Verkenning van het raakvlak van cybersecurity en AI

TNO Publiek ⟩ TNO 2024 R10768
30 september 2024

ICT, Strategy & Policy
www.tno.nl
+31 88 866 00 00
info@tno.nl

TNO 2024 R10768 – 30 september 2024

Verkenning van het raakvlak van cybersecurity en AI

Auteurs: Niels Brink, Bart Gijsen, Yori Kamphuis, Nina van Liebergen, Daan Opheikens, Jip van Stijn, Stefan Wijnja

Rubricering rapport: TNO Publiek

Titel: TNO Publiek

Rapporttekst: TNO Publiek

Aantal pagina's: 59 (excl. voor- en achterblad)
Aantal bijlagen: 0
Opdrachtgever: Ministerie EZ, DG Economie en Digitalisering - Digitale Economie

Projectnaam: Cybersecurity & AI
Projectnummer: 060.59671

Onze partners

Samenvatting

Kunstmatige intelligentie (of Artificial Intelligence: AI) is een snel ontwikkelende technologie die veel nieuwe kansen biedt. In de context van de digitale transitie speelt cybersecurity een belangrijke rol in het waarborgen van de continuïteit van digitale dienstverlening, maatschappelijke waarden en de nationale veiligheid. Op het raakvlak van cybersecurity en AI (security & AI) ontstaat behoefte aan strategisch beleid waarvoor een goede, actuele en toekomst-georiënteerde informatiebasis essentieel is.

Probleemstelling

Het doel van dit rapport is om beleidsmakers te voorzien van deze informatiebasis. TNO heeft onderzocht hoe het raakvlak van cybersecurity en AI eruitziet, waarbij de volgende vragen zijn beantwoord. Welke security & AI deelgebieden zijn te onderscheiden? Welke voorbeelden zijn illustratief voor die deelgebieden? Wat is de stand van zaken in de security & AI deelgebieden, wie zijn de relevante actoren en welke rol spelen zij? Welke invloed op publieke belangen is reeds te zien in de security & AI deelgebieden, en wat is de verwachting gezien de ontwikkelingen?

Beschrijving van de werkzaamheden

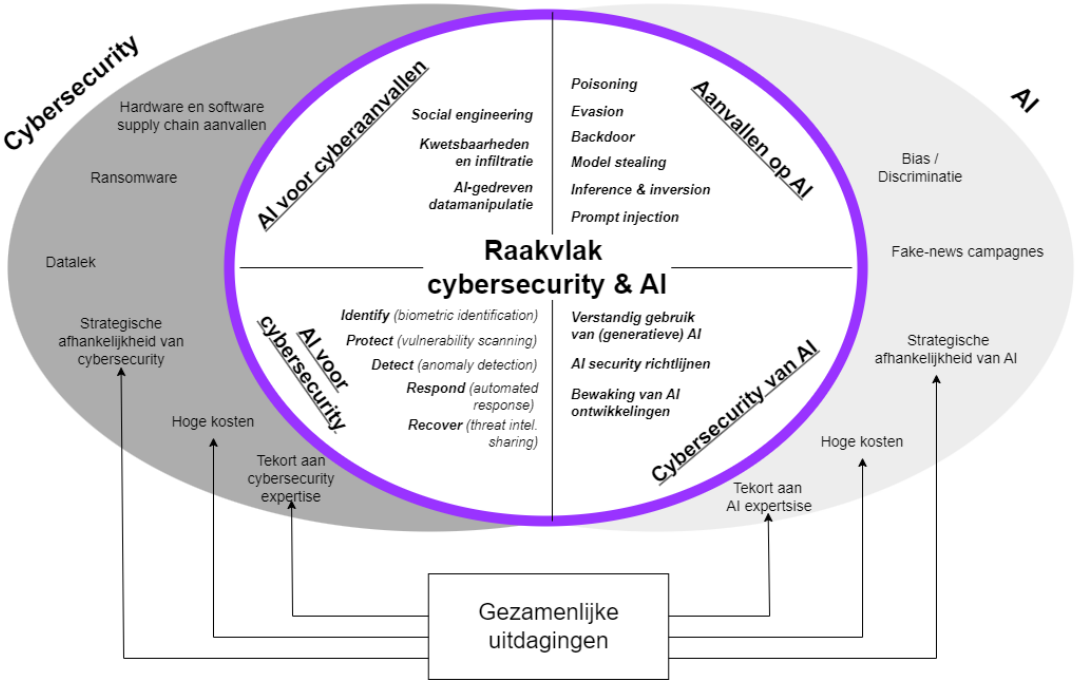
Dit rapport is tot stand gekomen door analyses van literatuur over AI, cybersecurity en het raakvlak van cybersecurity en AI. Dit omvat standaarden, taxonomieën, strategie- en beleidsdocumenten, wetenschappelijke artikelen en andere publicaties. Daarnaast zijn gesprekken met diverse stakeholders gevoerd. De verkregen kennis is verwerkt in een uiteenzetting van zowel AI als cybersecurity, met speciale aandacht voor security & AI.

Resultaten en conclusies

Zowel het expertisegebied cybersecurity als AI zijn sterk in beweging en kennen een breed scala aan toepassingen en uitdagingen. Deels hebben die gebieden hun eigen problematiek, maar in dit onderzoek is voornamelijk gefocust op de security & AI toepassingen en uitdagingen die een significante onderlinge wisselwerking hebben.

De gebieden van cybersecurity en AI en hun raakvlak zijn visueel inzichtelijk gemaakt in het venndiagram van Figuur 1. Binnen dit raakvlak zijn vier security & AI kwadranten te zien. Deze deelgebieden zijn: AI voor cybersecurity en AI voor cyberaanvallen, die beiden aan de linkerkant van het kwadrant staan, en Cybersecurity van AI en Aanvallen op AI, die beiden aan de rechterkant van het kwadrant te vinden zijn. Per kwadrant staan in de figuur enkele voorbeelden ter illustratie genoemd.

Naast de uitdagingen binnen het raakvlak spelen er ook enkele gezamenlijke uitdagingen die die minder onderlinge wisselwerking hebben. Hieronder vallen de afhankelijkheid van Amerikaanse aanbieders, schaarste aan expertise en hoge kosten.



Figuur 1: Overzicht van het raakvlak van cybersecurity en AI. Werk van de auteurs.

Van de relevante actoren kunnen overheden beleid en regelgeving gebruiken om zowel cybersecurity- als (veilige) AI-ontwikkelingen te stimuleren. De drie grootste markten zijn China, de VS en Europa, die ieder net een verschillende aanpak volgen. Daarnaast zijn er, veelal Amerikaanse, private aanbieders van (cyber)AI-producten, en dienstaanbieders, waarin ook sterke Nederlandse partijen zich bevinden. De derde groep relevante actoren is de chipindustrie, waarin Nederlandse bedrijven internationaal gezien een cruciale rol spelen.

Toepasbaarheid

Dit rapport biedt een basis om toe te werken naar een raamwerk waarmee (a) security & AI vraagstukken gedefinieerd en gecategoriseerd kunnen worden, (b) de met de vraagstukken gepaard gaande impact vanuit verschillende invalshoeken ingeschat kan worden, waarmee vervolgens (c) de relatie naar handelingsperspectief gelegd kan worden, inclusief innovatie- en kennisprogrammering.

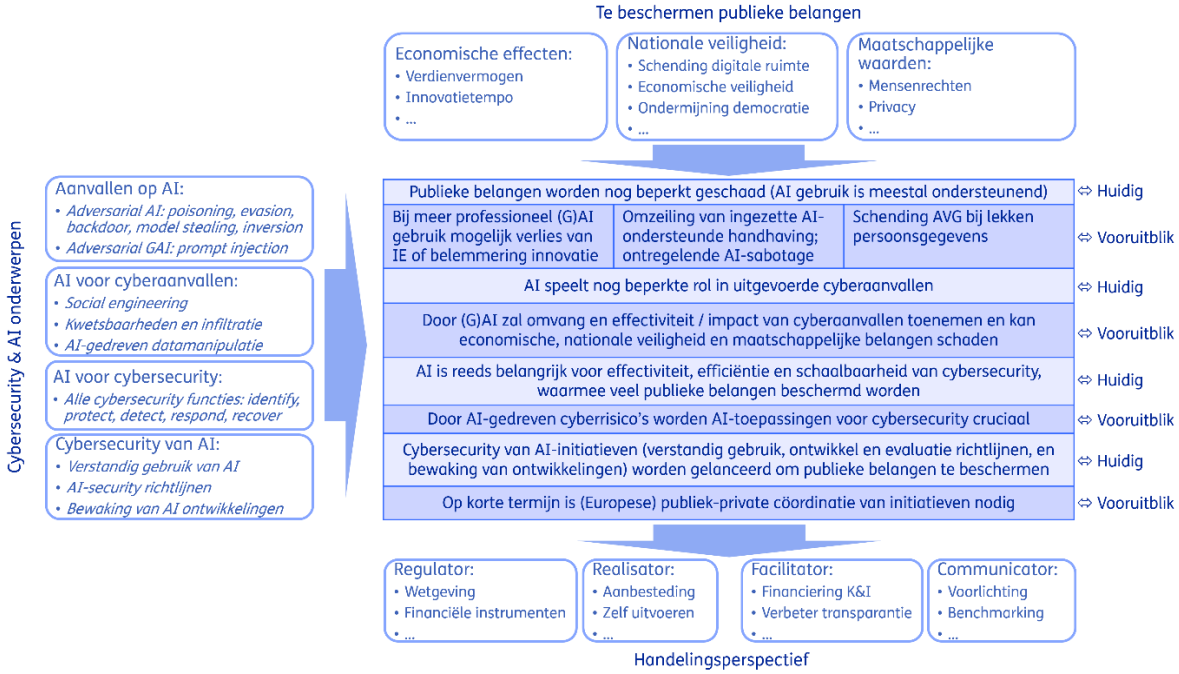
Door de snelle ontwikkelingen rondom AI-toepassingen is een op feiten-gebaseerde inschatting van het publieke belang een beperkt houdbare momentopname, en vraagt deze snelheid tegelijkertijd om een proactieve aanpak. Daarom is naast deze op feiten-gebaseerde inschatting ook een extrapolatie gemaakt naar voorstelbare impact in de nabije toekomst. Grofweg kunnen de te beschermen publieke belangen onderverdeeld worden in economische effecten, nationale veiligheid en maatschappelijke waarden. De kwalitatieve inschatting van de te beschermen publieke belangen is te vinden in Figuur 2. Hierbij is per onderdeel de inschatting voor de huidige situatie weergegeven (in licht blauw), gevolgd door de vooruitblik naar de nabije toekomst (donkerblauw).

De stand van zaken en ontwikkelingen verschillen sterk per security & AI deelonderwerp. Een duidelijk publiek belang is vastgesteld voor het deelonderwerp AI voor cybersecurity; de publieke belangen voor de andere deelonderwerpen worden vooralsnog minder geraakt. De inschatting is dat dit echter zeer snel kan veranderen (zeker zodra professioneel gebruik van

AI toeneemt). Gezien de verschillen tussen de deelonderwerpen ligt het niet voor de hand om beleid op te stellen dat gericht is op het raakvlak van cybersecurity en AI als geheel. Beter lijkt om de beleidsbehoefte voor security & AI onder te brengen in reeds opgestarte, nationale en internationale cybersecurity- en AI-beleidsinitiatieven. Meerdere van deze initiatieven en hun relatie naar het raakvlak is in kaart gebracht. De eerste vervolgstappen voor security & AI beleid betreffen:

a) het uitwerken van de beleidsbehoefte door kwantificering van publieke belangen,
b) het inpassen daarvan in opgestarte beleidsinitiatieven en
c) indien nodig aanvullende beleidsinitiatieven opstarten.

In de beschouwing aan het eind van dit rapport zijn enkele suggesties opgenomen over de wijze waarop deze vervolgstappen kunnen worden uitgevoerd.



Figuur 2: Overzicht van publieke belangen van het raakvlak van cybersecurity en AI. Huidige situatie in lichtblauw, gevolgd door vooruitblik nabije toekomst in donkerblauw. Werk van de auteurs.

Inhoudsopgave

1 Inleiding

Kunstmatige intelligentie (of Artificial Intelligence: AI) is een snel ontwikkelende technologie die veel nieuwe economische en maatschappelijke kansen biedt. Om deze kansen te benutten is vertrouwen in het gebruik van AI van groot belang voor burgers, bedrijven en de overheid. Daartoe is de Europese Commissie sinds 2018 gestart met het formuleren van een AI-strategie en het opstellen van regelgeving in de vorm van de AI-act (zie sectie 2.1).

Tegelijkertijd komt het cybersecuritydossier steeds nadrukkelijker op de agenda van beleidsmakers (zie sectie 2.2). In de context van de kansen die de digitale transitie biedt (in de volle breedte, dus inclusief toepassing van AI), speelt cybersecurity een belangrijke rol in het waarborgen van continuïteit van digitale dienstverlening, nationale veiligheid en maatschappelijke waarden. Ook hier zijn Europese en nationale beleidsmakers in actie gekomen door cybersecuritystrategieën en regelgeving op te stellen.

