaders wenselijk zijn, is een 25 tal interviews afgenomen met AI belanghebbenden werkzaam bij drie typen organisaties: publieke organisaties, private organisaties en (internationale) kennisinstellingen.

Op basis van de interviews blijkt er een duidelijke behoefte te zijn aan een NL AISI. In de interviews is informatie opgehaald over de gewenste taakstelling, kritieke succesfactoren en overige kaders. Daaruit komt naar voren dat een NL AISI alle drie de kerntaken van een AISI dient uit te voeren, met een focus op basis van de Nederlandse context.

- (i) Met onderzoek en innovatie nieuwe kennis, methoden, tools en datasets ontwikkelen voor AI-veiligheid, met focus op AI binnen voor Nederland relevante veiligheidsdomeinen (deze zijn nader te bepalen en onderhavig aan voortschrijdend inzicht).
- (ii) Monitoren, testen en evalueren van AI ontwikkelingen en risico's specifiek relevant voor Nederland, inclusief periodieke risicoduiding, voor ten minste de ministeries Binnenlandse Zaken (BZK), Buitenlandse Zaken (BZ), Justitie en Veiligheid (J&V), Defensie, Economische Zaken (EZK) en geïnteresseerde private organisaties. Een NL AISI dient zich juist niet te richten op handhaving en toezicht op wetgeving.
- (iii) Richting geven via gezaghebbende adviezen aan overheid (primair aan politieke en ambtelijke bestuurders) en het bredere publiek, maar ook door middel van praktische richtlijnen, standaarden en tools aan de private sector en uitvoerende publieke instanties.

Als kritieke succesfactoren wordt aangegeven dat een NL AISI mee moet kunnen bewegen in de sterke dynamiek van AI-veiligheid. Haar autoriteit dient het te ontlenen aan haar deskundigheid en inspirerende boegbeelden, waarmee draagvlak wordt verkregen bij publieke, private en kennisorganisaties. Om een leidende rol te kunnen vervullen dient focus gelegd te worden op sub-domeinen, moeten voldoende middelen gealloceerd worden en dient actief samen gewerkt te worden in het internationale AISI netwerk en met toonaangevende AI Labs.

Onderzoekspoor 4: inventarisatie AI-veiligheid initiatieven in Nederland
Ten slotte, is geïnventariseerd in hoeverre de taakstelling van een nationale AISI al is belegd binnen Nederland en bij welke organisaties en samenwerkingsverbanden.

De inventarisatie van Nederlandse AI- en cybersecurity initiatieven laat zien dat relevante activiteiten uitgevoerd worden, maar dat er geen organisatie is die de AISI-taakstelling in samenhang vervult of coördineert. Onderzoek, monitoring en advisering zijn verspreid over meerdere partijen zonder structurele verbinding. Daarbij worden Cybersecurity en AI activiteiten gescheiden van elkaar opgepakt, waardoor het thema AI-veiligheid niet in de volle breedte wordt geadresseerd. Ook valt op dat er geen expliciete focus ligt op *general-purpose* AI, terwijl op dit terrein de systemische risico's zich in snel tempo opstapelen, en de kansen voor economische groei onbenut blijven. Voor deze risico's is er geen voortdurende monitoring en duiding ingericht, waardoor handelingsperspectief voor het evidence dilemma ontbreekt.

Als gevolg hiervan ontbreekt in het huidige Nederlandse initiatieven aan een duidelijke organisatorische concentratie van taken en middelen voor het structureel monitoren, testen en duiden van risico's en kansen van AI. Hiermee ontbreekt het in Nederland aan slagkracht om tot handelingsperspectief voor AI-veiligheid te komen.

Conclusies
Het inzichtelijk maken van de ***kansen en dreigingen binnen het raakvlak van cybersecurity en AI*** is onderhevig aan het evidence dilemma, dat wil zeggen dat nieuwe ontwikkelingen zich zo snel manifesteren dat beleid niet kan leunen op volledige informatie. Daarom dienen de kansen en dreigingen op dit raakvlak continue gemonitord te worden. Een NL AISI lijkt een geschikt instrument daarvoor, welke in een aantal andere landen met succes zijn opgericht.

Voor het effectief ***monitoren van de ontwikkelingen*** binnen het raakvlak van cybersecurity en AI is een NL AISI taakstelling uitgewerkt, en er is opgehaald door middel van een marktconsultatie welke taken het meest wenselijk zijn volgens verschillende belanghebbenden. De belanghebbenden bestonden uit drie doelgroepen, namelijk publieke organisaties, private organisaties en (internationale) kennisinstellingen.

De oprichting van een NL AISI kan gezien worden als een ***duurzame beleidsmatige interventie*** om de verschillende belangen te balanceren. Zowel op het gebied van beheersing van ontsporende risico's in het raakvlak van cybersecurity, en voor het benutten van braakliggende kansen voor het verdienvermogen.

Aanbevelingen
Naast de eerder genoemde taakstelling voor een NL AISI zijn uit dit onderzoek ook kaders af te leiden voor de missie en organisatorische inrichting van een NL AISI, welke als aanbevelingen in meer detail uiteen zijn gezet.

Inhoudsopgave

Begrippenkader

AI: in dit rapport wordt onder “AI” verstaan wat in de literatuur ook aangeduid wordt als “general-purpose AI”. Deze term is verder gedefinieerd in sectie 1.1 van het International AI Safety report [5].

AISI: in dit rapport wordt de afkorting AISI gebruikt als verzamelterm voor AI Safety Instituten en AI Security Instituten. Internationaal worden beide benamingen gebruikt om te verwijzen naar instituties die gecentraliseerd technische en organisatorische capaciteit opbouwen om risico's van AI te evalueren en te vertalen naar richtinggevende inzichten voor beleidsmakers en andere betrokkenen.

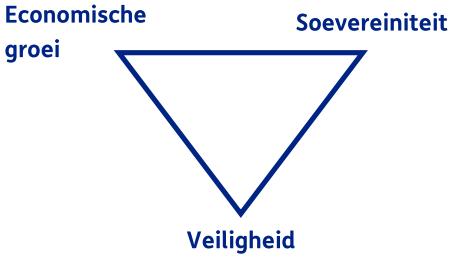
AI-veiligheid: in dit rapport wordt de term AI-veiligheid als verzamelterm gebruikt voor de vakgebieden AI Security en AI Safety. AI Security is in dit rapport gedefinieerd als het beschermen van AI systemen tegen aanvallen, manipulatie en cyberdreigingen. AI Safety is een breder werkveld dat gericht is op het voorkomen dat AI-systemen een gevaar vormen voor publieke belangen zoals (nationale) veiligheid, maatschappelijk welzijn en gezondheid, en mensenrechten.

Inleiding

AI als strategische technologie
Kunstmatige intelligentie (AI) is uitgegroeid tot een strategische technologie voor wereldmachten. Grote economieën investeren fors in AI, gedreven door de verwachting dat deze technologie economische groei, geopolitieke invloed en nationale veiligheid versterkt. Recente IMF-scenario's [8] suggereren dat AI het wereldwijde BBP-niveau in de komende 10 jaar met ongeveer 1,3% tot bijna 4% kan verhogen, waarbij de grootste stijging wordt voorzien voor de Verenigde Staten (VS). De VS zet dan ook volop in op de technologie en positioneert AI in de veiligheidsstrategie als sleuteltechnologie voor strategisch overwicht en mondiale invloed, waarvoor grootschalige inzet noodzakelijk is [9].

De opkomst van AI versterkt een gestaag groeiende strategische afhankelijkheid van Europa en Nederland van niet-Europese technologieën. Digitale dienstverlening waaronder social media, cloud en generatieve AI is grotendeels in handen van Amerikaanse spelers, terwijl China stevig investeert in geïntegreerde AI-ecosystemen en robotica [i]. Voor kritieke sectoren en overheid is deze afhankelijkheid steeds moeilijker te verenigen met nationale veiligheid, economische weerbaarheid en beleidsautonomie. Daarom staat strategische autonomie centraal in Europese en Nederlandse beleidsagenda's: het vermogen om in strategisch belangrijke digitale domeinen onafhankelijk te kunnen kiezen en opereren, als noodzakelijke stap richting digitale soevereiniteit [ii, iii].

Om de gemoeide publieke belangen te beschermen is nationale en Europese beleidsontwikkeling in volle gang. In het "Draghi rapport" [2] wordt gewaarschuwd dat Europa structureel achterloopt in digitale sleuteltechnologieën zoals AI, met toenemende afhankelijkheid van de Verenigde Staten en China tot gevolg. Nederland neemt hierin een ambivalente positie in: kwetsbaar door afhankelijkheden, maar strategisch sterk door de unieke rol in de mondiale halfgeleiderindustrie. Het Wennink-rapport [1], roept daarom op tot meer investeringen in de gehele waardeketen, een lijn die wordt onderschreven in het coalitieakkoord "Aan de slag (2026–2030)" [3], waarin de transitie in Nederland van 'pilotland' naar 'opschaalland' centraal staat.



Figuur 1: de strategische ambities voor AI van de EU en Nederland

Van belang is hierbij dat de ontwikkelingen van AI technologie sneller gaan dan deze zich bij bijvoorbeeld social media en cloud technologie hebben voorgedaan. Bovendien hebben deze digitale technologieën een onderling versterkend effect op het ontwikkelingstempo. Daarmee is de tijdspanne waarin afhankelijkheid van AI technologie groeit kleiner dan voor voorgaande technologische afhankelijkheden. De oproep in het Wennink-rapport om soeverein handelingsperspectief en duurzame innovatie kracht op te bouwen in de sterke AI dynamiek, is dus ook urgent.

AI en Cybersecurity
Parallel aan de AI (beleids-) ontwikkelingen is ook het beleid op de sleuteltechnologie Cybersecurity sterk in ontwikkeling. Voornamelijk via regelgeving en stimuleringsinitiatieven om de veiligheid van producten en diensten te vergroten, en de digitale weerbaarheid van burgers, bedrijven en publieke organisaties te versterken. Met de ontwikkelingen in zowel het AI als het Cybersecurity vakgebied ontstaat er steeds meer behoefte aan strategisch beleid op het raakvlak van beide velden.