Ook op het raakvlak van deze twee ontwikkelingen (AI en cybersecurity) ontstaat inmiddels behoefte aan strategisch beleid. ENISA heeft hierin een voortrekkersrol gespeeld door het publiceren van enkele rapporten waarmee eerste duiding wordt gegeven van het raakvlak van cybersecurity en AI (hierna aan te duiden als security & AI). Deze rapporten gaan bijvoorbeeld in op de stand van standaardisatie [1] en van onderzoek [2] daarin en het afvaardigen van good practices [3].

De bekendheid en begrijpbaarheid van deze informatie is belangrijk voor relevante beleidsmakers bij de rijksoverheid. Een goede en actuele informatiebasis is essentieel om een breed gedragen visie en beleid op te kunnen stellen voor dit zeer dynamische en complexe raakvlak van onderwerpen. Dit rapport beoogt om een gedragen, toekomst-georiënteerde informatiebasis te bieden waarop steekhoudend security & AI beleid gemaakt kan worden. Deze informatiebasis moet beleidsmakers in staat stellen om actuele security & AI vraagstukken te begrijpen en een handreiking te geven voor het maken van impactinschattingen daarover, zodat zij uiteindelijk goed onderbouwd security & AI beleid kunnen formuleren en daar draagvlak voor kunnen creëren. Dit beleid moet vervolgens (onder andere) een fundament bieden voor programmering van kennis en innovatie initiatieven t.a.v. security & AI door de overheid.

1.1 Doel van het onderzoek

Het doel van dit onderzoek is om beleidsmakers binnen de overheid een informatiebasis en handreikingen te bieden ter ondersteuning van degelijk onderbouwd security & AI beleid. Als input daarvoor richt dit onderzoek zich primair op beantwoording van de vragen:

- Welke security & AI deelgebieden zijn te onderscheiden? Welke voorbeelden zijn illustratief voor die deelgebieden?
- Wat is de stand van zaken in de security & AI deelgebieden (inclusief de kennis en innovatie), wie zijn de relevante actoren en welke rol spelen zij?
- Welke publieke belangen (b.v. economisch, maatschappelijk, nationale veiligheid) zijn gemoeid bij de security & AI deelgebieden? Welke impact is al te zien en wat is de verwachting gezien de ontwikkelingen?

Dit rapport biedt een basis om toe te werken naar een raamwerk waarmee (a) security & AI vraagstukken gedefinieerd en gecategoriseerd kunnen worden, (b) de met de vraagstukken gepaard gaande impact vanuit verschillende invalshoeken (b.v. economisch, nationale veiligheid en maatschappelijk) ingeschat kan worden, waarmee vervolgens (c) de relatie naar handelingsperspectief gelegd kan worden.

1.2 Aanpak

Ter inventarisatie van onderwerpen in het raakvlak van cybersecurity en AI is een uitgebreid literatuuronderzoek uitgevoerd. Hierin zijn meegenomen publicaties over[1]:

- AI en cybersecuritystandaarden, concepten en terminologie;
- Strategie- en beleidsdocumenten over AI, over cybersecurity en over het raakvlak daarvan door nationale en internationale (hoofdzakelijk overheids-)organisaties;
- Wetenschappelijke publicaties ter duiding van de stand van zaken ten aanzien van vakgebieden in het raakvlak van cybersecurity en AI;
- Diverse publicaties ter illustratie van security & AI vraagstukken en de impact ervan.

In de eerste helft van 2024 is dit literatuuronderzoek uitgevoerd en besproken met alle TNO-projectleiders die in de afgelopen jaren betrokken zijn geweest bij tientallen projecten in het raakvlak van cybersecurity en AI. Ook zijn daarbij eerste gesprekken gevoerd met opdrachtgevers van enkele van die projecten.

Op basis van deze inzichten is deze verkenning opgesteld, die dient als input voor discussie over het raakvlak van cybersecurity en AI met relevante belanghebbenden. De terugkoppeling en inzichten die in deze discussies naar voren zullen komen worden verwerkt in een volgende versie van dit rapport.

Afbakening

- AI is een breed begrip en een dynamisch werkveld. Dat geldt ook voor het specifiekere onderwerp van security & AI. Niet alle security & AI casussen zullen worden geadresseerd. Bijvoorbeeld, autonome wapensystemen en de beveiliging daarvan vallen buiten de afbakening van deze verkenning. Verder wordt in deze verkenning ingekaderd op basis de invloed die AI kan hebben en/of het belang van het domein waarin de AI wordt toegepast. Daar waar dergelijke inkadering is toegepast wordt dat in dit rapport toegelicht.
- Tijdens het project zal zoveel mogelijk gebruik worden gemaakt van reeds voorhanden, gepubliceerd materiaal (bijvoorbeeld de ENISA-rapporten) en AI en cybersecurityexpertise aanwezig bij TNO.

1.3 Leeswijzer

De huidige versie van dit rapport dient als basis voor discussie over het raakvlak tussen cybersecurity en AI. Na discussierondes met relevante belanghebbenden zal dit rapport in de tweede helft van 2024 worden aangescherpt met de daaruit voortkomende inzichten.

Hierbij dient opgemerkt te worden dat het onderwerp security & AI dermate in beweging is dat een ‘eindversie’ nog steeds een tijdsgebonden weergave van het raakvlak betreft. De ervaring opgedaan tijdens het schrijven van dit rapport leert dat sommige onderdelen van dit raakvlak zich dermate snel ontwikkelen, dat de stand van zaken in een tijdspanne van een half jaar al grotendeels veranderd kan zijn.

[1] De volledige lijst van publicaties is opgenomen in sectie **Fout! Verwijzingsbron niet gevonden.**.

2 AI, cybersecurity en beleid

2.1 AI

AI is een sterk ontwikkelend vakgebied, dat wordt gekenmerkt door een grote diversiteit aan technieken en toepassingsgebieden. Hoewel de eerste pioniers al meer dan een eeuw geleden speculeerden over 'denkende automaten' en 'robots', zijn het recentere doorbraken geweest die AI onder de aandacht van het bredere publiek hebben gebracht. De introductie van spraakgestuurde aanbevelingssystemen op smartphones vanaf 2011 (zoals SIRI, Google Now en Cortana) en AI-chatbots zoals ChatGPT in 2022 zijn hiervan prominente voorbeelden.

De veelheid aan technieken die bestempeld worden met het etiket AI en het brede bereik van toepassingen leiden tot het gebruik van dubbelzinnige AI-terminologie. Om hierin uniformiteit aan te brengen zijn AI-definities en taxonomieën gepubliceerd en in de loop der tijd aangepast aan de voortgang van AI. In dit rapport wordt uitgegaan van de definitie die wordt gehanteerd in de AI Act [4]. Met behulp van aanvullende taxonomieën worden in de volgende subsecties diverse aspecten van AI beschreven.

2.1.1 Definitie van AI

AI-systemen worden ontwikkeld voor het uitvoeren van taken die 'intelligentie'[2] vereisen. Binnen AI worden technieken gebruikt uit verschillende expertisegebieden zoals informatica, wiskunde, biologie, filosofie, linguïstiek, psychologie en cognitieve wetenschappen [5]. Er is geen uniforme, gestandaardiseerde grens tussen AI-systemen en conventionele, "niet-AI" systemen. Soms wordt de wiskundige of statistische basis als onderscheid genomen (het gebruik van bijvoorbeeld Machine Learning (ML) of zelfs Deep Learning (DL)). Dit kan leiden tot een nauwe, modelgebaseerde definitie die andere slimme en complexe systemen uitsluit en tot een definitie die sterk tijdsgebonden is. Voor beleidsmatige doeleinden biedt dergelijk technisch, inhoudelijk onderscheid weinig houvast. Vaak zijn verschillende kenmerken van AI- en traditionele systemen meer relevant. In [5] worden vier elementen aangegeven als kenmerkende verschillen tussen AI- en meer traditionele systemen:

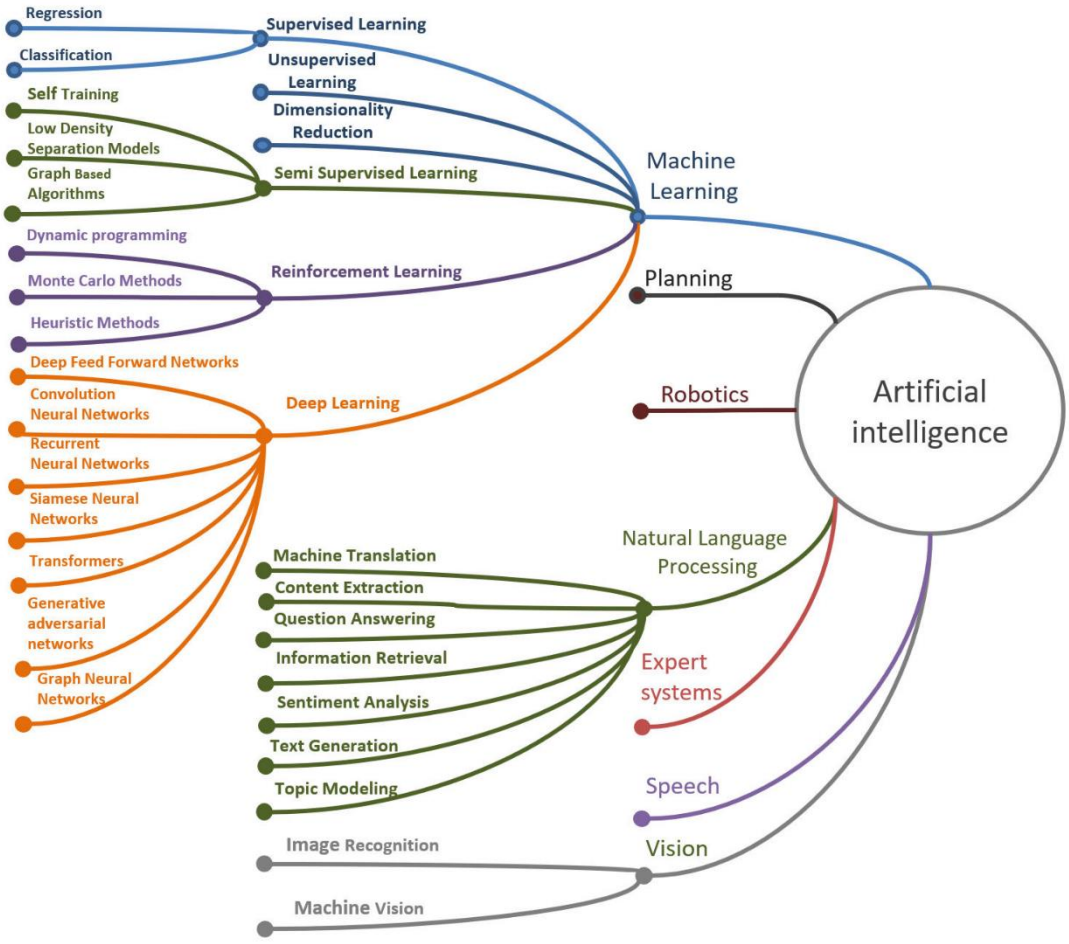
1. Interactief: de input van een AI-systeem verloopt via sensoren en/of een menselijke inbreng. Het systeem leidt hieruit een resultaat[3] af, dat vervolgens aan een menselijke gebruiker of machine wordt teruggekoppeld, en heeft zo effect op een fysieke of virtuele omgeving.

[2] Het begrip 'intelligentie' is onderhevig aan maatschappelijke en filosofische discussie. Van Dale definieert het als 'verstandelijk vermogen', andere definities benadrukken aspecten van leren, het begrijpen van nieuwe, ingewikkelde situaties, of juist het toepassen van kennis in de echte wereld, zoals probleemoplossend vermogen. Het kwantificeren van intelligentie door middel van bijvoorbeeld de IQ-test is geenszins alom geaccepteerd: een stroming van schrijvers stelt bijvoorbeeld dat er te veel nadruk ligt op analytische vermogens. Hierbij zijn termen zoals 'sociale intelligentie' geïntroduceerd om de waarde van andere elementen in het fenomeen intelligentie te benadrukken [105]. Tegelijkertijd is het de vraag of men geneigd is een antropocentrisch beeld van intelligentie aan te hangen: kunnen mensen zich een niet-menselijke, of zelfs niet-biologische vorm van intelligentie voorstellen?

[3] Voorbeelden van typen resultaat zijn voorspellingen, diverse media (content), aanbevelingen of beslissingen.

2. Contextueel: over het algemeen gebruikt AI verschillende soorten databronnen, waaronder gestructureerde en/of ongestructureerde digitale datasets. Ook sensorische input kan worden gebruikt.
3. Autonomie: AI-applicaties kunnen opereren met een toenemende mate van autonomie, waardoor minder intensief menselijk toezicht nodig is.
4. Adaptief: sommige AI-systemen kunnen real-time veranderende datasets analyseren en hun gedrag aanpassen op basis van nieuwe beschikbare data.

Wanneer in populaire media wordt gesproken over AI, worden verschillende termen gebruikt. Het is hierbij mogelijk om onderscheid te maken op basis van allerlei verschillende assen. In de academische wetenschap wordt vaak onderscheid gemaakt op basis van de onderliggende technologie. De onderstaande taxonomie in Figuur 3 is hier een voorbeeld van.