Sinds 2025 wordt in het International AI Safety Report [5] periodiek een wetenschappelijke onderbouwing gepubliceerd over de ontwikkelingen van AI risico's. Het rapport onderscheidt risico's van kwaadwillig gebruik (e.g. cyber offensieve capaciteiten), risico's door onbetrouwbaarheid van AI en systeem-risico's (zoals marktconcentratie). Deze risico's groeien mee met de toename van AI-capaciteiten en –adoptie. Ook wordt in het rapport benadrukt dat deze ontwikkelingen moeilijk te voorspellen zijn. Expliciet wordt gewezen op het 'evidence dilemma' dat dit meebrengt: het beleid moet meebewegen met snelle ontwikkeling, zonder te kunnen leunen op volledige informatie.

Nationaal Cybersecurity en AI beleid
Ter onderbouwing voor de ontwikkeling van cybersecurity & AI beleid door de Nederlandse overheid voerde TNO in 2024 een eerste verkenning uit [4]. Inzichtelijk werd gemaakt dat het raakvlak van Cybersecurity en AI op te delen is in vier deelgebieden: AI voor cybersecurity, AI voor cyberaanvallen, cybersecurity van AI, en aanvallen op AI. Elk van deze kwadranten kent verschillende dynamiek, actoren en publieke belangen, en wat mist is een integrale sturing. In de praktijk zijn verantwoordelijkheden vaak langs bestaande lijnen van "AI-beleid" en "cyberbeleid" georganiseerd, terwijl de wisselwerking duidelijk aangeeft dat er samenhang nodig is over de beleidsterreinen heen. De beleidsmatige uitdaging is om kennisgebieden en actoren rond AI en cyberveiligheid samen te brengen in een ecosysteem, om zo gezamenlijk van duiding naar handelingsvermogen te komen. In de context van de snelle ontwikkelingen rondom AI-toepassingen werd in 2024 al opgeroepen tot een proactieve aanpak hiervoor [4].

Onderzoeksvragen en aanpak
Dit rapport is opgesteld in opdracht van het Ministerie van Economische Zaken en Klimaat (EZK), vanuit de ambitie om de beleidsdomeinen van AI en Cybersecurity in Nederland beter te verbinden, zodat publieke belangen rond veiligheid, soevereiniteit en economische groei beter geborgd kunnen worden. Hierbij staan de volgende vragen centraal:

- Welke kansen en dreigingen in het raakvlak van cybersecurity en AI zijn het meest relevant t.a.v. verdien-vermogen en de publieke belangen?
- Hoe kunnen cybersecurity en AI ontwikkelingen en hun invloed op verdienvermogen en publieke belangen worden gemonitord?
- Welke beleidsmatige interventies zijn meest geschikt voor het balanceren van elk van die belangen?

Dit onderzoek is uitgevoerd in 4 onderzoeksporen: een scenarioanalyse, duiding AI-veiligheid Instituten (AISIs), een marktconsultatie en een inventarisatie van nationale AI-veiligheid initiatieven.

Onderzoekspoor 1: Scenarioanalyse
In het voorgaande onderzoek van TNO [4] zijn AI en Cybersecurity gerelateerde kansen en bedreigingen op hoofdlijnen geschetst en geïllustreerd met voorbeelden. In het International AI safety Rapport [5] is een beeld geschetst o.b.v. recente weten-schappelijke inzichten hierover. Hoewel deze bronnen een actueel beeld schetsen blijft onduidelijk welke concrete invloed die kansen en bedreigingen hebben op Nederlandse private en publieke organisaties, en burgers. En daarmee ontbreekt het zicht op concrete Cybersecurity & AI beleidsprioriteiten.
Om een meer concreet fundament te geven voor het opstellen van een doelgerichte aanpak voor Cybersecurity en AI, zijn in dit onderzoek scenario's opgesteld en getoetst bij private en publieke organisatie. Dit is in Onderzoekspoor 1 uitgewerkt.

Onderzoekspoor 2: AI-veiligheid Instituten (AISIs)
Ontwikkelingen in Cybersecurity en AI gaan in hoog tempo. In eerdere onderzoeken [4, 5, 6, 7] wordt daarom ook aangegeven dat de inschatting van kansen en bedreigingen geen éénmalige opgave is, maar continue monitoring vergt. Dit wordt internationaal erkent, waarbij sommige landen reageren met de inrichting van een nationaal AI-veiligheid Instituut (AISI). Welke landen een AISI hebben opgericht en hoe zij hier invulling aan geven is uitgewerkt in Onderzoekspoor 2.

> ***"# 46 Richt een Nationaal AI Impact Instituut op.***
>
> *Het instituut wordt onderdeel van een internationaal netwerk van publieke AI-instituten die kennis uitwisselen en neemt het Britse AI Security Institute als voorbeeld voor effectieve inrichting, werkwijze en organisatiecultuur."*
>
> **Bron: Nationaal AI Deltaplan [11]**

Onderzoekspoor 3: Marktconsultatie
Om op te halen of er ook binnen Nederland behoefte is aan een nationale AISI, en welke taakstelling en kaders wenselijk zijn, is een 25 tal interviews afgenomen verdeeld over drie doelgroepen: publieke organisaties, private organisaties en kennisinstellingen. Dit is uitgewerkt in Onderzoekspoor 3.

Onderzoekspoor 4: AI-veiligheid initiatieven in Nederland
Ten slotte, is onderzocht in hoeverre de taakstelling van een nationale AISI al is belegd binnen Nederland en bij welke organisaties en of samenwerkingsverbanden.

Het rapport eindigt met een samenvatting van de conclusies uit de vier onderzoeksporen, de beantwoording van de onderzoeksvragen en aanbevelingen voor vervolg.

Onderzoekspoor 1: AI-veiligheid Scenarioanalyse

Doel van Onderzoekspoor 1
Het doel van de scenarioanalyse is om een reëel beeld te verkrijgen over de toekomstige kansen en bedreigingen in het raakvlak van Cybersecurity en AI, zoals die door Nederlandse publieke en private organisaties worden ingeschat. Een uitdaging hierbij is dat de toepassing en adoptie van AI en de mate van Cybersecurity volwassenheid verschillen per sector (en in mindere mate ook tussen organisaties in sectoren). Daar komt bij dat in het jonge AI-veiligheid domein de meeste organisaties nog nauwelijks ervaring hebben, die relevant is voor hun specifieke bedrijfsvoering. Dit maakt het lastig om reële inschattingen te verkrijgen op basis van de dagelijkse praktijk. Scenarioanalyse biedt de mogelijkheid om gerapporteerde AI-veiligheid gebeurtenissen te vertalen naar de context van specifieke sectoren om daarmee beter aan te sluiten bij de dagelijkse praktijk.

Opstellen AI-veiligheid scenario's
Om een breed beeld van AI-veiligheid kansen en dreigingen in kaart te brengen zijn een behapbaar aantal scenario's opgesteld, die gespreid zijn over de sectoren energie, financieel, telecom, mobiliteit, overheid en RD&I (Research, Development & Innovatie). In de keuze voor deze sectoren is meegenomen dat de sector een aanzienlijk publiek belang dient (en vanuit de cyberbeveiligingswet een zorg- en rapportageverplichting hebben), of in het geval van de RD&I kennissector een speciale rol heeft t.a.v. de ontwikkeling en gebruik van AI.

Naast de spreiding over sectoren zijn de scenario's verdeeld over diverse kansen en dreigingen. Deze kansen en dreigingen komen voort uit de verkenning van het raakvlak van Cybersecurity en AI [4] en naderhand gepubliceerde AI-veiligheid gebeurtenissen (o.a. in [5]). Door deze aanpak die voortbouwt op de verkenning [4], dekken de scenario's de in [4] geïntroduceerde kwadranten af. In Figuur 2 zijn de opgestelde scenario's per kwadrant weergegeven.

Korte beschrijvingen van de scenario's zijn opgenomen in de bijlage. De vetgedrukte scenario's in Figuur 2 zijn in meer detail uitgewerkt. Voor die scenario's zijn fictieve 'story-lines' opgesteld aan de hand waarvan inschattingen zijn gemaakt van de impact op publieke belangen.

Toetsing van scenario's
De door het TNO onderzoeksteam opgestelde scenario's zijn besproken met de opdrachtgever en collega's van het programmateam AI, onderdeel van Economische Zaken. Vervolgens is getoetst bij vertegenwoordigers van sectorale belangenverenigingen, welke van de AI-veiligheid scenario's hen relevant lijken. Contact is opgenomen met: Nederlandse Vereniging van Banken, Topsector Energie en Coalitie voor Cyberveilige Energietransitie, de Verenging van Nederlandse Gemeenten en Cyberveilig Nederland. Daarbij is specifiek gezocht naar vertegenwoordigers wiens werkzaamheden namens hun belangenvereniging gericht zijn op onderwerpen AI en/of cybersecurity. Voor de sectoren Telecom en Mobiliteit is het niet gelukt om een geschikt contact te vinden, en voor RD&I heeft het TNO onderzoeksteam haar eigen beeld over de sector gebruikt.