Figuur 3: Taxonomie van AI. Het linker gedeelte van deze taxonomie is grotendeels gebaseerd op technische aspecten, terwijl het deel rechts (de hoofdcategorieën) voornamelijk functioneel van aard zijn [6].

Termen als *machine learning*, *neural networks* en *deep learning* zijn nuttig voor de technische ontwikkeling van toepassingen. Voor beleidsmatige doeleinden zijn ze echter vaak niet toereikend, omdat de onderliggende technologie weinig zegt over de manier waarop het model interacteert met en impact maakt op de wereld. Daarbij komt dat bijvoorbeeld voortgaande autonomisering van systemen (zoals voertuigen) ertoe leidt dat er in één systeem steeds meer, verschillende AI-technieken worden toegepast om een steeds

bredere combinatie van taken uit te voeren (zoals beeldherkenning, route-optimalisatie, voertuigbesturing en -bewaking).

Daarom kan het zinvoller zijn om een meer taakgerichte of functionele taxonomie van AI te hanteren. AI-technologieën kunnen op een schaal geplaatst worden van *narrow* AI tot *general* AI. *Narrow* AI is in staat is om een specifieke, afgebakende taak uit te voeren, terwijl *general* AI een groot aantal verschillende taken kan uitvoeren. Het algemene beeld is dat AI op dit moment vooral nog *narrow* is. Hoewel meningen verschillen over de mate waarin *general* AI mogelijk is, kijken veel onderzoekers met grote interesse naar de recente ontwikkelingen rondom Large Language Models (LLM's). Dit zijn modellen die getraind zijn op grote hoeveelheden tekst en daardoor de structuur van taal geleerd hebben. Hierdoor kunnen de modellen menselijke taal "begrijpen" en genereren. Op de schaal van *narrow* naar *general* AI hebben LLM's al een stap gemaakt richting *general* AI: hetzelfde model kan gebruikt worden om een kookrecept te vertalen, een fictief verhaal te schrijven in de stijl van een bekende schrijver, of advies te geven over het programmeren van software. Daarbij is de gedachte dat LLM's een rol kunnen spelen op deze dimensie door een brugfunctie te vervullen tussen verschillende specifieke (*narrow*) modellen. De resultaten van verschillende modellen kunnen dan, via menselijke tekst, geduid en vertaald worden, en vervolgens tot input leiden voor weer een ander specifiek model. Hierdoor kan het systeem als geheel meer verschillende taken kan vervullen.

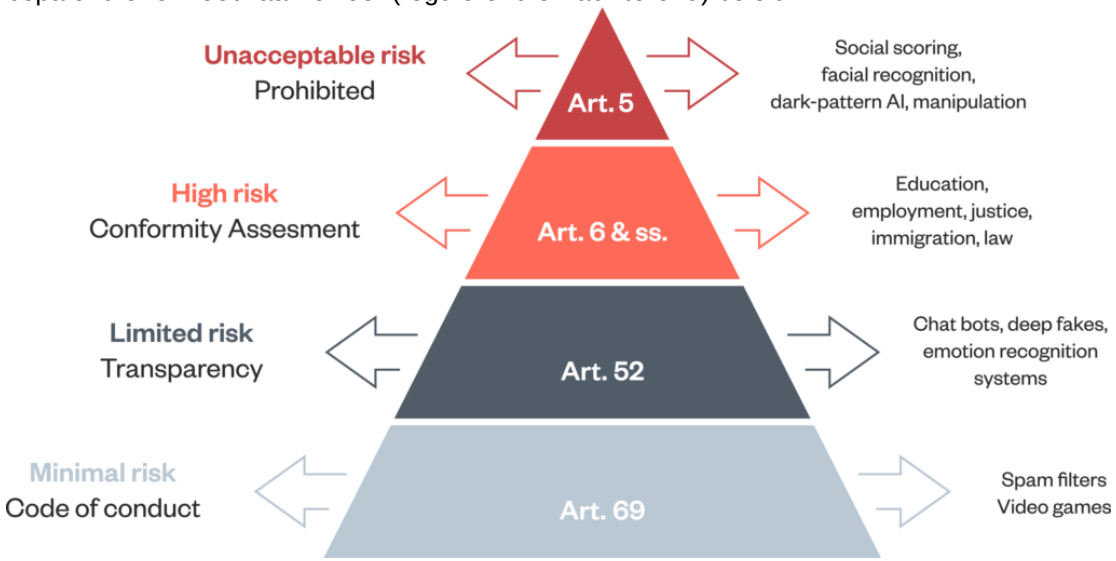
Een ander belangrijk onderscheid binnen functionele omschrijvingen van AI is dat tussen voorspellende AI en generatieve AI (niet te verwarren met *general* AI). De afgelopen decennia hebben technologische ontwikkelingen zich vooral gericht op voorspellende AI: modellen die op basis van grote hoeveelheden data accurate voorspellingen van toekomstige (of niet eerder beschouwde) situaties kunnen doen. Generatieve AI is juist in staat om originele voorbeelden (bijvoorbeeld stukken tekst of beelden) te creëren op basis van een bepaalde input. Deze nieuwe voorbeelden voldoen grotendeels aan de patronen die zijn geobserveerd in, wederom, grote hoeveelheden data.

De taken die aan een AI-systeem kunnen worden overgelaten zijn grofweg onder te verdelen in de volgende categorieën [7]:

- Classificatie (Wat is het?)
- Associatie (Waar lijkt het op? Waar hoort het bij?)
- Optimalisatie (Wat is de beste of meest efficiënte optie?)
- Voorspelling (Wat gaat het doen?)
- Creatie

Deze functionele taxonomie van AI is een stuk intuïtiever en bruikbaar voor uiteenlopende gebruikers, vergeleken met meer technische onderverdelingen van AI. De functionele taxonomie is echter niet per definitie geschikt om tot een goede inschatting van de effecten of de veiligheid van de toepassing te komen: AI-toepassingen met dezelfde functionele componenten kunnen, toegepast in hun eigen context, verschillende effecten teweegbrengen. In de EU AI Act [4] wordt onderscheid gemaakt tussen groepen toepassingen van AI, die in verschillende risico-categorieën passen op basis van het voorgenomen doel van de toepassing (zie ook Figuur 4). Uit deze categorisering blijkt dat er maar een zeer klein deel van AI-toepassingen als inherent kwalijk of risicovol worden gezien en dat het toepassingsdomein een bepalende factor is in de risico-inschatting. Zelfs de toepassingen die in de *"unacceptable risk"* categorie vallen, zijn binnen sommige domeinen gebruikelijk. Zo wordt *social scoring* binnen allerlei dienstverleningen gebruikt (bijv. Uber), maar wordt het onacceptabel gezien voor actoren om dit te gebruiken in een context waar de data is verzameld. Gezichtsherkenning wordt in smartphones en op grenscontroles

toegepast, maar is bijvoorbeeld niet toegestaan voor massa-surveillance. Kortom: de manier waarop en de context waarin een AI-systeem wordt toegepast is doorgaans het meest bepalend of er noodzaak is voor (regulerend of faciliterend) beleid.



Figuur 4: De risico-pyramide in de EU AI Act, volgens [8].

De veiligheid van een AI-toepassing hangt dus niet alleen af van de technologie op zichzelf, maar vooral ook van hóe deze wordt gebruikt in een organisatie of de maatschappij. Een discipline die deze constatering onderschrijft is het *Socio-Technical Systems* perspectief. Hierbinnen is ook aandacht voor het fenomeen "AI-toepassingen", waarbij de technologie als meer vervlochten met de menselijke (sociale) wereld wordt gezien. Hierbij kan een taxonomie opgesteld worden van mens-machineteams en kunnen archetypes van AI-toepassingen worden opgesteld [9] (zie ook Figuur 5).

Layer 1: Sociomaterial entities	Layer 2: Dimensions		Layer 3: Characteristics								
Human (human agency)	Human cognitive functions	NE	Perceiving	Reasoning	Predicting	Planning	Decision-making	Explaining	Interacting	Creating	Empathizing
	Interaction human to AI	ME	Facilitating			Verifying			Supplementing		
	Human focus	ME	Sensemaking		Creativity		Compassion			Flexibility	
AI (material agency)	AI cognitive functions	NE	Perceiving	Reasoning	Predicting	Planning	Decision-making		Interacting	Creating	
	Interaction AI to human	ME	Facilitating			Verifying			Supplementing		
	AI focus	ME	Automation				Augmentation				
Sociomaterial practices	Form of interworking	ME	Parallel			Sequential			Flexible		
	Mode of interworking	ME	Singular				Continuous				
	Learning	ME	None	AI learns		Human learns		Human and AI learn separately		Co-evolution	

ME = mutually exclusive, NE = non-exclusvie

Figuur 5: Karakteristieken waarmee mens-machineteams geclassificeerd kunnen worden vanuit het socio-technical systemsparadigma [9].

2.1.2 Beleid

Overheden realiseren zich de veelbelovende toekomst die een florerende AI-sector kan bieden en trachten in toenemende mate om dit te stimuleren. Dit omvat aspecten zoals het ondersteunen van de MKB-sector die vaak minder toegang heeft tot de benodigde

infrastructuur vergeleken met grotere bedrijven. Echter zijn overheden zich ook bewust van de mogelijke dreigingen van onbeteugelde AI-proliferatie. Daarom hebben een aantal belangrijke overheidsspelers op nationale en internationale niveaus beleid geformuleerd dat de ontwikkeling van AI stimuleert, mits die veilig en betrouwbaar is.

In China hebben zeven ministeries in 2023 een set van interim maatregelen uitgebracht aangaande generatieve AI. Vergeleken met andere landen, waar beleid overkoepelend is, focust China op het identificeren en reguleren van individuele problemen: algoritmes, deepfakes en generatieve AI [10]. Hierdoor kan de Chinese overheid een proactieve houding nemen, aangezien het op kortere termijn specifieke problemen kan prioriteren en adresseren. De komende jaren gaat de Chinese overheid werken aan een algehele AI-wet. Het relevante beleid voor cybersecurity, gefocust op generatieve AI, omvat maatregelen die de ontwikkeling van AI en veiligheid promoten. Door hardware en software, waaronder datasets, beschikbaar te stellen, hoopt de overheid AI-ontwikkeling te versnellen. Daarnaast omvat de wet een aantal vereisten jegens de overheid, zoals maatregelen om discriminatie te voorkomen en respect voor socialistische waarden, en gebruikers, zoals persoonlijke databeveiliging en maatregelen om verslaving te voorkomen. Naleving wordt gecontroleerd middels rapportageverplichtingen en inspecties. Belangrijk is dat AI-service aanbieders aangemerkt worden als producenten van onlinemateriaal. Daarom worden zij wettelijk verantwoordelijk gehouden voor materiaal dat hun gebruikers produceren. Zo zou een serviceprovider voor het genereren van plaatjes aansprakelijk zijn als een gebruiker die gebruikt om posters die oproepen tot separatisme te produceren.

In hetzelfde jaar produceerde president Biden een presidentieel decreet [11] waarmee de Amerikaanse overheid het veilig en verantwoord gebruik van AI wenst te bevorderen. Dit document omvat een achttal prioriteiten: de veiligheid van AI verzekeren, innovatie en competitie stimuleren[4], werknemers behoeden voor negatieve effecten en verder ondersteunen, rechtvaardigheid en burgerrechten bevorderen, belangen van gebruikers bevorderen, privacy-gerelateerde risico's mitigeren, risico's rondom het gebruik van (generatieve) AI door de federale overheid beheren, en de VS positioneren als globale leider op het gebied van verantwoord gebruik van AI. Dit beleidsstuk specificeert de benodigde acties van de Amerikaanse overheid om deze doelen te realiseren. Ook stelt het binnen welke tijdlijn deze moeten worden ondernomen.

Een andere grote speler op dit gebied is de EU, die in mei van 2024 de AI Act [4] goedkeurde. Dit is de eerste alomvattende wetgeving en beoogt om te garanderen dat de positieve mogelijkheden van AI geen afbreuk doen aan de openbare belangen rondom gezondheid, veiligheid en grondrechten. Hiervoor wordt geredeneerd vanuit een risicomanagement perspectief, waarbij AI-systemen worden beoordeeld en vervolgens worden ingedeeld op basis van de risico's die ze vormen. Afhankelijk van de risicocategorie worden er bepaalde eisen gesteld aan de AI-systemen. Systemen die een onacceptabel risico vormen, bijvoorbeeld wanneer ze ontworpen zijn om iemands onderbewustzijn te beïnvloeden, worden verboden. AI-systemen in de hoge risico categorie worden wel toegelaten, mits ze geëvalueerd worden voor toetreding tot de markt en gedurende de levenscyclus. Daarnaast doelt dit beleidsstuk om innovatie te stimuleren, met nadruk op kleine en middelgrote ondernemingen. Om deze wetten uit te voeren heeft de EU een organisatiestructuur ontworpen met afdelingen op nationaal niveau en EU-niveaus, waarbinnen overheden en andere belanghebbenden betrokken zijn. Op deze manier geeft de EU een invulling aan de wens om AI-wetgeving binnen de EU te harmoniseren, waarmee ondernemingen met

[4] Hieronder vallen maatregelen zoals het faciliteren van publiek-private samenwerking, het aantrekken van 500 nieuwe AI-experts per 2025 en stimuleren van de intellectuele eigendom positie.

naleving van één wet toegang verkrijgen tot de markt van de gehele EU. Specifieker nationaal beleid is onder andere vastgelegd in het strategische actieplan voor AI en de overheidsbrede visie generatieve AI. Een concreet beleidsinstrument zijn investeringen in economische groei met AI middels het nationale groeifonds AiNed en de ontwikkeling van GPT-NL.