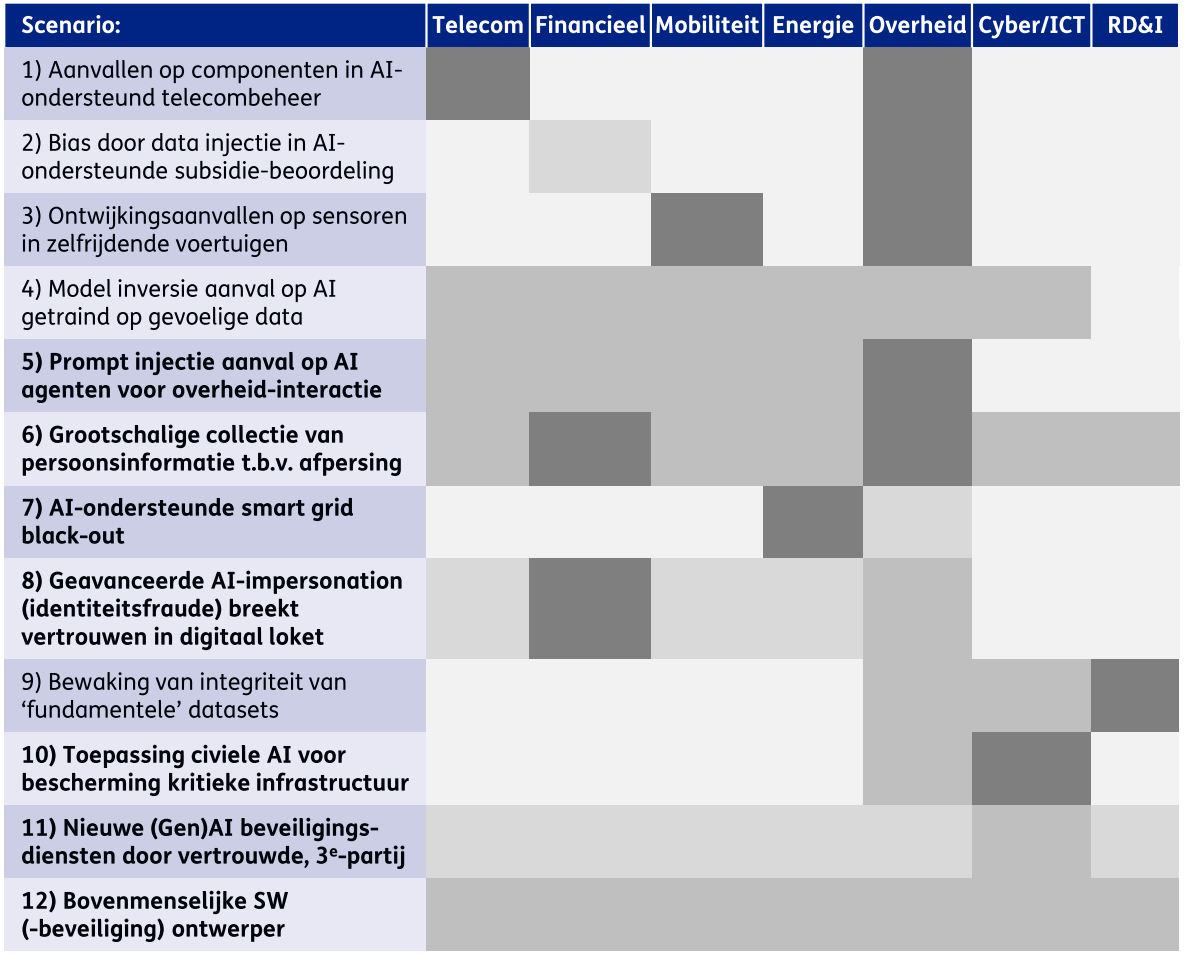
6) Grootschalige collectie van persoonsinformatie t.b.v. afpersing **7) AI-ondersteunde smart grid black-out** **8) Geavanceerde AI-impersonation (identiteitsfraude) breekt vertrouwen in digitaal loket** *AI voor cyberaanvallen*	1) Aanvallen op componenten in AI-ondersteund telecombeheer 2) Bias door data injectie in AI-ondersteunde subsidie-beoordeling 3) Ontwijkingsaanvallen op sensoren in zelfrijdende voertuigen 4) Model inversie aanval op AI getraind op gevoelige data **5) Prompt injectie aanval op AI agenten voor overheid-interactie** *Aanvallen op AI*
AI voor cybersecurity **10) Toepassing civiele AI voor bescherming kritieke infrastructuur** **12) Bovenmenselijke SW (-beveiliging) ontwerper**	*Cybersecurity van AI* 9) Bewaking van integriteit van 'fundamentele' datasets **11) Nieuwe (Gen)AI beveiligings-diensten door vertrouwde, 3e-partij**

Figuur 2: Opgestelde AI-veiligheid scenario's per kwadrant

De acties om de scenario's te valideren en impact inschattingen vanuit de dagelijkse praktijk te verkrijgen waren beperkt succesvol. Vanuit de VNG en Topsector Energie is aangegeven welke scenario's voor hen het meest relevant zijn (zie de donkerste grijstint in de kolommen Energie en Overheid), en op basis van hun feedback zijn scenario's aangescherpt. Voor Cyberveilig Nederland is de inschatting van meest relevante scenario's gemaakt door het TNO team en zijn de betreffende scenario beschrijvingen met de leden van de belangenvereniging gedeeld. Hier is geen specifieke feedback op ontvangen. Binnen de financiële sector wordt aangegeven dat er al een aantal lopende samenwerkingen zijn waarbinnen de (cybersecurity) risico's van AI worden besproken, inclusief scenario's welke specifiek van toepassing zijn op de financiële sector, onder andere geleid door de NVB. De kansen en risico's van AI en cybersecurity worden niet in samenwerkingsverband besproken en pakt iedere financiële organisatie voor zich op.

Toepassing van scenario's

Buiten de afbakening van dit onderzoek zijn de scenario's ook gebruikt voor andere AI-veiligheid onderzoeken en events. Deze waren gericht op specifieke onderwerpen zoals 'de rol van AI t.a.v. de weerbaarheid van het energiesysteem', 'AI-gedreven cyberoperaties en gerubriceerde gegevens' en 'AI & Digitale veiligheid'. Kortom, de opgestelde AI-veiligheid scenario's blijken een nuttig hulpmiddel, maar vooral in een specifieke context.



Figuur 3: AI-veiligheid scenario's toegespitst op sectoren
(Meer donkere grijstint geeft hogere relevantie voor de betreffende sector weer)

Observaties
Hoewel de scenarioanalyse niet volledig is uitgevoerd, komen er een aantal observaties uit naar voren.

Ten eerste, blijkt AI-veiligheid bij organisaties wel op de agenda te staan, maar ieder voor zich worstelt met de vraag of en hoe dat aan te pakken. De spreiding van relevante scenario's over de sectoren in Figuur 3 illustreert dat AI-veiligheid zich voor organisaties verschillend manifesteert. Dit terwijl veel scenario's (i.h.b. 4, 5, 6, 8 en 12) sterk vergelijkbaar zijn tussen sectoren en de mogelijkheid om van elkaar te leren gemist lijkt te worden. Ook is het contrast tekenend tussen de beperkte respons op de brede uitvraag naar de impact van AI-veiligheid scenario's, t.o.v. de ruime interesse in de behandeling van specifieke scenario's.

Een andere observatie betreft vragen over de scenario's in de context van opkomende wetgeving: in hoeverre worden de gevolgen van de scenario's voldoende beheerst via de AI-verordening of de cyberweerbaarheidswet? De invoering van de wetgeving lijkt echter niet bij te gaan dragen aan AI-veiligheid. Het raakvlak van Cybersecurity en AI wordt vanuit beide dossiers gezien als een beleidsmatige 'niche' en krijgt vanuit beide invalshoeken niet de aandacht die nodig lijkt te zijn.

Ten derde is tijdens de uitvoering van de scenarioanalyse gebleken hoe dynamisch het AI-veiligheid domein zich ontwikkelt. Waar de discussie over sommige scenario's zich vóór de zomer van 2025 nog toespitste op het te futuristische gehalte, werd in het najaar reeds gepubliceerd over het voltrekken van diezelfde scenario's. Specifieke voorbeelden zijn de publicatie door de AFM over financiële oplichting via generatieve AI identiteitsfraude (scenario 8) en de introductie van autonome software ontwikkeling met Claude Code (Security) en de waarschuwingen voor de cyber risico's van deze ontwikkeling (scenario 12). Verder viel de Anthropic's publicatie [10] over de eerste AI georkestreerde cyber spionage campagne op; dit was begin 2025 als scenario overwogen maar niet verder uitgewerkt omdat het te futuristisch zou overkomen.

Conclusie
Deze geleerde lessen uit de scenarioanalyse bevestigen het evidence dilemma voor beleidsmakers. AI-veiligheid is gebaat bij een gecoördineerde, sector-overstijgende aanpak en deze ontstaat niet vanzelf. Verder vergt de dynamiek van AI-veiligheid continue monitoring en adaptieve (beleids-) maatregelen (mogelijkerwijs door middel van een sign-posts of change methode zoals voorgesteld in [6]), want wetgeving is belangrijk maar simpelweg te traag.

Onderzoekspoor 2: Opkomst AI-veiligheid Instituten

Doel van Onderzoekspoor 2
De scenarioanalyse in Onderzoekspoor 1 laat zien dat ontwikkelingen in AI-veiligheid zich in snel tempo voltrekken met specifieke risico's voor (kritieke) sectoren. Het duiden van risico's en het vertalen naar handelingsperspectief blijkt geen eenmalige exercitie, maar vraagt continue monitoring die sector overstijgend georganiseerd moet zijn om systemische effecten te adresseren. Deze conclusie past bij eerdere oproepen van de WRR en het Nationaal AI Deltaplan voor een orgaan dat AI-veiligheid kennis kan bundelen, richting kan aanbrengen in de voor de overheid relevante vraagstukken en op internationaal niveau kan samenwerken met vergelijkbare AI-veiligheidorganen.

Tegen deze achtergrond is Onderzoek-spoor 2 gericht op de aanpak van dit beleidsprobleem in andere landen. Daarbij blijkt een belangrijke rol weggelegd voor AI Safety/Security Instituten (AISI's). Dit zijn gespecialiseerde organen die technische evaluatie en risicoduiding organiseren en die inzichten vertalen naar richtinggevende input voor beleid en veilige adoptie. Het doel van dit onderzoekspoor is in kaart brengen hoe AISI's in andere landen zijn georganiseerd, met welke doelstelling, taakstelling, financiële middelen en mandaat.

Achtergrond – eerdere oproepen voor AI (Veiligheid) instituut
Reeds in 2021 riep de WRR op tot de oprichting van een AI coördinatie instituut in haar rapport 'Opgave AI. De nieuwe systeemtechnologie' [7]. In het onderzoekswerk wordt AI beschreven als systeemtechnologie die net als elektriciteit en de verbrandingsmotor een transformerende werking gaat hebben op economie en samenleving. De onderzoekers schetsen dat AI het 'onderzoekslab' begint te verlaten en dat het in steeds meer toepassingen zijn weg vindt naar buiten. Als systeemtechnologie is AI breed toepasbaar en overheden en bedrijven zien deze technologie als bron van toekomstige economische groei en betere dienstverlening. Tegelijkertijd is er ook een grote impact op publieke waarden en de (inter)nationale veiligheid en is de dynamiek hiervan onvoorspelbaar.

Om de doorwerking van AI als nieuwe systeemtechnologie te begeleiden pleit de WRR voor het opbouwen van een beleidsinfrastructuur voor AI, te beginnen met het oprichten van een "AI-coördinatiecentrum voorzien van politieke verankering middels een ministeriele onderraad". Het voorgestelde onafhankelijke centrum zou moeten dienen als 'contactpunt voor internationale organisaties' en voor de Nederlandse overheid zou het samenwerking moeten stimuleren en focus moeten aanbrengen in relevante risico's en kansen. Qua organisatie zou het centrum gevoed moeten worden door een 'externe AI-Raad van prominente experts' die de overheid informeren met kennis vanuit de wetenschap en het bedrijfsleven.