2.1.3 Belanghebbenden

Het spectrum van belanghebbenden binnen het domein van AI is omvangrijk en divers. Een groot aantal burgers en ondernemingen maakt in verschillende gradaties gebruik van AI-diensten [12].

Burgers komen in aanraking met AI-toepassingen via diverse kanalen, waaronder persoonlijke digitale assistenten (zoals Siri of ChatGPT), navigatieapplicaties (bijvoorbeeld Google Maps) en gepersonaliseerde streamingdiensten. In de zakelijke sector wordt AI ingezet voor onder meer klantenservice-optimalisatie, geavanceerde marketing- en verkoopstrategieën (middels gepersonaliseerde aanbevelingen) en voorspellende analyses ten behoeve van efficiënt voorraadbeheer.

Het ontwikkelingsveld van AI kenmerkt zich eveneens door een grote verscheidenheid aan actoren. Multinationale technologiebedrijven, waaronder Google, Microsoft en IBM, investeren aanzienlijke middelen in AI-onderzoek en -ontwikkeling. Sociale mediabedrijven ontwikkelen AI-systemen voor contentmoderatie, aanbevelingsalgoritmen en advertentie-optimalisatie. Veel van deze aanbieders beschikken over grootschalige cloudinfrastructuren, en daarmee over belangrijke AI 'productiemiddelen' zoals enorme hoeveelheden data en rekenkracht. Academische en onderzoeksinstellingen leveren substantiële bijdragen aan zowel fundamenteel als toegepast AI-onderzoek. Daarnaast zijn er talrijke andere partijen actief in AI-ontwikkeling, waaronder startups, overheidsorganisaties en diverse industriële sectoren.

Gezien de heterogeniteit van zowel gebruikers als ontwikkelaars van AI-diensten, worden in dit rapport drie belangrijke subcategorieën nader belicht:

- Beleid en regelgeving
- Aanbieders/ontwikkelaars van publiek toegankelijke generatieve AI-diensten
- De chipindustrie

Voor de laatste twee categorieën belanghebbenden, worden verschillende aanbieders van AI en actoren in de chipindustrie genoemd. Dit is geen uitputtende lijst, maar dient als illustratieve lijst.

2.1.3.1 Beleid en regelgeving

Voor Nederland zijn er twee relevante beleidsmakers op het gebied van AI: de EU en de Nederlandse overheid. De EU heeft een uitgebreide aanpak ontwikkeld voor het reguleren van AI met de AI Act (zie voor verdere uitleg sectie 2.1.2). In deze sectie volgt een beknopte opsomming van elementen uit AI-beleid, die het meest relevant zijn in het raakvlak van cybersecurity en AI.

In Nederland wordt het beleid voor AI vormgegeven door een samenspel van verschillende overheidsinstanties en organisaties, elk vanuit hun eigen beleidsterrein. Ter illustratie worden hieronder enkele organisaties en hun rol uitgelicht:

- Het Ministerie van Economische Zaken en Klimaat (EZK) speelt hierin een coördinerende rol en is gericht op economische meerwaarde van AI. Ook speelt de Rijksdienst voor Digitale Infrastructuur, vallend onder EZK, een rol in het faciliteren van de digitale randvoorwaarden voor AI-ontwikkeling.
- Ministeries van Binnenlandse Zaken en Koninkrijksrelaties en van Justitie en Veiligheid zich richten op specifieke aspecten zoals digitale overheid AI-ethiek en toepassingen van AI voor hun taakstellingen (bijvoorbeeld ter verbetering van de handhaving van de openbare orde).
- Ook het Ministerie van Defensie draagt bij aan het AI-beleid, met name op het gebied van nationale veiligheid en militaire toepassingen. De Nederlandse AI Coalitie fungeert als een publiek-private samenwerking die diverse stakeholders samenbrengt. Onafhankelijke instanties zoals het Rathenau Instituut en de Autoriteit Persoonsgegevens dragen bij aan het beleid door advies te geven over maatschappelijke impact en toezicht te houden op privacy kwesties. De Inspectie Leefomgeving en Transport is betrokken bij de controle en toepassing van AI-systemen binnen haar werkterrein van infrastructuur, milieu en transport.

2.1.3.2 Aanbieders/ontwikkelaars van generatieve AI

De markt voor publiek toegankelijke generatieve AI wordt momenteel gedomineerd door een select aantal grote technologiebedrijven (en tevens cloudaanbieder) en gespecialiseerde AI-ondernemingen. Voorop staat OpenAI, bekend van GPT-modellen en ChatGPT, dat met zijn geavanceerde taalmodellen een leidende positie inneemt. Google, met zijn Bard-assistent en PaLM-model is eveneens toonaangevende spelers. Microsoft heeft zijn positie versterkt door samenwerking met OpenAI en integratie van diens technologie in diverse producten.

Volgens de Forbes AI 50-lijst [13], die de meest veelbelovende particuliere AI-bedrijven opsomt, zijn de toepassingen van AI breed. De drie hoogst gewaardeerde bedrijven op deze lijst - OpenAI ($86 miljard), Anthropic ($18,4 miljard), dat zich richt op veilige en betrouwbare AI, en Databricks ($43 miljard), bekend om zijn data-analyse en machine learning platform - bedienen klanten variërend van financiële instellingen en adviesbureaus tot overheidsinstanties en energiebedrijven.

Naast het benoemen van deze Amerikaanse bedrijven, is het van belang te erkennen dat ook Europa een groeiende rol speelt in de AI-sector. Europese AI-bedrijven trekken aanzienlijke investeringen aan. Het Duitse Aleph Alpha, gespecialiseerd in transformatieve AI zoals grote taalmodellen en multimodale modellen, haalde in 2023 meer dan 500 miljoen euro op. In Frankrijk wist Mistral AI, een startup voor generatieve AI, in minder dan zes maanden tijd 490 miljoen euro aan investeringen binnen te halen. [14]

Zoals eerder genoemd zijn deze bedrijven een illustratieve lijst en zijn er nog vele andere aanbieders en ontwikkelaars.

2.1.3.3 De chipindustrie

De chipindustrie vervult een onmisbare functie in de ontwikkeling en implementatie van AI-technologieën, waarbij een aantal sleutelspelers de markt domineert.

Internationaal is TSMC (Taiwan Semiconductor Manufacturing Company) een sleutelspeler als 's werelds grootste onafhankelijke chipfabrikant. NVIDIA heeft zich gepositioneerd als marktleider in AI-specifieke chips, met hun GPU's (Graphics Processing Unit) die essentieel zijn voor het trainen van complexe AI-modellen. Intel en AMD, traditioneel sterk in CPU's

(Central Processing Unit) hebben hun focus verlegd naar AI-geoptimaliseerde processors. Apple ontwikkelt eigen AI-chips voor hun consumentenelektronica, terwijl Google met hun Tensor Processing Units (TPU's) de efficiency van AI-berekeningen in cloudomgevingen verbetert. De dynamiek in de chipsector heeft verstrekkende gevolgen voor de toegankelijkheid, schaalbaarheid en energieverbruik van AI-toepassingen, wat directe implicaties heeft voor de Nederlandse open strategische autonomie (zie sectie 3.6.1).

In de Nederlandse context spelen ASML, ASM International, Nexperia en NXP Semiconductors een belangrijke rol. ASML is een cruciale speler in de halfgeleiderindustrie. Het bedrijf levert geavanceerde lithografiesystemen die essentieel zijn voor de productie van microchips. Deze technologie stelt chipfabrikanten in staat om steeds kleinere en krachtigere halfgeleiders te produceren, een bijdrage levert aan de vooruitgang van AI-systemen. Nexperia fabriceert ook halfgeleiders. ASM International richt zich op de ontwikkeling van specifieke productieprocessen. NXP levert AI-geoptimaliseerde chips voor automobiele en IoT-toepassingen.

Binnen de chipindustrie zijn er nog vele andere spelers, de opsomming van bovenstaand is natuurlijk niet volledig.

2.2 Cybersecurity

Cybersecurity betreft de beveiliging van digitale informatie, systemen en netwerken tegen cyberaanvallen daarop. De toepassing van cybersecurity is een steeds belangrijker onderdeel geworden van de waarborging van de meerwaarde die voortvloeit uit de voortgaande digitale transitie. Hierbij speelt cybersecurity een rol in zowel het vertrouwen dat nodig is voor de adoptie van digitale innovaties, als in het beteugelen van de schadelijke effecten van cybercriminaliteit. Hoewel kwantitatieve schattingen van deze effecten zeer lastig te maken zijn,[5] lijkt er consensus te bestaan dat de kosten van cybercriminaliteit enkele procenten van het BNP beslaan[6] . Verder zijn er aanwijzingen dat het vertrouwen aanzienlijke invloed kan hebben op de omzet van bepaalde digitale technologie; het kan zelfs oplopen naar tientallen procenten[7]. Dergelijke indicaties illustreren de significante economische impact van cybersecurity.

Sinds de opkomst van computersystemen wordt er geëxperimenteerd met technieken zoals virussen om de kwetsbaarheid van die systemen te testen. De opkomst van het internet en toenemende economische waarde van digitale dienstverlening heeft geleid tot malafide gebruik van dergelijke technieken. Om deze cyberaanvallen te pareren zijn cybersecuritytechnieken ontwikkeld wat vervolgens heeft geleid tot een voortdurende wedloop tussen cyberaanval en -verdediging. In de loop der tijd heeft aan weerszijden automatisering zijn intrede gedaan, met als meest recente ontwikkeling de introductie van geavanceerde automatisering door middel van AI.

[5] Met cybercriminaliteit gaan directe kosten gepaard (zoals herstelkosten, betaling van losgeld, etc.), maar ook indirecte kosten zoals gedervde inkomsten en in uitzonderlijke gevallen failissementen. Voorts zijn en kosten gerelateerd aan het voorkomen van cyberincidenten. Verder ontbreekt een goed overzicht van alle cyberincidenten, omdat lang niet alle incident gemeld hoeven te worden. Dergelijke factoren maken het zeer lastig om een goede kwantitatieve inschatting van de kosten van cybercriminaliteit te maken.

[6] Gerefereerd wordt vaak naar het hackerpocalypse rapport van Cybersecurity Venture (dat vrij gedateerd is) [106], maar ook Statistica komt tot dergelijke conclusie [107].

[7] Een door McKinsey uitgevoerd onderzoek onder een brede groep van industriële organisaties komt uit op een percentage van 28% ‘extra omzet’ in de Internet of Things sector als cybersecurityrisico's beheersbaar zijn [108].

2.2.1 Beschikbaarheid, Integriteit en Vertrouwelijkheid

De effecten van cyberaanvallen worden vaak onderverdeeld in drie typen: de schending van de beschikbaarheid, de integriteit, en de vertrouwelijkheid van digitale gegevens en dienstverlening. Deze onderverdeling wordt aangeduid met BIV (Engels: CIA triad) [15, 16]. Hoewel deze onderverdeling primair voortkomt vanuit het deelgebied informatiebeveiliging, wordt in dit rapport een bredere definitie gehanteerd van de beoogde bescherming geboden door cybersecurity:

- Beschikbaarheid – waarborging van tijdige en betrouwbare toegang tot en gebruik van digitale diensten en systemen;
- Integriteit – bescherming tegen onbedoelde modificatie of vernietiging van digitale gegevens en waarborging van de onweerlegbaarheid en authenticiteit daarvan;
- Vertrouwelijkheid – voorbehouden van toegang tot en ontsluiting van digitale diensten en gegevens aan daartoe geautoriseerde personen, inclusief de bescherming van privacy- en bedrijfsvoeringgevoelige diensten en informatie.

Merk op dat deze 'bredere definitie' van cybersecurity ruimte laat om het te interpreteren als de steeds gangbaarder wordende term cyberweerbaarheid. Bijvoorbeeld, de term 'beschikbaarheid' is niet alleen gedefinieerd onder omstandigheden waarin opzettelijke aanvalspogingen gedaan worden om een systeem onbeschikbaar te maken (sabotage), maar ook maatregelen om een digitaal systeem beschikbaar te houden tijdens niet-moedwillige verstoringen (bijvoorbeeld verstoring van een sensorsysteem door falende AI) vallen volgens de definitie onder cybersecurity. Desalniettemin wordt de term cyberweerbaarheid vaak nog breder gedefinieerd dan de in dit rapport gebruikte term cybersecurity.