Inmiddels zijn vijf jaar verstreken en in de tussentijd is er veel veranderd. Met de komst van generatieve AI modellen zoals ChatGPT, sprong AI vanuit het lab de samenleving in en werd het (nog meer) een strategische technologie voor wereldmachten. De roep om een instituut voor AI lijkt daarmee meer weerklank te vinden. Het onlangs gepubliceerde Nationaal AI Deltaplan beschrijft een 'impact instituut' dat zich internationaal kan aansluiten bij publieke AI-instituten zoals het Britse AI Security Instituut [11]. Hiervoor moet het instituut experts uit verschillende domeinen bij elkaar brengen, om AI-ontwikkelingen te monitoren en gericht advies te brengen aan kabinet en overheid. Het Britse AI Security Instituut (UK AISI) staat daarbij niet op zichzelf, maar is onderdeel van een netwerk van internationale AI-veiligheid instituten.

Internationaal netwerk van AI Safety/Security Instituten
Verschillende landen hebben AI Safety/Security Institutes (AISI's) opgericht om het evidence dilemma praktisch te adresseren met een adaptieve aanpak van onderzoek, testen en bieden van handelingsperspectief. Sinds 2023 hebben het Verenigd Koninkrijk, Japan, Canada en Singapore officiële instituten opgericht. Daarnaast heeft de Europese Unie via het AI Office een vergelijkbare veiligheidsfunctie ingericht, zij het dat zij ook een expliciete toezichthoudende rol hebben bij de handhaving van de AI-verordening [12]. Gezamenlijk vormen deze instituten een internationaal netwerk dat zich richt op gezamenlijk onderzoek, gedeelde testmethoden en best-practices, waardoor een raamwerk ontstaat waarin overheden hun technische capaciteit bundelen om AI-veiligheid te waarborgen en verantwoord innovatiebeleid te ondersteunen.

Doelstelling van AISI's
Het Amerikaanse Center for Strategic & International Studies (CSIS) beschrijft dat de opkomst van AI Safety/Security Institutes (AISI's) voortkomt uit een groeiende behoefte om grensverleggende AI-systemen beter te kunnen testen en evalueren, risico's beheersbaar te houden en internationale samenwerking te versterken [12]. Die redenering sluit aan bij de WRR-positionering van AI als systeemtechnologie: een breed toepasbare technologie met groot economisch potentieel, maar ook met een onvoorspelbare doorwerking op publieke waarden en veiligheid, waardoor structurele inbedding en regie randvoorwaardelijk worden om kansen te kunnen benutten [7].

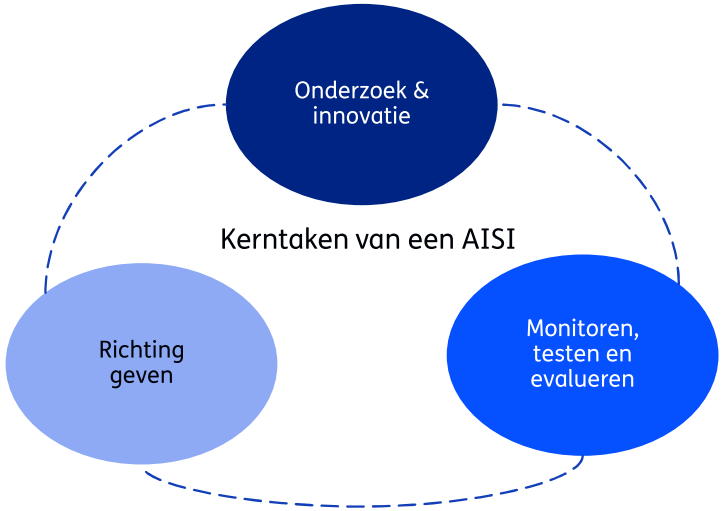
De doelstelling van AISI's is daarom om de veiligheid van AI te bevorderen door het ontwikkelen van de 'remmen' die nodig zijn om verantwoord te kunnen versnellen. In een vroege fase kan er veel misgaan door gebrek aan ervaring en duidelijke regels, wat vraagt om instituties die adaptief kunnen monitoren, evalueren en duiden. Die doelstelling laat zich beleidsmatig opsplitsen in **(i) technische veiligheid**: model- en systeemgericht begrip opbouwen van gedrag, robuustheid en risico's via onderzoek, testen en evaluatie; en **(ii) procesmatige veiligheid**: de bestuurlijke en maatschappelijke inbedding organiseren die nodig is om inzichten tijdig te vertalen naar normen & richtlijnen en naar concreet handelingsperspectief, inclusief internationale afstemming. Zo versterken AISI's het vertrouwen in de technologie en het vermogen om de snelle transitie in veilige banen te begeleiden.

Taakstelling van AISI's
Om de veiligheid van AI te bevorderen werken de AISI's aan de drie samenhangende kerntaken van onderzoek, testen en richting geven (zie Figuur 4).

- **Onderzoek & innovatie**: AISIs ontwikkelen een wetenschappelijke basis voor AI-veiligheid door onder-zoek te doen naar AI-model-gedrag, risico's, veiligheidsmechanismen en mogelijke interventies.
- **Monitoren, testen en evalueren**: AISIs testen nieuw aangeboden AI-modellen om hun capaciteiten, risico's en onvoorspelbaar gedrag te begrijpen.
- **Richting geven aan publieke en private sectoren**: AISIs vertalen deze inzichten naar bruikbare duiding voor beleid en praktijk, zodat overheden en private organisaties geïnformeerde besluiten kunnen nemen. De inzichten ondersteunen de besluitvorming en bevorderen consensus over testmethoden en evaluatiepraktijken. Ze slaan de brug tussen technische analyse en concreet beleid.

Organisatie en randvoorwaarden
Het CSIS-rapport schetst dat het AISI-netwerk in de startfase bestaat uit circa tien betrokken partijen. Voor de zeven partijen waarvoor publiek de meeste organisatie-informatie beschikbaar is, is een overzicht opgenomen in Tabel 1.

Naast deze zeven AISI's noemt CSIS ook Frankrijk, Kenia en Australië als (beoogde) netwerkleden, maar daarvoor is de beschikbare informatie beperkt. Aanvullend op het CSIS-rapport geldt dat Frankrijk in januari 2025 INESIA heeft opgericht met de missie om nationale veiligheidsspelers te verbinden rond AI-veiligheid [13], en dat Australië in november 2025 aankondigde begin 2026 een AI-veiligheidsinstituut te lanceren [14].



Figuur 4: De kerntaken van een AISI, bron: [12]

Organisatorisch worden de AISI's in het netwerk door CSIS gepositioneerd als publiek gefinancierde onderzoeks- en uitvoeringscapaciteit, ondergebracht bij publieke instanties, met als doel om voor overheden in-house technische expertise en organisatorische capaciteit op te bouwen voor het evalueren en monitoren van risico's van AI-modellen voor publieke en nationale veiligheidsbelangen. Deze uitvoering wordt gedaan met organisaties van enkele tientallen (technische), tot 100+, medewerkers. In 2024 varieerden de investeringen voor de oprichting van nationale AISI's tussen de $7,2 en $65 miljoen per jaar (om te rekening naar circa €5 tot €55 miljoen). Het Verenigd Koninkrijk meldt daarbij dat het beschikking heeft tot 1,5 miljard pond aan prioriteitsfinanciering voor extra rekenkracht (compute) voor intensieve AI-model-testen. Tot slot benadrukt CSIS dat effectieve taakuitvoering vraagt om nauwe samenwerking met frontier AI-ontwikkelaars om testen en evaluaties voordat ze breed worden ingezet mogelijk te maken.

Conclusie

De beleidsuitdaging van snel veranderende AI-risico's en beperkte informatie om beleid op te baseren, wordt internationaal herkend. Daarom kiezen koplopers voor het opbouwen van permanente capaciteit die risico's continu kunnen beoordelen en duiden. Het AISI-model is daarmee in de praktijk een antwoord op het evidence dilemma. AISI's zijn ontworpen om twee werelden te verbinden. Enerzijds leveren ze technisch diepgaand inzicht: onderzoek, testen en evalueren van AI. Anderzijds maken ze die inzichten bruikbaar door handelingsperspectief te geven aan overheid en de private sector.

Hiermee zijn AISI's een middel om de mate van veiligheid te realiseren, die randvoorwaardelijk is voor snelle innovatie met en van AI. Organisatorisch laat het internationale beeld zien dat een AISI bestaat uit een compact maar hoogkwalitatief kernteam dat is voorzien van voldoende budget. Effectiviteit hangt daarnaast af van toegang tot relevante test-mogelijkheden en van werkbare samenwerking met ontwikkelaars van toonaangevende AI labs om vroegtijdige evaluaties mogelijk te maken. Het internationale netwerk versterkt dit door kennis en methoden te delen en door toe te werken naar meer standaardisatie van evaluatiepraktijken.

	Verenigde Staten	Verenigd Koninkrijk	Europese Unie	Japan	Singapore	Zuid-Korea	Canada
Naam	Centre for AI Standards and Innovation (CAISI)	AI Security Institute (UK AISI)	EU AI Office	Japan AISI	Singapore AISI	Korea AISI	Canada AISI
Oprichting	Februari 2024	November 2023	Mei 2024	Februari 2024	Mei 2024	November 2024 *	November 2024 *
Budget (USD & eigen valuta)	$ 10 miljoen (FY24)	>$65 miljoen/jaar (>£ miljoen/jaar)	$51 miljoen (€46.5 miljoen)		$7.5 miljoen/jaar (S$10 miljoen/jaar)	$7.2-14.4 miljoen/jaar (W10-20 miljard/jaar)	$38.5 miljoen (C$ miljoen)
Staf	c. 20	100+ technische staf *	c. (AI-veiligheid onderdeel)	c. 23		Min. 30	

Tabel 1: Organisatie informatie over AISI instituten uit het internationale netwerk. Tabelinformatie is overgenomen uit CSIS rapport. Informatie in cellen gemarkeerd met * is geüpdatet in Februari 2026.

Onderzoekspoor 3: Marktconsultatie

Doel van Onderzoekspoor 3
Het doel van de marktconsultatie is om te achterhalen in hoeverre Nederlandse organisaties waarde zien in de mogelijke oprichting van een Nederlands AI-veiligheid instituut (NL-AISI). Hiervoor zijn 25 interviews afgenomen met belanghebbenden uit verschillende domeinen. De belanghebbenden zijn organisaties die een aantoonbaar belang hebben bij de totstandkoming, werking of uitkomsten van een NL-AISI, omdat zij AI ontwikkelen of gebruiken, dan wel AI risico's beheersen of beleid en toezicht hiervoor vormgeven. De geïnterviewden zijn specifiek gevraagd naar het belang dat zij hechten aan de NL-AISI(-taken).

Doelgroepen
De belanghebbenden zijn daarbij ingedeeld in drie doelgroepen:
(1) publieke partijen en autoriteiten die faciliterend optreden in AI-veiligheid advies en toezicht, dan wel richting geven door beleid- en strategievorming.
(2) private partijen en publieke organisaties die vanuit hun rol vooral gebruikers of aanbieders zijn van diensten voor de veilige inzet van AI. Hieronder vallen ook publieke uitvoeringsorganisaties die diensten gebruiken om AI-veiliger te maken en geen beleidsgerichte rol vervullen ten aanzien van AI of cybersecurity.
(3) kennisinstellingen die gericht zijn op onderzoek en innovatie, zoals universiteiten en AI laboratoria. Hierbij is onderscheid gemaakt tussen partijen die werken vanuit de invalshoek AI of vanuit cybersecurity.

Zie Tabel 2 voor de verdeling van de interviews per doelgroep.

Interviewopzet
De interviews hadden een duur van 45 minuten en zijn online uitgevoerd. De gesprekken volgden een vaste structuur: (i) algemene openingsvragen over AI gebruik en AI risicobeheersing, (ii) een gestructureerde discussie over de AISI subtaken, en (iii) afsluitende vragen.

Bij de algemene openingsvragen is samen met de respondent vastgesteld welke rol de betreffende geïnterviewde organisatie speelt binnen het AI-veiligheidsdomein.

Deze input is gebruikt om verdere vraagstelling en verdieping in het gesprek goed te laten aansluiten bij de rol van de organisatie, maar ook om de uiteindelijke categorisering van de doelgroepen te maken, zoals weergegeven in Tabel 2.

Om de interviews beter te kunnen structureren en efficiënt te kunnen afnemen is een AISI takenkader uitgewerkt, voor onderdeel (ii) van de interview opzet. Het takenkader is een uitbreiding van de kerntaken (1) onderzoek, 2) monitoren, testen en evalueren en 3) richting geven, zoals die in onderzoekspoor 2 zijn gedefinieerd [5]. Deze focus gebieden zijn aangepast tot acht specifieker geformuleerde subtaken, die zijn weergegeven in Tabel 3.

Geïnterviewden is gevraagd het belang van deze acht taken voor Nederland te beoordelen op een Likert-schaal van 1 (helemaal niet belangrijk) tot 5 (heel belangrijk). De resultaten zijn geaggregeerd en vergeleken tussen de verschillende doelgroepen.

De overige resultaten zijn kwalitatief beoordeeld.

Doelgroep	Rol	Aantal interviews
Publiek (beleid, strategie en toezicht)	Beleid en strategie	4
	Advies en toezicht	4
Privaat en publiek (uitvoerend)	Veilige AI gebruiker	4
	Veilige AI aanbieder	5
Kennisinstelling & internationale AI-veiligheid organisaties	Cybersecurity	1
	AI	7

Tabel 2: Aantal interviews per doelgroep

AISI kerntaak	AISI subtaken	Publiek (beleid, strategie en toezicht) (N=8)	Privaat en publiek (uitvoerend) N=9	Kennisinstellingen & internationale AI-veiligheid organisaties N=8
Onderzoek en innovatie Nieuwe kennis, methoden, tools en datasets ontwikkelen voor AI-veiligheid	1. Duiden, verbeteren en ontwikkelen van AI-veiligheids- en evaluatietools	3,2	3,2	4,7
	2. Stimuleren van binnenlandse innovatie met veilige AI	2,5	1,8*	4,1*
Monitoren, testen en evalueren Testen, evalueren en monitoren van relevante AI ontwikkelingen en risico's, inclusief periodieke risicoduiding.	3. Monitoren en identificeren van nieuwe capabilities, ontwikkelingen en risico's o.g.v. AI technologie	3,8	3,5*	3,9
	4. Testen en evalueren van AI-systemen op risico's, cyberveiligheid en robuustheid	3,3	3,2	3,5
	5. Beoordelen van conformiteit aan AI en/of Cybersecurity regelgeving	1,8	2,9*	1,7*
Richting geven Richting geven via gezaghebbende adviezen aan overheid, sectoren en het bredere (internationale) ecosysteem.	6. Publiceren en promoten van AI-veiligheidsrichtlijnen, standaarden en evaluatietools	3,1	3,7	3,0**
	7. Informeren en adviseren van beleidsmakers en publiek over risico's en veilig gebruik van AI	4,1	3,3	4,4*
	8. Internationale belangenbehartiging op het gebied van AI-veiligheid	2,6	2,3	3,6

Totaal niet belangrijk (1 - 1,4)	Niet belangrijk (1,5 - 2,4)	Neutraal (2,5 – 3,4)	Belangrijk (3,5 – 4,4)	Heel belangrijk 4,5 – 5)

Tabel 3: gemiddeld belang van de negen AISI taken voor Nederland volgens een 25-tal geïnterviewden vanuit drie verschillende doelgroepen. * 1 ontbrekende score, ** 2 ontbrekende scores

Analyse van resultaten
Alle drie de doelgroepen erkennen het belang van een NL-AISI, maar er zit verschil in welke taken men het belangrijkst vindt.

De *publieke organisaties* met een verantwoordelijkheid op het gebied van Nederlands AI-veiligheid beleid, strategie en/of toezicht, vinden vooral belangrijk dat de politieke en ambtelijke leiding van Nederland wordt voorzien van hoge kwaliteit informatie en adviezen op het gebied van AI en de bijkomende risico's, wat tot uiting komt in een gemiddeld score van 4,1 op Taak 7 *informeren en adviseren van beleidsmakers en publiek over de risico's, en het veilig gebruik, van AI.* Ook wordt het monitoren van nieuwe capabilities, risico's en ontwikkelingen op het gebied van de meest recente AI als randvoorwaardelijk en belangrijk geacht met een gemiddelde score van 3,8. Zo stelde één van de respondenten dat een NL-AISI het scenario denken binnen Nederland kan stimuleren om zo de toekomstige impact van AI op de maatschappij scherper te maken, waarop beleid en strategie bij overheid en private sector beter onderbouwd kan worden.

De *private en uitvoerende publieke organisaties*, ofwel de organisaties die AI-veilig in de praktijk moeten toepassen, zijn primair geïnteresseerd in concrete en praktisch toepasbare richtlijnen, standaarden en technologie met een gemiddelde score van 3,7 op Taak 5 *publiceren en promoten van AI-veiligheid richtlijnen, standaarden en evaluatietools.* Dit kan economische groei verder aanjagen door drempels op het gebied van veilige toepassing van AI te verlagen. Ook deze doelgroep erkent het belang van een gemeenschappelijke capaciteit om trends en risico's te monitoren, met een gemiddelde score van 3,5 op Taak 3 *het monitoren van nieuwe capabilities, risico's en ontwikkelingen.* Men erkent namelijk dat geen enkele organisatie de razendsnelle AI ontwikkelingen helemaal kan bijhouden en doorgronden. Daarentegen is in sommige interviews ook aangegeven dat het gezamenlijk monitoren van risico's wel wenselijk is voor private partijen, maar dat de kansen die AI biedt echt een kwestie van eigen verantwoordelijkheid is in een competitieve markt.

Een aantal van de belanghebbenden in de eerst genoemde doelgroep, publieke organisaties werkzaam op het gebied van beleid en strategie, heeft aangegeven dat een NL-AISI zonder de betrokkenheid van de uitvoerende doelgroep (privaat en publiek uitvoerend) niet wenselijk is.

De *kennisinstellingen en internationale AI-veiligheid organisaties* vinden bijna alle taken belangrijk, met uitzondering van Taak 5 (score van 1,7), waarbij Taak 1 *duiden, verbeteren en ontwikkelen van AI-veiligheid en evaluatietools* er voor hen echt uitspringt met een gemiddelde score van 4,7 op een schaal van 5. Het is een duidelijk signaal dat deze organisaties het belang inzien van een NL-AISI, aangezien deze organisaties zich bovenal bezig houden met de technologie van de toekomst, inclusief de meest geavanceerde vormen van AI. Deze doelgroep vindt het belangrijk dat Nederland anticipeert op de verdere ontwikkeling van geavanceerde AI en voorziet nu al allerlei gevolgen, zoals de impact op de arbeidsmarkt, uitdagingen op het gebied van menselijke controle over een technologie die mogelijk "slimmer" wordt dan de mens, en de mogelijkheden tot misbruik van deze technologie.

Opvallend is dat het beoordelen van conformiteit aan wetgeving, ofwel enige vorm van toezichthoudende rol, door de meeste ondervraagden, en in alle doelgroepen, wordt gezien als niet wenselijk voor een NL-AISI. De publieke organisaties en kennisinstellingen scoren dit bovendien als allerlaagst, met gemiddelde scores van 1,8 en 1,7, respectievelijk. Bovendien wordt er door enkele respondenten genoemd dat dit afleidt van de primaire taken van een AISI, en dat het tevens de vertrouwensrelatie met andere organisaties waarmee idealiter samengewerkt wordt onder druk kan zetten.

Tenslotte, door meerdere respondenten is aangegeven dat de verschillende taken samenhangend zijn en eigenlijk niet helemaal los van elkaar uitgevoerd kunnen worden. Bijvoorbeeld, zonder onderzoek en innovatie is het lastiger om de ontwikkelingen van AI technologie echt te doorgronden en te duiden, wat het lastiger maakt om vervolgens gericht te monitoren en te testen, wat vervolgens weer noodzakelijk is om gegrond advies te geven aan politieke en private beleidsmakers en bestuurders.

Belangrijke statistieken

Informeren en adviseren van beleidsmakers wordt vaak genoemd als gewenste primaire taak van NL-AISI- 48% van alle respondenten scoort Taak 7 als Heel Belangrijk (score = 5).

Monitoren van nieuwe capabilities en risico's wordt vaak genoemd als primaire taak maar ook als randvoorwaardelijk om Taak 7 goed uit te voeren - 40% van alle respondent scoort Taak 3 als Heel Belangrijk (score = 5).

Beoordelen van conformiteit aan regelgeving wordt gezien als niet wenselijk voor een NL-AISI - 52% van alle respondenten scoort Taak 5 als Totaal Niet Belangrijk (score = 1).