2.2.2 Cybersecurity functies

Cybersecurity omvat een breed scala aan digitale beveiligingsmaatregelen. Een veel gebruikt raamwerk is het Cybersecurity Framework 2.0 van NIST [17], dat een onderverdeling maakt van cybersecurity in de "functies": identificeren, beschermen, detecteren, reageren, en herstellen. In de recentste versie is hier governance aan toegevoegd. Deze functies zijn verder onderverdeeld in categorieën, zie Figuur 6. Dit raamwerk bevat generieke richtlijnen voor de inrichting van cybersecuritymaatregelen. Deze worden door organisaties gebruikt als basis voor cybersecuritymaatregelen die worden toegespitst op de organisatie specifieke bedrijfsvoering omstandigheden. Hiermee draagt het raamwerk ook bij aan een zekere mate van standaardisatie en overdraagbaarheid van cybersecuritymaatregelen tussen soortgelijke organisaties. In sectie 3.5 wordt nader toegelicht voor welke van deze functies AI wordt toegepast.

Function	Category	Category Identifier
Govern (GV)	Organizational Context	GV.OC
	Risk Management Strategy	GV.RM
	Roles, Responsibilities, and Authorities	GV.RR
	Policy	GV.PO
	Oversight	GV.OV
	Cybersecurity Supply Chain Risk Management	GV.SC
Identify (ID)	Asset Management	ID.AM
	Risk Assessment	ID.RA
	Improvement	ID.IM
Protect (PR)	Identity Management, Authentication, and Access Control	PR.AA
	Awareness and Training	PR.AT
	Data Security	PR.DS
	Platform Security	PR.PS
	Technology Infrastructure Resilience	PR.IR
Detect (DE)	Continuous Monitoring	DE.CM
	Adverse Event Analysis	DE.AE
Respond (RS)	Incident Management	RS.MA
	Incident Analysis	RS.AN
	Incident Response Reporting and Communication	RS.CO
	Incident Mitigation	RS.MI
Recover (RC)	Incident Recovery Plan Execution	RC.RP
	Incident Recovery Communication	RC.CO

Figuur 6: Cybersecurityraamwerkfuncties en categorieën; bron: NIST CSF 2.0 [17].

2.2.3 Belanghebbenden

Het speelveld van belanghebbenden ten aanzien van cybersecurity kan op hoofdlijn worden onderverdeeld in:

- 'cyberdoelwitten': aanbieders en gebruikers van digitale diensten,
- cyberaanvallers en –verdedigers, en
- regelgevers en toezichthouders.

2.2.3.1 Cyberdoelwitten: aanbieders en gebruikers van digitale diensten

Digitalisering speelt in onze hedendaagse maatschappij een dermate grote rol dat zo goed als elke aanbieder of gebruiker van digitale diensten en systemen doelwit kan worden van een cyberaanval. Wel valt er onderscheid te maken in de mate van het belang van cybersecurity voor een specifieke belanghebbende.

Aan de aanbodzijde van digitale dienstverlening zijn vanuit Europese en nationale regelgeving specifieke doelgroepen aangewezen op basis van het belang van (cybersecurity voor) hun dienstverlening. Voorbeelden hiervan zijn de nationale aanbieders van kritieke processen [18] en de in de CER (Critical Entities Resilience directive) [19] opgenomen kritieke entiteiten. Ook cloudaanbieders zijn in de CER aangewezen als kritieke entiteiten. Grote, Amerikaanse aanbieders spelen een speciale rol in de zin dat zij niet alleen een kritieke entiteit zijn, maar ook veel cybersecurity- en AI-diensten aanbieden. Naast cybersecurityregulering voor deze kritieke doelgroepen speelt cybersecurity in toenemende mate een rol in gereguleerde sectoren, bijvoorbeeld in de automotive sector, en voor

aanbieders van producten met digitale componenten [20]. Ook de digitale overheid dient haar cybersecurity op orde te hebben, temeer omdat de digitale overheid tot de meest aangevallen organisaties behoort (zie Figuur 8 in [21]).

De gebruikers van digitale diensten spelen ook een rol in de beheersing van hun cyberrisico's en het treffen van passende maatregelen. Het creëren van de bewustwording van gebruikers over cybersecurity wordt gestimuleerd met bijvoorbeeld voorlichtingscampagnes door aanbieders van digitale diensten en door de overheid. Het instrument van voorlichtingscampagnes is echter niet eenvoudig te hanteerbaar: de relatie tussen voorlichting over cybersecurity en het beoogde effect is complex en lastig meetbaar [22]. Mede hierdoor is gezocht naar aanvullende beleidsinstrumenten om via wetgeving (bijvoorbeeld de CRA) meer cybersecurityverantwoordelijkheid te verschuiven richting aanbieders van digitale producten en diensten.

2.2.3.2 Cyberaanvallers en -verdedigers

Actoren die cyberaanvallen plegen doen dat met verschillende motieven en met verschillende middelen en capaciteiten. Cyberaanvallers zijn onderscheidbaar in bepaalde stereotypen die acteren in een cybercrimineel ecosysteem van toeleveringsketens, digitale dienstverlening en infrastructuur. Veelgebruikte stereotypering van malafide actoren onderscheidt o.a. script kiddies, georganiseerde cybercriminelen, statelijke actoren, hacktivisten en 'insiders' (b.v. kwaadwillende medewerkers). Op het darkweb kunnen actoren terecht om de voor hun doeleinden benodigde, al dan niet AI ondersteunde, middelen en cybercrime-as-a-service diensten te verkrijgen.

De cybersecuritysector staat de aanbieders en gebruikers van digitale diensten ter zijde in hun strijd tegen de cyberaanvallers. De sector bestaat uit een cybersecurityproductmarkt en een cybersecuritydienstenmarkt. De productmarkt wordt gekenmerkt door complexe productontwikkeling, waarin ook AI een steeds grotere rol speelt, en vergt grote voorinvestering en ervaring met softwareontwikkeling. In de productmarkt hebben private partijen uit de VS een aanzienlijke marktmacht, met enkele Nederlandse spelers in specifieke niches (bijvoorbeeld EclecticIQ in de deelmarkt van Cyber Threat Intelligence). De cybersecuritydienstenmarkt is sterker klant-specifiek en nationaal georiënteerd, met sterke Nederlandse spelers zoals Fox-IT en Northwave.

Cybersecurityregelgevers- en toezichthouders

De behoefte aan cybersecurity – en regelgeving daarvoor – komt in de praktijk voort uit digitalisering. Omdat de wijze en tempo van digitalisering per sector verschilt, is cybersecurityregelgeving veelal sectoraal ingebed. Een overzicht van wet- en regelgeving cybersecurity is in 2021 opgesteld vanuit het Programma Nederland Digitaal Veilig (door ministerie van Justitie en Veiligheid) [23].

Mede om meer uniformiteit aan te brengen in de versnipperde cybersecurityregelgeving zijn er sinds het vorige decennium Europese en nationale cybersecurity overheidsorganisaties ingesteld. Naast de taak voor uniformering van regelgeving, kregen deze ook de taak om een constructieve rol te vervullen voor de cyberweerbaarheid van de Europese en de Nederlandse samenleving. In Nederland wordt deze taak uitgevoerd door het Nationaal Cyber Security Centrum (NCSC) en deze werkt in Europees verband samen met andere nationale Computer Emergency Response Teams (CERTs) onder regie van de European Union Agency for Cybersecurity (aangeduid met ENISA, naar de voorgaande naam: European Network and Information Security Agency). Deze aan ENISA toebedeelde rol is vastgelegd in de Cybersecurity Act (CSA, [24]). Een meer facilitaire (en minder operationele)

rol is toebedeeld aan het ECCC (European Cybersecurity Competence Center) voor het bevorderen van Europa's cyberweerbaarheid en haar concurrentievermogen.

2.2.4 Beleid

Ten aanzien van cybersecurity heeft de overheid een aantal taken. Voor de uitvoering van deze taken is door de overheid (onder coördinatie van ministerie Justitie en Veiligheid) de Nederlandse Cybersecurity Strategie 2022-2028 (NLCS) [25] opgesteld. Deze kent vier pijlers:

- Digitale weerbaarheid van de overheid, bedrijven en maatschappelijke organisaties,
- Veilige en innovatieve digitale producten en diensten,
- Tegengaan van digitale dreigingen van staten en criminelen en
- Cybersecurity-arbeidsmarkt, onderwijs en digitale weerbaarheid van burgers.

Omdat de digitale samenleving zich afspeelt voorbij geografische grenzen, sluit het nationale beleid aan op Europees beleid. De kern daarvan is vastgelegd in de EU Cybersecurity Strategy [26] en deze bestaat uit drie actiegebieden:

- Veerkracht, technologische soevereiniteit en leiderschap,
- Operationele capaciteit om te voorkomen, af te schrikken en te reageren,
- Samenwerking om een wereldwijde en open cyberruimte te bevorderen.

In lijn met deze strategie zijn Europese cybersecurityrichtlijnen uitgevaardigd voor Network and Informatie Security (NIS 2, [27]) en de Critical Entities Resilience (CER, [28]). De NIS 2 beoogt een versterking van het cybersecurityniveau in Europa door het instellen van een zorgplicht (bedrijven moeten een cyberrisicoanalyse uitvoeren en op basis daarvan proportionele cybersecuritymaatregelen treffen) en een incidentmeldplicht. Deze verplichtingen worden opgelegd aan in de CER benoemde partijen (de annex van de CER bevat een volledige lijst). Verder vereist de NIS 2 dat een nationaal cybersecuritycentrum (in Nederland is dat het NCSC) toezicht zal houden op deze zorg- en meldplicht. De NIS 2 en CER richtlijn moeten in de tweede helft van 2024 zijn omgezet in nationale wetgeving en van kracht worden.

In parallel aan deze operationele cybersecurity- en weerbaarheidrichtlijnen voor (onder andere) digitale infrastructuuraanbieders, werkt de Europese Commissie aan de Cyber Resilience Act (CRA, [20]). Deze verordening verplicht leveranciers van digitale producten om te voldoen aan Europees geharmoniseerde cybersecurityeisen gedurende de volledige levensduur. Hierbij is onderscheid gemaakt in een aantal productcategorieën van toenemende kritikaliteit. Digitale producten gebruikt in de digitale infrastructuur (bijvoorbeeld firewalls, operating systems en systeembesturingssystemen) vallen daarbij in de categorie met de strengste eisen en auditverplichting. De CRA-verordening is in behandeling bij de Europese Commissie en het duurt mogelijk nog enkele jaren voordat deze daadwerkelijk van kracht wordt.

In deze cybersecuritystrategieën, richtlijnen en wetten komen nauwelijks specifieke referenties naar AI voor[8]. Dit neemt niet weg dat cybersecuritymaatregelen effect kunnen hebben op AI en vice versa. Het raakvlak tussen cybersecurity en AI staat dan ook prominent in het vizier van relevante uitvoerende organisaties zoals ENISA en ECCC.

[8] De EU Cybersecurity Strategy bevat weliswaar een verwijzing naar het gebruik van AI: “The Commission also proposes to launch a network of Security Operations Centres across the EU, powered by artificial intelligence (AI), which will constitute a real ‘cybersecurity shield' for the EU …”

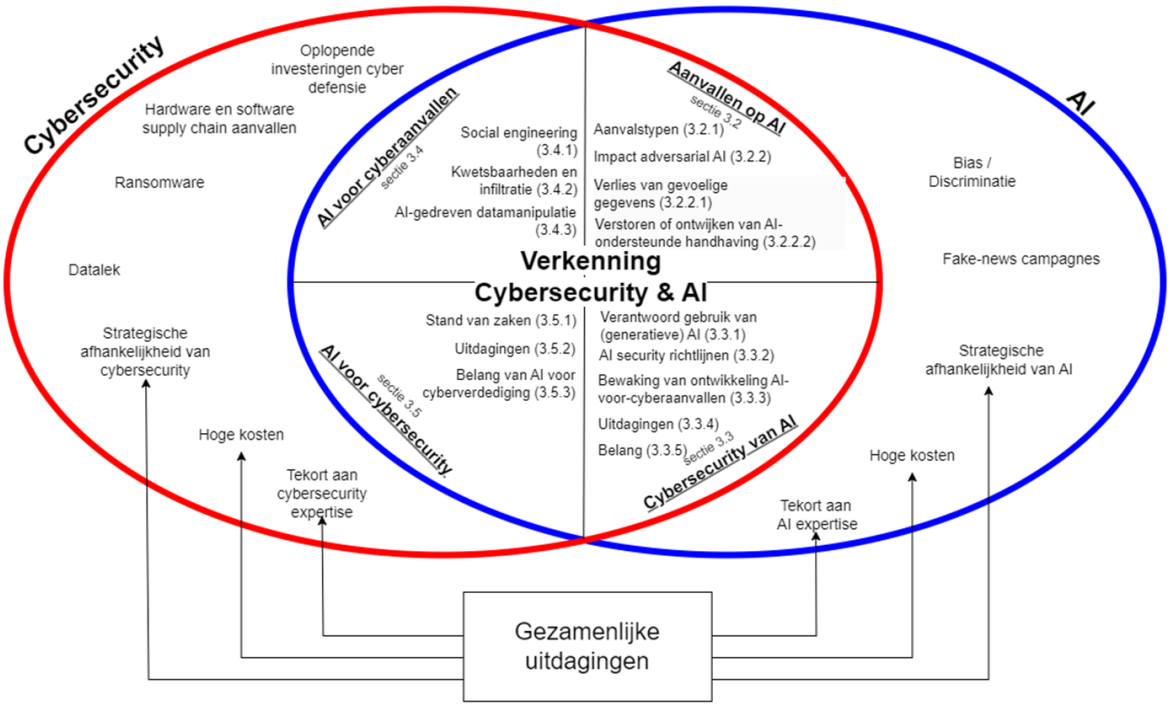
3 Raakvlak tussen cybersecurity en AI

Zowel het expertisegebied cybersecurity, als AI zijn sterk in beweging en kennen een breed scala aan toepassingsmogelijkheden en ook uitdagingen. Een deel van die mogelijkheden en uitdagingen zijn relevant voor het ene expertise gebied, maar niet voor beiden. Denk hierbij aan met AI gecreëerde fake-news campagnes voor beïnvloeding van verkiezingen, potentieel verlies van vertrouwelijke gegevens door cryptografische post-quantum uitdaging of discriminatie ten gevolge van bias in AI. Elk van deze uitdagingen zijn relevant binnen hun cybersecurity of AI-expertisegebied, maar spelen in het andere expertisegebied nauwelijks een rol. Deze uitdagingen vallen buiten de afbakening van dit onderzoek.