Ontbrekende taken in het takenkader
Tijdens de interviews is ook gevraagd of de respondenten nog belangrijke taken missen in het takenkader voor een NL AISI.

Een concrete suggestie door meerdere respondenten genoemd, is een periodieke (e.g. jaarlijks) publicatie over de *State of AI in Nederland*, wat in principe een concrete invulling is van Taak 3 en 7 samen. Dit zou de resultaten van een International AI Safety Report [5] en het Frontier AI Trends Report [15] als inspiratie kunnen nemen aangevuld met voor Nederland specifieke ontwikkelingen en risico's.

Verder is in een interview genoemd dat een NL AISI een geschikt middel kan zijn om warme contacten bij Frontier AI labs (e.g. Mistral, OpenAI, Anthropic, Google Deepmind) te onderhouden en daarmee internationaal invloed uit te oefenen, wat ook weer een concretere invulling is van Taak 8. Mogelijk zullen dit soort bedrijven zich zelfs vestigen in Nederland wat economische activiteit verder kan aanjagen. Tevens, kan een NL AISI een vliegwiel functie vervullen voor innovatie coalities, en start-ups/ scale-ups op het gebied AI-veiligheid.

Ook is genoemd dat een NL AISI een agenda zettende rol zou kunnen en moeten nemen om richting te geven aan meer fundamenteel onderzoek, zoals de UK AISI doet met haar Alignment Project [16].

Tenslotte is in enkele interviews is gevraagd in hoeverre NL AISI taken beperkt worden tot publieke analyses en adviezen, of dat de expertise ook ten dienste kan staan aan gerubriceerde analyses en adviezen die bijdragen aan nationale veiligheids-vraagstukken en/of militaire toepassingen. Of en hoe dit het geval kan zijn is niet verder onderzocht.

Kritieke succesfactoren
Bij het afnemen van de interviews is gevraagd naar de kritieke succesfactoren voor een NL AISI. Onderstaand zijn de meest genoemde succesfactoren uiteengezet.

- **Inhoudelijke autoriteit door middel van toptalent**: de inhoudelijke koers van een NL AISI dient te worden gedragen door een gezaghebbend kernorgaan van inspirerende boegbeelden en toptalent, met zichtbaarheid, erkenning en connecties in het internationale AI-veiligheid werkveld. Inhoudelijke boegbeelden zouden kunnen worden aangetrokken van Frontier AI labs (bijvoorbeeld Mistral, Anthropic, Google Deepmind of OpenAI), omdat het waardevol wordt geacht dat een deel van de experts praktische ervaring heeft op het gebied van de ontwikkeling van de meest recente AI.
- **Wendbaarheid**: een NL AISI dient zeer wendbaar, niet bureaucratisch, georganiseerd te worden om het tempo van AI ontwikkelingen bij te kunnen houden. Als voorbeeld voor een wendbare organisatie verwezen meerdere respondenten naar UK AISI, dat op een effectieve manier invulling geeft aan een nationale AISI capaciteit.
- **Effectieve politieke en bestuurlijke verankering**: de richtinggevende inzichten voor beleid en veilige adoptie van AI dienen effectief te landen bij de relevante politieke en ambtelijke bestuurders. Een mogelijkheid is om een NL AISI het mandaat te geven om (ongevraagde) adviezen te agenderen bij de ministerraad en/of Secretarissen-Generaal (SG) overleg. Tevens kan dit geborgd worden in het leiderschap van een NL AISI, door ook een publiek/ambtelijk boegbeeld aan te trekken.
- **Interdepartementale belangen:** een NL AISI zal meerdere ministeries tegelijk moeten bedienen met actuele informatie en beleidsadviezen, voor tenminste: BZK, BZ, EZK, Defensie, en J&V.
- **Economisch belang en betrokkenheid private sectoren:** nauwe betrokkenheid en toewijding van private en uitvoerende publieke organisaties is cruciaal om AI-veiligheidsconcepten in de praktijk te brengen. Drempels voor veilige toepassingen van AI verlagen draagt stimuleert economische groei. Echter, het is ook van belang om AI (veiligheid) talent niet weg te kapen bij deze organisaties. Hiervoor zou een mogelijke liaisonstructuur van part-time bijdrage door experts een oplossing kunnen zijn. Ook is een optie om naast een inhoudelijk en publiek/ambtelijk boegbeeld ook een privaat boegbeeld met een goed netwerk in de private sector aan te trekken.
- **Aansluiting op internationale netwerk:** een NL AISI dient te fungeren als brug tussen nationale belanghebbenden en het internationale AI Safety & Security werkveld, waarbij een NL AISI idealiter onderdeel wordt van het International Network of AISIs, en nauwe relaties opbouwt bij de aanbieders van toonaangevende AI diensten (zogenaamde Frontier AI labs). Een NL AISI zou kunnen bijdragen aan het behartigen van het Nederlandse belang in de wereldwijde discussie omtrent AI Governance en militaire AI toepassingen.
- **Complementair aan bestaande autoriteiten:** nationale opbouw van AI-veiligheid capaciteit is cruciaal, echter een NL AISI zal complementair moeten opereren aan het EU AI Office en andere bestaande autoriteiten zoals toezichthouders. Met het EU AI Office dient tevens ook nauw samengewerkt te worden. Ook dient afgestemd te worden dat NL AISI taken niet overlappen met activiteiten van lopende AI en cybersecurity instanties (zoals bijvoorbeeld AIC4NL, NCSC en HSD; zie ook onderzoeksspoor 4).
- **Focus op specifieke subdomeinen:** het is verstandig om in te zetten op subdomeinen waar Nederland een sterke positie op kan verkrijgen. Dit kan helpen bij het complementair opereren aan bestaande organisaties, maar kan ook helpen om zowel nationaal als in het internationale AI-veiligheid werkveld inhoudelijke autoriteit op te bouwen en uit te stralen. De AI-veiligheid subdomeinen waar Nederland leidend in kan zijn dient nader onderzocht te worden, maar gedacht kan worden aan focus op het gebied van het evalueren van AI modellen op het gebied van CBRN risico's, het evalueren van AI modellen op het gebied van offensieve cyber capabilities, of de ontwikkeling van veilige hardware/compute binnen de AI supply chain.

- **Focus op general-purpose AI:** de meeste respondenten zijn van mening dat een NL AISI zich primair moet focussen op de laatste trends en op dat moment meest geavanceerde vormen van AI.
- **Financiering:** meerdere respondenten benoemen dat een NL AISI een wereld leidende rol kan pakken in AI-veiligheid. De realisatie van die ambitie vergt wel de benodigde middelen, waaronder financiering. In een interview werd een jaarbudget genoemd van € miljoen. Die budgetinschatting ligt in lijn met funding van bestaande AISI's in andere landen (zie ook onderzoeksspoor 2).
- **Toegang tot rekenkracht:** ten aanzien van benodigde middelen is ook genoemd dat NL AISI prioriteit dient te krijgen bij toegang tot bestaande rekenkracht (compute) faciliteiten, om technische evaluaties uit te kunnen voeren. Tevens kan een NL AISI bijdragen met kennis en expertise aan een nationale strategie op het gebied van AI rekenkracht (compute).

Conclusies

Door een 25 tal interviews af te nemen is duidelijk geworden dat er onder de respondenten, bestaande uit publieke organisaties, private organisaties en (internationale) kennisinstellingen, een duidelijke behoefte is aan een NL AISI. Tevens is er veel feedback opgehaald over de gewenste taakstelling en overige kaders welke hiernaast in samenvatting zijn weergegeven.

Als kritieke succesfactoren wordt aangegeven dat de NL-AISI mee moet kunnen bewegen in de sterke dynamiek van AI-veiligheid. Haar autoriteit dient het te ontlenen aan haar deskundigheid en inspirerende boegbeelden, waarmee draagvlak wordt verkregen bij publieke, private en kennisorganisaties. Om een leidende rol te kunnen vervullen dient focus gelegd te worden op sub-domeinen, moeten voldoende middelen gealloceerd worden en dient actief samen gewerkt te worden in het internationale AISI netwerk en met toonaangevende AI Labs.

Taakstelling van een NL AISI

De NL AISI dient alle drie de kerntaken van een AISI uit te voeren maar legt daarbij wel de juiste nuances en nadruk.

(i) Met onderzoek en innovatie nieuwe kennis, methoden, tools en datasets ontwikkelen voor AI-veiligheid, met nadrukkelijke focus op geavanceerde AI, en specifieke AI-veiligheid subdomeinen (nader te bepalen; e.g. veilige AI compute/hardware, misbruik voor offensieve cyber operaties, misbruik voor CBRN doeleinden, maar ook toepassing van AI voor defensieve doeleinden).

(ii) Monitoren, testen en evalueren van AI ontwikkelingen en risico's specifiek relevant voor Nederland (o.a. nationale duiding van International AI Safety Report [5]), inclusief periodieke risicoduiding, voor ten minste de ministeries BZK, BZ, J&V, Defensie en EZK en geïnteresseerde private organisaties. Een NL AISI dient zich juist niet te richten op handhaving en toezicht van wetgeving.

(iii) Richting geven via gezaghebbende adviezen aan overheid (primair aan politiek en ambtelijk leiderschap, e.g. ministerraad en/of SG overleggen) en het bredere (nationale en internationale) publiek, maar ook door middel van praktische richtlijnen, standaarden en tools aan de private sector en uitvoerende publieke instanties.

Onderzoekspoor 4: AI-veiligheid initiatieven in Nederland

Doel van onderzoekspoor 4
De geconstateerde behoefte aan een NL AISI roept de vraag op in hoeverre de beschreven taken reeds opgepakt worden in bestaande, nationale cybersecurity een AI initiatieven. Hiertoe is een beknopte inventarisatie uitgevoerd van relevante initiatieven door Nederlandse organisaties.

Inventarisatie-opzet
De inventarisatie van initiatieven is gefocust op organisaties met een publieke taak of stelselverantwoordelijkheid op het gebied van AI en cybersecurity, die ook voldoende structurele capaciteit hebben om verschillende AISI-taken duurzaam uit te voeren. In totaal zijn websites en publicaties van 23 relevante, Nederlandse organisaties/ samenwerkingsverbanden meegenomen.*

In deze inventarisatie zijn gepubliceerde uitingen van die organisaties over hun missie, taakstelling en organisatorische invulling kwalitatief vergeleken met de AISI subtaken opgenomen in Tabel 3. De publieke uitingen bestaan uit website-informatie, jaarverslagen en relevante rapportages. Naast een vergelijking met AISI subtaken is ook bepaald of de missie van de organisatie primair gericht is op (a) cybersecurity, (b) AI of (c) AI-veiligheid. Ten derde is bepaald of er in de publicaties verwezen wordt naar specifieke typen AI (zoals general-purpose AI).

Deze drie aspecten (zie Figuur 5) vormen samen de basis van de inventarisatie, die in kaart brengt welke functies in Nederland al belegd zijn en waar structurele capaciteit mist om handelingsperspectief voor het AI-veiligheidsvraagstuk te bieden.

Taakstelling
Welke onderdelen van de algemene AISI-taakstelling de organisatie invult

Missiedomein
Het inhoudelijke focusgebied van de organisatie, uitgesplitst naar cybersecurity, AI of AI-veiligheid

Primaire focus op type AI
Mate waarin het focusgebied van de organisatie specifiek gericht is op een bepaald type AI.

Figuur 5: Aspecten van de inventarisatie

Observaties
In de inventarisatie is er geen organisatie aangetroffen die vanuit haar huidige rol invulling geeft aan het geheel van de AISI subtaken. Sterker, er is geen organisatie die als missie de spanwijdte omvat van de drie overkoepelende AISI-kerntaken:
i. Onderzoek & innovatie,
ii. Monitoren, testen en evalueren; en
iii. Richting geven (adviseren en promoten).

Hoewel individuele organisaties één of meerdere van deze functies vervullen, worden zij niet in samenhang opgepakt of structureel met elkaar verbonden. Hierdoor ontbreekt een duidelijke plek waar technische inzichten, monitoring van ontwikkelingen en gezaghebbende advisering samenkomen en worden vertaald naar handelingsperspectief voor beleidsmakers en andere belanghebbenden.

Uit de inventarisatie komt ook naar voren dat cybersecurity en AI initiatieven in Nederland grotendeels gescheiden worden ingericht. Activiteiten vinden voornamelijk plaats binnen afzonderlijke gebieden van cybersecurity of AI, terwijl het snijvlak van beide slechts beperkt structureel is ingevuld. Dit betekent dat kansen en bedreigingen die voortkomen uit de wisselwerking tussen cybersecurity en AI (zoals geschetst in de scenarioanalyse) beleidsmatig en organisatorisch niet vanzelf samenkomen en geadresseerd worden.

Tenslotte valt op dat in Nederlandse initiatieven weinig expliciete nadruk ligt op general-purpose AI. Hoewel meerdere organisaties activiteiten ontplooien op het gebied van AI, omvat deze categorie in de praktijk veel werk op het gebied van narrow-AI en taakgebonden systemen. Daarmee vormt general-purpose AI een deelonderwerp binnen bredere AI- en digitaliseringsagenda's. Dit is relevant, omdat juist bij general-purpose AI-systemen momenteel de grootste technologische versnellingen plaatsvinden en de potentiële risico's sector- en domein overstijgend zijn. Het ontbreken van een expliciete focus op general-purpose AI bemoeilijkt daarmee een samenhangende aanpak van de systemische risico's die uit deze ontwikkelingen voortkomen, en beperkt de mogelijkheid om deze risico's structureel te monitoren, technisch te evalueren en beleidsmatig te duiden.

Conclusies
Uit de inventarisatie volgt dat er diverse Nederlandse cybersecurity- en AI-initiatieven ondernomen worden, maar dat deze slechts beperkt aansluiten op de samenhangende AISI-taakstelling zoals uitgewerkt in de voorgaande onderzoeksporen. Er is geen organisatie geïdentificeerd die de verschillende, gerelateerde AISI-taken vervult of coördineert. Hiermee ontbreekt het in Nederland aan slagkracht om tot handelingsperspectief voor AI-veiligheid te komen.

** Voetnoot: de inventarisatie beoogt NIET de vaststelling van een uitputtende lijst van relevante AI-veiligheid organisaties. In deze context, en doordat de publieke uitingen niet bij de organisaties zijn geverifieerd, worden de resultaten hier anoniem gepresenteerd. Voor specifiekere details over deze inventarisatie kan contact opgenomen worden met de auteurs van dit rapport.*

Conclusies en Aanbevelingen

Conclusies
Dit rapport is opgesteld in opdracht van het Ministerie van Economische Zaken en Klimaat, vanuit de ambitie om de beleidsdomeinen van AI en Cybersecurity in Nederland beter te verbinden, zodat publieke belangen rond veiligheid, soevereiniteit en economische groei beter geborgd kunnen worden. Hierbij staan de volgende vragen centraal:

- Welke kansen en dreigingen in het raakvlak van cybersecurity en AI zijn het meest relevant t.a.v. verdienvermogen en de publieke belangen?
- Hoe kunnen cybersecurity en AI ontwikkelingen en hun invloed op verdienvermogen en publieke belangen worden gemonitord?
- Welke beleidsmatige interventies zijn meest geschikt voor het balanceren van elk van die belangen?

Het inzichtelijk maken van de ***kansen en dreigingen binnen het raakvlak van cybersecurity en AI*** is onderhevig aan het evidence dilemma, dat wil zeggen dat nieuwe ontwikkelingen zich zo snel manifesteren dat beleid niet kan leunen op volledige informatie. Daarom dienen de kansen en dreigingen op dit raakvlak continue gemonitord te worden. Een NL AISI lijkt een geschikt instrument daarvoor, welke in een aantal andere landen met succes zijn opgericht.

Voor het effectief ***monitoren van de ontwikkelingen*** binnen het raakvlak van cybersecurity en AI is een NL AISI taakstelling uitgewerkt, en er is opgehaald door middel van een marktconsultatie welke taken het meest wenselijk zijn volgens verschillende belanghebbenden. De belanghebbenden bestonden uit drie doelgroepen, namelijk publieke organisaties, private organisaties en (internationale) kennisinstellingen.

De oprichting van een NL AISI kan gezien worden als een ***duurzaam beleidsmatig initiatief*** om de verschillende belangen, zowel op het gebied van beheersing van de risico's binnen het raakvlak van cybersecurity en AI, als de kansen voor het verdienvermogen, te balanceren.

Aanbevelingen
Uit dit onderzoekt blijkt dat er bij publieke, private organisaties en kennisinstellingen veel belang wordt gehecht aan de taken van een nationale AISI, en dat deze taakstelling vooralsnog onvolledig en versnippert wordt opgepakt binnen Nederland. In dit onderzoek zijn duidelijke kaders opgehaald voor de verdere inrichting van een NL AISI, met onder andere een aantal kritieke succesfactoren, hiernaast in samenvatting weergegeven.

Deze opgehaalde kaders kunnen worden gebruikt om verdere invulling van een NL AISI te verkennen. Vooral de organisatorische inrichting van een NL AISI behoeft verdere verdieping.

De opgehaalde kaders voor een NL AISI

Doelstelling

NL AISI dient in haar missie drie strategische belangen te dienen, namelijk (i) economische groei door drempels voor veilige toepassing van AI te verlagen, (ii) nationale veiligheid door risico's en misbruik van AI tijdig te duiden, en (iii) strategische soevereiniteit door een sterke Nederlandse kennis- en marktpositie op te bouwen waarmee internationaal aangesloten kan worden, en vooroplopen binnen subdomeinen binnen AI-veiligheid werkveld.

Taakstelling

De NL AISI dient alle drie de kerntaken van een AISI uit te voeren maar legt daarbij wel de juiste nuances en nadruk.

i. Met onderzoek en innovatie nieuwe kennis, methoden, tools en datasets ontwikkelen voor AI-veiligheid, met nadrukkelijke focus op geavanceerde AI, en specifieke AI-veiligheid subdomeinen (nader te bepalen; e.g. veilige AI compute/hardware, misbruik voor offensieve cyber operaties, misbruik voor CBRN doeleinden, maar ook toepassing van AI voor cyberdefensieve doeleinden).
ii. Monitoren, testen en evalueren van AI ontwikkelingen en risico's specifiek relevant voor Nederland (o.a. nationale duiding van International AI Safety Report [5]), inclusief periodieke risicoduiding, voor ten minste de ministeries BZK, BZ, J&V, Defensie en EZK en geïnteresseerde private organisaties. Een NL AISI dient zich juist niet te richten op handhaving en toezicht van wetgeving.
iii. Richting geven via gezaghebbende adviezen aan overheid (primair aan politiek en ambtelijk leiderschap, e.g. ministerraad en/of SG overleggen) en het bredere (nationale en internationale) publiek, maar ook door middel van praktische richtlijnen, standaarden en tools aan de private sector en uitvoerende publieke instanties.

Organisatiekaders

De NL AISI, met een jaar begroting van tientallen miljoen euro's, is een zeer wendbare organisatie, die mee kan bewegen met de snelheid en dynamiek van het internationale werkveld op het gebied van AI. De organisatie is een autoriteit op het gebied van AI-veiligheid, en wordt geleid door meerdere inspirerende boegbeelden (inhoudelijk, publiek/ambtelijk, en privaat). Er wordt nauw samengewerkt met de EU AI Office, binnen het International Network of AISIs, maar ook met de Frontier AI labs. Verschillende toonaangevende Nederlandse publieke en private organisaties (uit meerdere sectoren) werken samen met NL AISI om bijvoorbeeld een jaarlijkse State of AI in Nederland rapport te publiceren.