Een ander deel van de uitdagingen is relevant in zowel het cybersecurity als het AI-expertisegebied, maar zijn niet direct aan elkaar gerelateerd. Voorbeelden hiervan zijn schaarste aan professionals en problematiek ten aanzien van afhankelijkheid van Amerikaanse leveranciers. Deze uitdagingen doen zich zowel op het gebied van cybersecurity als op het gebied van AI voor, maar beïnvloeden elkaar slechts in beperkte mate. Ook is niet het geval dat oplossingen voor een uitdaging in één van de gebieden biedt niet direct een oplossing voor de uitdaging in het andere expertisegebied.

Tenslotte zijn er ook toepassingsmogelijkheden en uitdagingen die wel een behoorlijke wisselwerking kenen tussen beide gebieden. Bijvoorbeeld, toepassingsmogelijkheden van AI die een cybersecurityoplossing of -uitdaging vormen, of cybersecuritykwetsbaarheden die *trustworthy* AI ondermijnen. Het zijn deze toepassingsmogelijkheden en uitdagingen die in dit onderzoek centraal staan, omdat ze het raakvlak van cybersecurity en AI vormen.

In Figuur 7 zijn voorbeelden weergegeven van onderwerpen die in het raakvlak van cybersecurity & AI vallen, en onderwerpen die daaraan raken of daarbuiten vallen. De voorbeelden in het raakvlak geven ook de inhoudsopgave van dit hoofdstuk weer.



Figuur 7: Venndiagram ter illustratie van het cybersecurity & AI-raakvlak. Werk van de auteurs.

3.1 Verkenning raakvlak cybersecurity en AI

Het raakvlak tussen cybersecurity en AI is onder te verdelen in meerdere vraagstukken. Deze vraagstukken zijn als volgt te categoriseren:

1. Aanvallen op AI: Adversarial AI
2. Cybersecurity van AI-toepassingen
3. AI voor cyberaanvallen - complexe automatisering
4. AI voor cybersecurity
5. Gezamenlijke uitdagingen.

Deze categorieën zijn geïnspireerd door de onderverdeling van ENISA [1], met als aanvulling het derde punt: AI voor cyberaanvallen. Het vijfde punt complementeert het 'cybersecurity en AI-raakvlak kwadrant' met uitdagingen die zich zowel in cybersecurity als in AI voordoen.

In de volgorde zoals hierboven opgesomd worden deze security & AI categorieën in het vervolg van deze sectie uitgewerkt. Voor elk van de categorieën wordt een toelichting van de categorie en de huidige stand van zaken aan de hand van concrete voorbeelden. Ook wordt een indicatie gegeven van de impact op publieke belangen, alsmede een blik op verwachte toekomstige ontwikkelingen.

3.2 Adversarial AI

Het scala aan mogelijke cyberaanvallen op AI-toepassingen is groot en omvat aanvallen die op traditionele digitale toepassingen uitgevoerd kunnen worden. Immers, AI is in wezen ook een digitale toepassing met grotendeels vergelijkbare implementatieonderdelen als traditionele ICT (bijvoorbeeld internettoegang, gebruikersautorisatie, cloud-gebaseerde implementatie, etc.). Deze implementatieonderdelen kennen dezelfde kwetsbaarheden en

leiden onder dezelfde cyberaanvallen (bijvoorbeeld DDoS, digitale identiteitsfraude, etc.). In de context van deze verkenning worden deze traditionele aanvallen op AI verder buiten beschouwing gelaten en wordt ingezoomd op AI-specifieke aanvallen: adversarial AI.

Met adversarial AI wordt gedoeld op het doelbewust misleiden, manipuleren of uitschakelen van een AI-product: het AI-product werkt incorrect of helemaal niet. Een gangbare toelichting van het adversarial AI-werkveld bestaat uit een onderverdeling op basis van verschillende typen adversarial AI-aanvallen. De meeste typen aanvallen kunnen gericht zijn tegen zowel voorspellende als generatieve AI, maar sommige aanvallen zijn specifiek gericht op generatieve AI. De typen adversarial AI worden beschreven in sectie 3.2.1.

Voor beleidsmakers is de *manier waarop* adversarial AI wordt toegepast vaak minder relevant dan de *impact* die een aanval kan hebben en de *handelingsperspectieven* ter voorkoming of beheersing van de impact. De potentiële impact van adversarial AI wordt beschreven in sectie 3.2.2. Ook wordt daarbij ingegaan op mogelijke toekomstige effecten ten gevolge van adversarial AI-ontwikkelingen.

3.2.1 Aanvalstypen

3.2.1.1 Aanvallen op voorspellende AI

Adversarial AI richt zich op het AI-deel van een product en is tot op zekere hoogte toepassings-agnostisch: het kan zich uiten in veel verschillende gebieden waar AI wordt ingezet. Bij adversarial AI voor voorspellende AI worden over het algemeen vijf typen aanvallen onderkend: *poisoning* aanvallen, *input* of *evasion* aanvallen, *backdoor* aanvallen, *model stealing* of *model reverse engineering* aanvallen en tot slot *inference & inversion* aanvallen[9,10]. Eerst volgt een korte beschrijving en een voorbeeld bij elk type aanval, waarna dieper wordt ingegaan op bekende aanvallen, mogelijke scenario's en wat mogelijk is in de nabije toekomst.

Bij een *poisoning* aanval wordt het AI-systeem vergiftigd doordat een aanvaller aanpassingen aan de trainingsdata doet, waardoor het AI-systeem fouten gaat maken. Bijvoorbeeld een spamfilter die getraind is op gemanipuleerde data, en zo toch nog bepaalde spam e-mails doorlaat.

Bij een *input*- of *evasion* aanval voegt een aanvaller hele kleine bewerkingen toe aan een input zodat een AI-systeem wordt misleid: het trekt een foute conclusie. Een voorbeeld hiervan is het plakken van een gele post-it op een stopbord, waardoor een auto met AI-gebaseerde omgevingsherkenning het bord niet meer goed kan herkennen en zijn snelheid aanpast.

Bij een *backdoor* aanval is een 'achterdeurtje' in een AI-model aanwezig: de aanvaller heeft toegang tot het AI-model en kan het manipuleren. Door een speciale sleutel mee te geven bij een input, kan de aanvaller beslissen wat voor output het AI-model teruggeeft. Een voorbeeld hiervan is dat een model bedoeld voor het herkennen van nummerborden van auto's van criminelen, door een speciale toevoeging de nummerborden van specifieke

[9] [29] en [103].

[10] Deze vijf typen aanvallen grijpen een AI-product aan in de verschillende fases van hun life cycle. Poisoning en backdoor aanvallen hebben betrekking op de voorbereidingsfase waarin het model wordt ontworpen en data wordt verkregen en gecureerd. Zowel input- en evasion als inference aanvallen worden in de inputfase van een operationeel model uitgevoerd. Model stealing en model reverse engineering worden uitgevoerd op de gekoppelde input- en outputfase van een model.

criminelen niet meer herkent. En ze dus ongestoord door het herkenningssysteem heen blijven komen [29].

Bij *model stealing* of *model reverse engineering* brengt een aanvaller in kaart hoe een AI-model exact in elkaar zit. Hierdoor kan een aanvaller een *trade secret* of intellectueel eigendom achterhalen. Ook kan het opnieuw opgebouwde model worden gebruikt om kwetsbaarheden te vinden of een tegenreactie te bedenken.

Met *inversion* of *inference* aanvallen kan een aanvaller achterhalen wat voor (mogelijk geheime) trainingsdata is gebruikt. Zo kunnen gevoelige informatie worden blootgelegd, waaronder privacygevoelige gegevens en intellectueel eigendom.

3.2.1.2 Aanvallen op generatieve AI

In aanvulling op de kwetsbaarheden die zijn beschreven voor voorspellende AI heeft generatieve AI te maken met additionele kwetsbaarheden. Dit zijn directe en indirecte prompt injectie. Bij prompt injectie gebruikt de aanvaller input waardoor een LLM of chatbot onbedoelde acties uitvoert of acties waarvoor geen toestemming is [30]. Met indirecte prompt injectie kan hetzelfde worden bereikt, maar is geen directe interactie met een LLM of chatbot nodig. Deze aanvallen kunnen leiden tot onbeschikbaarheid, aangetaste integriteit, het vrijgeven van persoonsgegevens en misbruik [31].

3.2.2 Impact Adversarial AI

Bovenstaande adversarial AI-methodes kunnen verschillende impact hebben. Veelal is de impact gerelateerd aan het verlies van gevoelige gegevens of aan het verstoren of ontwijken van AI-functionaliteit.

3.2.2.1 Verlies van gevoelige gegevens

Bij aanvallen op AI-systemen kunnen gevoelige gegevens vrijkomen, wat aanzienlijke gevolgen kan hebben voor individuen en organisaties. Twee soorten gegevenslekken worden onderscheiden: de trainingsdata waarop het model is getraind en data van gekoppelde systemen. Trainingsdata vormen de kern van AI-modellen en kunnen verschillende soorten gevoelige informatie bevatten zoals persoonsgegevens en intellectueel eigendom. Anderzijds kan er data van gekoppelde systemen lekken. AI-systemen worden vaak geïntegreerd met andere systemen, wat de potentiële reikwijdte van gegevenslekken vergroot. Zowel trainingsdata als gekoppelde systeemdata kunnen gevoelige persoonsgegevens of intellectueel eigendom bevatten[11].

Het lekken van gevoelige gegevens kan leiden tot verschillende problemen. Ten eerste is er kans op financieel verlies voor getroffen organisaties. Het lekken van de data zorgt voor directe financiële schade door fraude, diefstal of de noodzaak om systemen te herstellen en beveiligingsmaatregelen te verbeteren. Ook kan verlies van gevoelige persoonsgegevens leiden tot overtreding van de Algemene Verordening Gegevensbescherming (AVG), en eventuele sancties die daaruit kunnen volgen. Daarnaast is er de mogelijkheid tot reputatieschade, zoals vertrouwensverlies bij klanten, partners en het grote publiek, wat kan leiden tot verlies van marktwaarde en verdienvermogen.

[11] Gevoelige persoonsgegevens zijn gegevens die kunnen worden gebruikt om een individu te identificeren, zoals namen, adressen en andere persoonlijke details. Onder intellectueel eigendom vallen bedrijfsgeheimen, onderzoeksgegevens en andere vormen van intellectueel eigendom.

Er zijn meerdere incidenten bekend waarbij ongewenst data is vrijgekomen door de toepassing of ontwikkeling van een AI-model:

- In december 2022 werd bekend dat de PyTorch-nightly versie was gecompromitteerd door een supply chain aanval [32]. Een kwaadwillende actor had een schadelijke versie van de torchtriton package met dezelfde naam geüpload naar de Python Package Index (PyPI). Deze package werd geïnstalleerd door gebruikers die dachten de officiële versie te downloaden, wat leidde tot het lekken van gevoelige data zoals systeeminformatie, gebruikersgegevens, en configuratiebestanden via versleutelde DNS-queries naar een malafide domein. Hoewel de precieze impact onbekend is, werd de schadelijke package 3000 keer gedownload voordat het verwijderd werd [33].
- In 2023 ontdekten onderzoekers dat gebruikersinformatie kon lekken via de Large Language Model (LLM) van Bing Chat door een *prompt injectie* aanval [34]. Wanneer gebruikers toestemming gaven, kon Bing Chat andere geopende websites lezen. Aanvallers plaatsten geheime instructies op deze websites, waardoor Bing Chat gevoelige informatie zoals namen, adressen en bankgegevens doorgaf aan de aanvallers. Dit gebeurde door de gegenereerde tekst om te zetten in HTML-elementen zoals afbeeldingen, waarbij de URL's gegevens exfiltreerden zonder dat de gebruiker dit doorhad. Microsoft heeft sindsdien maatregelen genomen om dit probleem te verhelpen door onder andere het implementeren van strengere Content Security Policies.