Referenties

Nummer	Referentie
[1]	“De Route Naar Toekomstige Welvaart”, “Rapport Wennink”, December 2025, https://www.rapportwennink.nl/downloads/rapport_wennink_12december2025.pdf
[2]	“The Future of European Competitiveness, “Draghi Rapport”, September 2024, https://commission.europa.eu/topics/competitiveness/draghi-report_en
[3]	Coalitie Akkoord 2026 - 2030: Aan de Slag, Januari 2026, https://www.kabinetsformatie2025.nl/documenten/2026/01/30/aan-de-slag---coalitieakkoord-2026-2030
[4]	“Verkenning van het raakvlak van cybersecurity en AI”, TNO, rapport TNO-2024-R10768, oktober 2024, https://publications.tno.nl/publication/34643247/fRGFjaaX/TNO-2024-R10768.pdf
[5]	“International AI Safety Report”, website gepubliceerd door een internationaal expert adviespanel gefinancieerd door de overheid van het VK, https://internationalaisafetyreport.org/ Meest recente (en in dit rapport gebruikte) versie: https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf
[6]	“De toekomst van cyberaanvallen met Large Language Models.”, December 2023, NCSC en TNO, https://www.ncsc.nl/artificial-intelligence/de-toekomst-van-cyberaanvallen-met-large-language-models
[7]	“Opgave AI. De nieuwe systeemtechnologie”, WRR rapport nr. 105, November 2021, https://www.wrr.nl/documenten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie
[8]	“The Global Impact of AI: Mind the Gap”, WP/25/76, April 2025, https://www.imf.org/-/media/files/publications/wp/2025/english/wpiea2025076-print-pdf.pdf
[9]	“National Security Strategy of the United States of America”, November 2025, https://www.whitehouse.gov/wp-content/uploads/2025/12/2025-National-Security-Strategy.pdf
[10]	“Disrupting the first reported AI-orchestrated cyber espionage campaign”, Anthropic, November 2025, https://www.anthropic.com/news/disrupting-AI-espionage
[11]	“Nationaal AI Deltaplan”, November 2025, https://aiplan.nl/deltaplan
[12]	“The AI Safety Institute International Network: Next Steps and Recommendations”, Center for Strategic & International Studies, oktober 2024, https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations
[13]	“La France se dote d’un Institut national pour l’évaluation et la sécurité de l’intelligence artificielle (INESIA)”, Februari 2025, economie.gouv.fr
[14]	“Establishment of Australian AI Safety Institute, November 2025, https://www.minister.industry.gov.au/ministers/timayres/media-releases/establishment-australian-ai-safety-institute
[15]	“Frontier AI Trends Report”, December 2025, UK AISI, https://www.aisi.gov.uk/frontier-ai-trends-report/pdf
[16]	“The Alignment Project”, UK AISI, https://alignmentproject.aisi.gov.uk/research-agenda

Bijlage 1: AI-veiligheid scenario's (in Engels) (1/3)

Deze bijlage bevat korte beschrijvingen van de opgestelde AI-veiligheid scenario's. Een overzicht van de scenario's is opgenomen in figuur 2. Van de met "*" gemarkeerde scenario's zijn uitgebreider scenario uitwerkingen opgesteld; deze zijn opvraagbaar bij de auteurs.

1) Adversarial AI attack on AI-assisted telecom network management
With the advent of more AI-assisted technology in management systems of critical infrastructure, the AI attack surface increases. For example, an adversary supplier may implant a hidden backdoor into AI systems used to manage telecommunication equipment. These AI systems support automated, real-time network management functions, such as zero touch networking (in telecommunication) and AI-powered smart grids (in energy grids). The backdoor could be activated remotely to disable monitoring, reroute traffic, or grant covert access to network resources.

2) Discriminatory data injection in AI-assisted social care services
An attacker targets AI models that are used to inform or direct public services, e.g. prioritizing vulnerable citizens for subsidiary benefits or identifying early signs of fraud. Such attacks reduce the objective effectiveness of the targeted social service(s), but also public perception of its fairness and effectiveness.

3) Evasion attack on autonomous system's sensors
An attacker manipulates the input to an autonomous vehicle's perception system (e.g., through visual adversarial patterns, LiDAR spoofing, or signal interference), causing the system to misinterpret its surroundings. These evasion attacks allow the attacker to bypass object detection or traffic rules enforcement, potentially leading to collisions, mis-navigation, or denial-of-service of critical transport functions. Such attacks exploit weaknesses in AI models' sensor processing and decision-making, often without triggering fail-safes.

4) Inversion attack on AI service trained on highly sensitive data
An inversion attack targets AI models trained on sensitive datasets—such as proprietary research data, trade secrets, or confidential experimental results. By querying the AI service (e.g., a public API for scientific modeling or innovation support), an attacker attempts to reconstruct or infer original training data. These attacks can compromise intellectual property, undermine competitive advantage, and expose confidential information embedded in the model, especially when privacy safeguards are weak.

5)* Prompt injection attack on government AI agents
With the breakthrough of commercial LLMs in 2020, AI has become significantly more accessible for both individuals and organizations. So much so, that many large organizations are planning on or already using LLMs for internal or even external applications. As the trust in these models grows, organizations are willing to give them more and more responsibility paving the way for agentic AI or AI agents. AI agents consist of LLMs integrated with tools such that they can interact with other digital systems and perform actions. The idea is that these agents would be able to function autonomously, splitting up tasks and taking decisions based on the input they receive from both humans and the digital environment they have access to. The potential use cases for AI agents seem endless and are promising to significantly increase productivity and reduce costs.

5-Generic) Prompt injection attack on professional AI agents
Recently an increase is seen in the usage of AI agents that autonomously perform business tasks. In most cases such AI agents interact with publicly available LLM models and cybersecurity issues may be overlooked in the rush of development. For example, AI agent-based wallets developed by FinTech companies and chatbots in for example digital services of the government.

6)* Mass collection of blackmailable information
AI can be abused to analyze behavior and build profiles of people. Agentic AI (AI agents with more autonomy) is reportedly used for selecting victims (including politicians, executives, industry professionals, whistleblowers, or any of their relatives) and search for extortion-able information such as profiles on dating (and cheating) websites, building bonds to win trust and collect more personal information, as well as collecting harmful information from compromised accounts or leaked databases. The malicious objective is not confined to theft, but to steal leverage. Many cases are reported in which AI composed profiles were used for for targeted coercion, reputational pressure, decision manipulation across many sectors, and even sextortion and self-harm. Abusing Agentic AI in this way is easily replicable, scalable and can be done at very limited cost.

Bijlage 1: AI-veiligheid scenario's (in Engels) (2/3)

7)* AI-assisted smart grid blackout
Dutch companies and households have become critically dependent on electrical power services. A sustained power outage can have far stretching consequences. Meanwhile, the challenge to keep the power grid balanced is becoming more complex. For example, the increasing proportion of renewable energy is welcomed from the ecological perspective, but it also creates a production system that fluctuates with changing weather conditions. In addition to the increased complexity of balancing the grid, the Dutch power grid is at the boundary of its capacity and while TSO and DSOs are working hard to expand the power grid it could take years for that to be resolved.

To cope with power grid management, AI-based innovations are being investigated and deployed. Both in central smart grid management as well as decentral home systems. In turn, such AI innovations introduce (new) challenges. Unreliable data, non-robust or unexplainable AI algorithms may create instable grid behavior, and adversarial AI attacks are a complementary cyber security vulnerability. Management of these risks is needed to sustain a stable, national power grid.

7-Generic) AI-enabled attacks to destabilize critical infrastructure (e.g. power grid)
An adversary uses AI agents to target cyber defenses of critical infrastructure. AI agents probe defenses and expose vulnerabilities at far higher rate than human hackers can, resulting in significantly more successful incursions. As soon as vulnerabilities are noticed and addressed by cyber defenses, the adversary's AI agents adjust their attack to find new vulnerabilities.

8)* Advanced AI impersonation and falsification erodes trust in digital counter services
Society has grown used to massive, digital interaction, both for social and for commercial use. Where webshops and online banking were the first commercial examples, and digital counter services increasingly replaced call centers. The recent introduction of LLM chatbots (and in future: AI agents) is further fueling digitalization of interactions between consumers and their service providers.

Many of such digital counter services still lack strong authentication for creating accounts and logging into these, that is sufficient to protect them against GenAI impersonation and falsification attacks. A massive increase in such type of attacks and/or prolonged period of exploitation may undermine the trust required for using digital counter services.

9) Preserving integrity of foundational research data sets and models
As foundational AI datasets in R&D domains such as scientific modeling, environmental forecasting, socio-economic correlations and healthcare innovation become more central, their integrity becomes a (national) asset. In future, the importance of datasets for training AI-based cyber security products will also increase. Introducing biases into these datasets skews the behavior of downstream models and scientific conclusions. This can lead to public policy, safety assessments, or medical advisories in ways that perpetuate inequality or scientific blind spots. The impact may only become visible years later, as skewed knowledge propagates through scientific and technological systems.

10)* Expand AI-assisted, civil protection technology to critical infrastructure protection
In the context of increasing geopolitical tension and increasing investment in Defense capabilities, AI technology may be expanded from civil applications to protection of critical infrastructure. For example, 'operation spider web' has increased the interest of the ministry of Defense and other NATO partners in AI-assisted radar systems that are used for bird detection, and apply them for military purpose. Other examples are distributed acoustic and temperature sensing systems that are used for civil purposes. In the aftermath of the Nord stream incident and cable breaches in the Baltic Sea, this technology is finding expanded interest in applications for surveilling critical infrastructure such as undersea cable, oil and gas pipeline infrastructure and railroad surveillance. Yet another example concerns drone platforms and services that are used for inspection and maintenance of private facilities, widespread high-voltage cables and pipelines. These examples illustrate the encompassing opportunity 'scenario' of the potential for market expansion from civil to critical infrastructure application of AI-assisted surveillance and maintenance systems.

11)* New trusted 3rd party security services for (gen)AI products and services
Regulation including the AI act and the CRA require secure development and assessment of AI products and services throughout their life cycle. Tools, techniques and facilities for AI security verification are still in their infancy and business models for third party assessment have yet to be developed. Also, in recent initiatives launched to create European AI factories there is a demand for cyber security expertise. These developments create a new field of economic activity for trusted security service providers.

Bijlage 1: AI-veiligheid scenario's (in Engels) (3/3)

12)* AI-automated software development – pushing human developers out of the job market

The quality of software generation capabilities of GenAI have been increasing significantly in the past years. It is claimed that ~30% of the Google's software was AI generated, for now enabling Google to ship new software products and new features in existing products at a relentless pace according to their own words. However, this may also result in a decline of software developer roles, particularly junior roles. BBC refers to a study on a % decline in tech job adverts between 2019 and 2025, with the rise of AI claimed to be a significant contributing factor. At this moment in time, GenAI is not good enough to replace software developer altogether, but it is definitely capable enough to produce small pieces of software within a larger human constructed code base, still with human verification of software quality in the software development lifecycle. Also, adversaries are using GenAI for producing malware more and more.

GenAI is not yet capable enough to fully replace humans in the software development lifecycle, however it does have human-competitive capabilities that are good to take into consideration, namely 1) the speed at which new software can be generated and 2) the fact that these models can work 24/7 (without paychecks and employee rights) are qualities that humans can never compete with. These capabilities manifest especially when GenAI is implemented in an agentic fashion.

This scenario offers both opportunities and challenges for Dutch software development industry, and more specifically also the cyber security sector.