3.2.2.1.1 Impact op publieke belangen

Momenteel zijn er geen gevallen bekend van incidenten van informatielekkage door AI-diensten die publieke belangen in gevaar hebben gebracht. Economische veiligheid[12] of het verdienvermogen zou in gevaar kunnen komen wanneer zeer gevoelige informatie wordt geüpload naar AI-diensten of wanneer deze diensten zijn gekoppeld aan systemen die zeer gevoelige informatie kunnen lekken. Denk hierbij bijvoorbeeld aan een ontwikkelaar van vertrouwelijke overheidsapplicaties die ChatGPT gebruikt als ondersteuning voor het genereren van code en diens code invoert. Als een aanvaller toegang kan verkrijgen tot deze chathistorie, dan is deze code toegankelijk voor de aanvaller en verkrijgt deze inzicht in de kwetsbaarheden van de overheidsapplicatie. De mogelijke gevolgen van informatielekkage via AI-diensten kan in de toekomst twee publieke belangen raken.

Ten eerste is de economische veiligheid in het geding, voornamelijk door het lekken van intellectueel eigendom dat centraal staat in het verdienvermogen van bedrijven. Daarbij is intellectueel eigendom belangrijk voor de economische stabiliteit van markten. Als de economische veiligheid te veel in het geding komt, kan het de bereidheid van bedrijven verminderen om nog te willen investeren in het doen van onderzoek. Hierbij wordt zowel de economische veiligheid als het intellectueel eigendom geraakt. Incidenten met dataverlies via AI kunnen leiden tot terughoudendheid van het gebruik van AI (voor diverse doeleinden, waaronder ook onderzoek), wat op termijn het duurzame verdienvermogen kan ondermijnen.

Ten tweede heeft het lekken van vertrouwelijke persoonsgegevens directe gevolgen op de privacy van burgers, wat een schending van grondrechten inhoudt.

[12] In dit rapport wordt gerefereerd aan 'economische veiligheid' als gedefinieerd in de leidraad geïntegreerde risicobeoordeling nationale veiligheid [100].

3.2.2.2 Verstoren of ontwijken van AI-functionaliteit

Een tweede type risico's van adversarial AI zijn toepassingen gericht op het verstoren of ontwijken van de functionaliteit van AI. Bijvoorbeeld voor cybersecuritytoepassingen van AI, zoals biometrische authenticatie voor fysieke of digitale toegang tot vertrouwelijke ruimten of gegevens, voor het ontwijken van malware detectie, of voor het misleiden van automatische nummerplaat herkenning. Ook niet moedwillige veroorzaakte, maar simpelweg falende AI, kunnen een risico opleveren hoewel dit in het algemeen niet onder de noemer adversarial AI wordt geplaatst.

De problemen die voortvloeien uit adversarial AI zijn zeer divers en afhankelijk van de toepassingen waarvoor de getroffen AI wordt ingezet. Veelal is adversarial AI gericht op het ontwijken van AI die wordt ingezet ten behoeve van bewaking, beveiliging en handhaving. Daarmee is adversarial AI vaak niet een probleem in zichzelf, maar meer een instrument in het verhullen van malafide praktijken of strategische doelen (zoals militaire systemen). Met de toename van AI voor dergelijke doeleinden (waaronder cybersecurity) neemt ook de toepassing van adversarial AI daartegen toe (zie ook [21][13]).

Er zijn veel illustratieve voorbeelden van adversarial AI ten behoeve van ontwijking. Een specifiek voorbeeld is de misleiding van een gezichtsherkenningsoftware door middel van speciaal ontworpen accessoires, zoals een adversarial AI-bril [35]. Hierbij was het mogelijk om de eigenlijke identiteit niet te laten vaststellen. Maar ook om als een specifieke andere persoon geïdentificeerd te worden. Als deze gezichtsherkenningssoftware wordt toegepast op automatische paspoortcontrole bij grensovergangen, zou dit in potentie kunnen leiden tot het inreizen of uitreizen van terroristen of criminelen. Verder is adversarial AI in opkomst als 'digitale camouflage' van militaire systemen [36].

3.2.2.2.1 Impact op publieke belangen

De gevolgen van niet correct functioneren van AI hangen af van de toepassing waarvoor de AI wordt gebruikt en de mate waarin de toepassing afhankelijk is van de AI-functionaliteit. Bijvoorbeeld, AI-algoritmen die onderdeel zijn van SPAM filters of media content aanbevelingssystemen (b.v. door Spotify of Netflix) worden volgens de *AI Act* geclassificeerd als laag-risico toepassingen. Het onjuist functioneren van dergelijke AI kan vervelend zijn, maar niet veel meer dan dat. De gevolgen van onjuist functionerende *high risk* AI zijn groter. Bijvoorbeeld, onjuist functionerende AI toegepast in zelfrijdende auto's heeft in de VS geleid tot tientallen fatale ongevallen [37]. Zelfs van AI-toepassingen in auto's met een beperktere mate van autonomie dan zelfrijdend kan foutief functioneren al tot levensgevaarlijke situaties leiden. Een bekend voorbeeld is 'phantom braking' door een Advanced Driving Assistance System (ADAS)[14] [37] dat standaard in veel hedendaagse auto's aanwezig is.

Een andere factor die van invloed is op de impact die foutief functionerende AI heeft, is de mate van het gebruik van de AI. In het voorgaande voorbeeld is het aantal fatale ongelukken met zelfrijdende auto's in de VS nog beperkt, omdat het gebruik van zelfrijdende auto's nog zeer beperkt wordt toegestaan. In Europa is deze AI-gedreven technologie nauwelijks toegestaan, waardoor de impact ervan nog geringer is dan die van 'phantom braking' waar iedere weggebruiker mee te maken kan krijgen[15]. Aangezien de opkomst van

[13] In ENISA's Threat Landscape wordt de term 'adversarial attack' gebruikt i.p.v. adversarial AI.
[14] 'Phantom breaking' is het fenomeen dat een auto die op cruise-control rijdt plotseling en voor de bestuurder op onverklaarbare wijze remt.
[15] Opmerking t.a.v. AI-toepassingen in automotive: hoewel dergelijke AI in de AI-act is aangemerkt als 'high risk', wordt de regulering (mede gericht op de beheersing van de risico's ervan) opgepakt in de sectorale automotive context.

AI in steeds meer toepassingen zijn intrede doet, zal een inschatting van de impact op publieke belangen vrij snel achterhaald kunnen zijn.

Praktijkvoorbeelden van negatieve AI verstorende effecten door specifieke adversarial AI aanvallen zijn nog beperkt. Hoewel ontwijking van gezichtsherkenning- of biometrische identificatiesoftware en van AI-gedreven militaire observatietechnieken inzetbaar is, blijkt dit in de praktijk lastig realiseerbaar [36]. Maar ook hier geldt dat voor enkele gevallen waarin adversarial wel succesvol wordt toegepast, het niet is uit te sluiten dat die resulteren in een serieuze impact.

Zoals voorgaande voorbeelden aangeven kan verstoorde of ontweken AI publieke belangen raken, zoals fysieke en digitale veiligheid en handhaving van de openbare orde. De gevolgschade door verstorende / ontwijkende effecten van adversarial AI kunnen bovendien nog meer typen publieke belangen raken, zoals maatschappelijke onrust of economische schade.

3.3 Cybersecurity van AI

Veiligheid van AI wordt met het toenemende gebruik van AI steeds belangrijker[16]. Het onderwerp vergt ook specifieke aandacht, omdat traditionele cybersecurityrichtlijnen ontoereikend zijn voor het ontwikkelen, toetsen en gebruiken van AI. ENISA (The European Union Agency for Cybersecurity) benoemt bijvoorbeeld het risico op ontraceerbaarheid van data en AI-componenten, en geeft aan dat voor sommige vormen van AI geheel andere testprocedures nodig zijn ten opzichte van traditionele IT-systemen [1]. Daarom is ontwikkeling van nieuwe, AI-specifieke richtlijnen voor het gebruik, de ontwikkeling, en de toetsing van cybersecurity aspecten van AI op gang gekomen. In deze sectie worden ontwikkelingen geschetst (niet uitputtend) ten aanzien van het verantwoord gebruiken van AI en richtlijnen voor veilige ontwikkeling en evaluatie van AI. Tenslotte worden initiatieven geschetst gericht op het ‘bewaken’ van ontwikkelingen op het gebied van AI-voor-cyberaanvallen.

3.3.1 Verantwoord gebruik van (generatieve) AI

Sinds de introductie van Large Language Models (LLMs) zoals ChatGPT, Google Bard, Midjourney en DALL-E, vaardigen steeds meer organisaties richtlijnen uit voor verantwoord gebruik daarvan. De rationale daarachter komt deels voort uit cybersecurity gerelateerde aspecten. Verschillende organisaties uit diverse sectoren hebben restricties ingevoerd of zelfs een compleet verbod vastgesteld op het gebruik van bepaalde LLMs. Enkele prominente voorbeelden zijn Amazon, Apple en Samsung:

- Apple heeft recentelijk het gebruik van ChatGPT en andere AI-tools door werknemers beperkt [38]. Dit besluit komt voort uit bezorgdheid over de mogelijke lekken van vertrouwelijke bedrijfsinformatie en de ontwikkeling van hun eigen AI-technologie. Apple wil voorkomen dat gevoelige data die door werknemers wordt ingevoerd, wordt gebruikt om de modellen verder te trainen en mogelijk opnieuw naar buiten komt als antwoord op vragen van andere gebruikers.
- Samsung heeft een algeheel verbod ingesteld op het gebruik van generatieve AI-tools zoals ChatGPT [39]. Dit besluit volgde nadat interne, gevoelige gegevens per

[16] Opmerking: in publicaties, zoals [11], wordt vaak de bredere term “Trustworthy AI” gebruikt. Onder trustworthy AI wordt volgens ISO’s AI terminologie verstaan (zie sectie 5.15 in [5]): AI-robuustheid, betrouwbaarheid, weerbaarheid, controleerbaarheid, verklaarbaarheid, voorspelbaarheid, transparantie en ‘bias and fairness’. De term trustworthy AI is gerelateerd aan cybersecurity van AI met deels (doch niet geheel) overlappende betekenis, maar rijkt ook grootdeels buiten het raakvlak van cybersecurity en AI. Daarom wordt deze term in deze verkenning verder buiten beschouwing gelaten.

ongeluk waren gelekt naar ChatGPT. Samsung heeft aangegeven dat gegevens die naar dergelijke AI-tools worden gestuurd, vaak worden opgeslagen op externe servers, wat problemen kan veroorzaken met betrekking tot toegang, verwijdering en onbedoelde deling van deze gegevens.

- Amazon heeft eveneens het gebruik van ChatGPT door hun medewerkers beperkt [38]. Dit besluit is genomen om te voorkomen dat bedrijfsgevoelige informatie wordt ingevoerd in de AI-modellen, wat kan leiden tot potentiële datalekken en misbruik van vertrouwelijke informatie. Daarnaast heeft Amazon een waarschuwing uitgegeven aan hun werknemers om voorzichtig te zijn met het gebruik van door derden gegenereerde AI-inhoud voor werk gerelateerde doeleinden.
- Naast deze prominente voorbeelden zijn ook Nederlandse bedrijven [40] en overheidsinstanties [41] met dergelijke richtlijnen aan de slag. Ook in deze voorbeelden wordt aangedrongen op verantwoord gebruik d.m.v. terughoudendheid om LLMs te gebruiken als daar gevoelige gegevens voor nodig zijn. Tevens wordt benadrukt dat LLMs slechts als hulpmiddel gebruikt kunnen worden: er kan niet vanuit worden gegaan dat antwoorden juist zijn en de verklaarbaarheid van de resultaten is beperkt.

Naast restricties t.a.v. het gebruik van LLMs vanuit organisaties, publiceren de aanbieders van LLMs ook richtlijnen voor verantwoord gebruik van hun AI-diensten [42, 43]. Grofweg worden daarin vrij evidente 'gebruiksvoorwaarden' aangegeven, namelijk dat de gebruiker (a) zich moet houden aan geldende wetgeving, (b) geen content mag genereren om te gebruiken voor kwaadwillige activiteiten, en (c) dat de LLM-dienst niet gehackt mag worden.

Vanuit de rijksoverheid is de 'overheidsbrede visie generatieve AI' opgesteld [44]. In deze visie (en in het daarin aangegeven "bestaande beleid") wordt het raakvlak met cybersecurity beperkt geadresseerd. Zorgen over gegevensbescherming en veilige AI[17] worden in de visie benoemd, maar verder komt cybersecurity van generatieve AI niet aan bod. Dit beeld komt overeen met de tot voor kort beperkte aandacht die aan cybersecurity werd besteed vanuit AI-initiatieven zoals NL-AIC, APPL.ai, AiNed, NOLAI en ICAI. Deze initiatieven richten zich in eerste instantie op (het stimuleren van) innovaties m.b.v. AI.

Al met al is beleid dat is gericht op verantwoord gebruik van (generatieve) AI nog beperkt tot specifieke organisatie-interne maatregelen, waarin cybersecurityoverwegingen een beperkte rol spelen. De initiatieven die genomen worden zijn gebaseerd op de risicoinschatting van individuele organisaties over gebruik van AI. Om de juiste risicoinschatting te kunnen maken is kennis en ervaring met AI-gebruik nodig, die nog beperkt is bij veel organisaties en individuen. Het delen van kennis en ervaringen is nuttig om te komen tot een uniformer richtlijn voor verantwoord gebruik van AI.

3.3.2 AI-security richtlijnen: ontwikkeling en evaluatie

Om ervoor te zorgen dat de AI-toepassingen minder kwetsbaar zijn voor adversarial AI-attacks, zijn er door verschillende organisaties, zoals het ministerie van Binnenlandse Zaken en Koninkrijksrelaties (AIVD), ENISA, NCSC-UK en BSI, richtlijnen voor de ontwikkeling van systemen met AI-componenten opgesteld. Ter illustratie: de AIVD heeft de folder "AI-systemen: ontwikkel ze veilig" uitgebracht, die mede gebaseerd is op onderzoek van TNO [29]. Hierin zijn een vijftal algemene richtlijnen opgenomen voor AI- en cybersecurityexperts bij het ontwikkelen van AI-applicaties. Deze richtlijnen zijn gerelateerd aan de soorten

[17] Hierbij wordt onder 'veilige AI' verstaan dat ingezette AI in bijvoorbeeld de zorg veilig is voor de patient, of AI in zelfrijdende auto's veilig is voor de inzittenden.

aanvallen op AI-systemen (zoals besproken in 3.2.1) en deze zijn: houd je datakwaliteit op orde, zorg voor validatie van je data, houd rekening met supply chain security, maak je model robuust tegen aanvallen, en zorg dat je model controleerbaar (auditable) is.

Naast richtlijnen voor het veilig ontwikkelen van AI is gestart met het opstellen van richtlijnen voor het evalueren van de veiligheid van AI, hoewel nog weinig hanteerbare richtlijnen zijn gepubliceerd. Onder andere in onderzoeksprogramma's van TNO wordt een AI-security normenkader opgesteld en getoetst[18]. Ook worden voor het uitvoeren van dergelijke AI-security evaluaties software tools ontwikkeld voor automatisering van evaluatie en het monitoren van AI-security. In deze onderzoeken wordt ook Europese samenwerking gezocht met bijvoorbeeld het Duitse Bundesambt für Sicherheit in der Informationstechnik (BSI), die enkele fundamentele uitgangspunten over security van AI hebben gepubliceerd [45]. In de VS heeft MITRE recent het AI Assurance and Discovery Lab opgericht. Dit lab richt zich op het ontdekken, beoordelen en mitigeren van risico's in AI-systemen gedurende de levensloop van AI-producten [46].

3.3.3 Bewaking van ontwikkeling AI-voor-cyberaanvallen

Het werkveld van AI voor cyberaanvallen is sterk in beweging (zie sectie 3.4). Om grip te houden op deze ontwikkeling (en op AI-ontwikkelingen in breder zin) zijn inmiddels een aantal methoden opgesteld, waaronder Signposts of Change, Frontier Safety Framework en AI Observatory.

3.3.3.1 Signposts of Change

In het speelveld van AI voor cyberaanvallen spelen een aantal factoren die een richting van de ontwikkelingen aangeven. Door toepassing van de Signpost of Change-methode (wegwijzer van verandering) kunnen belangrijke ontwikkelingen op waarde worden geschat. In deze methode worden waarneembare gebeurtenissen of ontwikkelingen beschreven zodat trends of veranderingen die aandacht behoeven tijdig worden onderkend. Hiervoor wordt periodiek naar de Signposts of Change gekeken [47].

De momenteel te onderkennen factoren zijn:

- Snelheid: ontwikkelingen volgen elkaar in rap tempo op. In een tijdsbestek van enkele maanden konden LLM-agenten eerst enkel bestaande kwetsbaarheden in specifieke omgevingen uitbuiten [48]; inmiddels kunnen ze in teams samenwerken, dit ook in echte systemen uitbuiten en zelfs al kwetsbaarheden ontdekken [49, 50, 51].
- Kosten: wetenschappers verwachten een kostendaling van 95% voor het inzetten van social engineering in de komende paar jaar [52]. Het inzetten van AI-agenten voor het exploiteren van bekende kwetsbaarheden kost nu in de orde van enkele tientallen euro's, waarbij een verdere kostendaling van een factor drie tot zes binnen twee jaar wordt verwacht [51].
- Opschalen: het opschalen van complexe aanvallen betekent nu meestal een vergroting van het personeelsbestand. Extra AI-agenten kunnen eenvoudig aangeroepen worden waarmee een drempel voor het opschalen van cyberaanvallen weggenomen wordt [51].
- Asymmetrie: in phishing geeft het toepassen van LLMs met name de aanvaller een voordeel, gezien hierbij mensen in plaats van techniek het doelwit zijn [52]. De asymmetrie tussen aanvallers en verdedigers kan verder worden vergroot doordat

[18] Het AI-security normenkader kan gezien worden als een inkadering van bredere kaders, zoals het 'algoritmekader' dat door het ministerie van Binnenlandse Zaken wordt opgesteld [111].

de schaalbaarheid van met name aanvallen en de verwachte snelle daling van kosten. Buiten phishing zou de focus moeten liggen op het identificeren van kwetsbare doelwitten voor specifieke situaties [53].

3.3.3.2 Frontier Safety Framework

Google heeft in dezelfde trant het Frontier Safety Framework opgezet [54]. Het idee hierachter is dat steeds meer geavanceerde AI-modellen risico's kunnen vormen op het gebied van veiligheid en privacy. In dit framework kijkt men naar de mogelijkheden van huidige state-of-the-art AI-modellen en onderzoekt men of deze modellen een mogelijk gevaar kunnen vormen mocht deze ontwikkeling doorgaan. Hiervoor stelt het framework een "Critical Capability Level" voor; een niveau waarop AI-modellen een groter risico vormen.

Deze Critical Capability Levels worden vastgesteld door onderzoekers bij Google door te kijken naar domeinen in de industrie met hoge risico's. Er wordt dan bekeken hoe een AI-model hierin schade kan aanrichten, en welke set van mogelijkheden een AI minimaal moet kunnen beheren om dit uit te voeren. Dit is al gedaan voor de domeinen Autonomy, Biosecurity, Cybersecurity, en Machine Learning R&D.

Als deze niveaus zijn vastgesteld, gaat men kijken naar hoe ver een AI-model al is richting dit niveau. Hierbij wordt gekeken naar welke mogelijkheden en vaardigheden het AI-model al bezit. Dit wordt elke drie maanden gedaan, en met behulp van de Deep Learning Scaling Law [55].

Wanneer een model te snel ontwikkelt en het richting een Critical Capability Level gaat, wordt de ontwikkeling stopgezet. Dit is niet altijd mogelijk voor iedereen; Google kan hun eigen ontwikkeling wel stopzetten, maar niet die van anderen. In het bijzonder kunnen activiteiten van criminelen en buitenlandse, statelijke actoren slechts in beperkte mate tegengehouden worden.

Eenzelfde raamwerk is ontwikkeld door OpenAI: het preparedness framework [56]. Daarin worden ook specifieke niveaus gedefinieerd voor de inschatting van cybersecurity risico's voortkomend uit de door OpenAI ontwikkelde generatieve AI-diensten.

Het voordeel van raamwerken zoals het Frontier Safety Framework en Preparedness Framework ten opzichte van Signposts of Change is dat het ruimte en tijd geeft om in te grijpen voordat het te laat is. Terwijl de Signposts of Change voornamelijk aangeven dat een verandering heeft plaatsgevonden, gekenmerkt door symptomen, gaat het Frontier Safety Framework uit van een proactieve respons. De toepassing van de raamwerken op te lanceren generatieve AI-diensten vergt echter wel toegang daartoe voordat deze publiek worden gemaakt. Daarom is toetsing van te lanceren generatieve AI-diensten wel voorbehouden aan de leveranciers daarvan (zoals Google en OpenAI).

3.3.3.3 AI Observatory

Gladstone, een AI-bedrijf in de V.S., stelt voor een AI Observatory te maken [57]. Dit zal dan worden uitgevoerd door het DHS (Department of Homeland Security). De voorgestelde functies van dit AI Observatory zijn:

- Horizon Scanning: hierin wordt gekeken naar huidige ontwikkelingen op het gebied van AI. Men kijkt hiervoor naar openbaar onderzoek, commercieel beschikbare datasets, discussies met onderzoekers, cybersecurity and Infrastructure Security

Agency (CISA) rapporten, en eigen evaluaties van publiekelijk toegankelijke AI-systemen.
- Emergency Preparedness: het creëren van responsplannen voor incidenten en investeringen in DHS Science and Technology (S&T) AI Safety and Security prioriteit te geven.
- Information-sharing and coordination: het delen van informatie en samenwerken met andere departementen en groepen binnen de overheid.

Gladstone's suggestie voor de uit voeren activiteiten in een AI Observatory omvat (zie bijlage H in [57]):

1. Volg geavanceerde AI-startups en organisaties wanneer deze bekend worden via publiekelijk beschikbaar nieuws. Hiermee wordt eveneens gekeken naar grote investeringen en aankopen.
2. Reviewen van literatuur en publicaties op het gebied van AI om belangrijke trends en veranderingen te identificeren.
3. Het creëren van een lijst van vooraanstaande AI-onderzoekers. Hierbij wordt de vergelijking gemaakt met The Bulletin of Atomic Scientists [58], maar dan op het gebied van AI.
4. In samenwerking met de SEC (Securites and Exchange Commission) het volgen van AI-ontwikkelingen in de financiële sector, met name hedge funds. Volgens Gladstone spelen hedge funds een belangrijke rol in de ontwikkeling van "catastrofale" AI.
5. Samenwerken met open-source experts en alternatieve dataleveranciers om internationaal een beeld te kunnen schetsen van AI-ontwikkelingen.
6. Een relatie opbouwen met de open-source AI-gemeenschap. Voorbeelden hiervan zijn: Eleuther AI [59], BigScience [60], Hugging Face [61], Together [62], en Ontocord [63].
7. Het identificeren van belangrijke knooppunten in AI-ontwikkeling: onderzoekers, academische en industriële instituten, gediplomeerden in relevante disciplines or onderzoek doen naar AI in universiteiten, publicatie clusters, GPU-aandelen in verschillende geografische gebieden, etc.
8. Het evalueren van publiekelijk beschikbare AI-systemen. Zowel open-access systemen, als systemen die via ene API benaderbaar zijn (bijv. OpenAI's GPT-4). Hierbij maakt Gladstone een vergelijking met onderzoekers op het gebied van massavernietigingswapens die kunnen helpen om te begrijpen of de ontwikkeling van AI dezelfde impact zou kunnen hebben.
9. Publiceer deze resultaten regelmatig in een "Global AI Risk Report"; een globaal AI-risico rapport.

Wanneer men deze methoden in combinatie gebruikt kan het helpen om te identificeren op welke manieren een AI-model een gevaar kan vormen, waar dit model aan moet voldoen, hoeveel tijd men heeft om hier iets aan te doen, en om bij te kunnen houden welke veranderingen hebben plaatsgevonden en dus al een risico vormen.

3.3.4 Uitdagingen

Samenvattend kan gesteld worden dat cybersecurity van AI een relatief nieuw expertisegebied is. Er zijn initiatieven gestart voor de ontwikkeling van nieuwe, AI-specifieke richtlijnen voor het veilige gebruik, veilige ontwikkeling en evaluatie van AI-security, maar continuering van die initiatieven is nodig om tot een volwassen niveau van cybersecurity van AI te komen. ENISA benadrukt hierbij de behoefte aan de ontwikkeling van gestandaardiseerde evaluatiekaders (en -tooling) voor het behoud van privacy,

vertrouwelijkheid van informatie(stromen), en het ontwerpen van AI-systemen. Tevens wordt aangegeven dat er behoefte bestaat aan een 'observatorium' voor AI en cyberdreigingen daarvoor en daarvan [2] (zie ook het initiatief voor het inrichten van een AI observatory vanuit de VS beschreven in voorgaande sectie).

3.3.5 Belang van cybersecurity van AI

Op soortgelijke wijze als met de introductie van eerdere digitale producten en technieken (zoals e-commerce en de digitale overheid) gaat het belang van cybersecurity gepaard met toenemend gebruik van AI en de economische en maatschappelijke waarde daarvan. Ook voor cybersecurity van AI blijkt dit het geval. Dit roept de vraag op of het gebruik van AI inmiddels dermate substantieel is dat ook de cybersecurity aspecten ervan zwaarder benadrukt moeten worden, en voor welk type gebruik van AI dat geldt.

Onderzoeken hiernaar geven een verschillend beeld. Vanuit onderzoeken in 2019, 2021 en 2024 naar het gebruik van AI door Nederlandse overheidsinstanties blijkt dat er ruim anderhalf keer meer AI-toepassingen zijn gevonden in de afgelopen vijf jaar [64]. Ander internationaal onderzoek van Reuters Institute en de universiteit van Oxford geeft aan dat de verwachtingen van onlinegebruikers over generatieve AI-toepassingen voor nieuwsgaring op een termijn van vijf jaar hoog zijn, maar dat het huidige gebruik van deze diensten vooralsnog zeer beperkt is [12]. Sommige critici trekken zelfs de toekomstige potentie van generatieve AI voor professioneel gebruik in twijfel [65, 66].

In welke mate het gebruik en (economisch) belang van AI zich ook gaat ontwikkelen, het lijkt raadzaam om tijdig voorbereid te zijn op de uitdagingen van cybersecurity van AI. De snelheid van (cybersecurity aspecten van) AI-ontwikkelingen is dermate hoog dat een passieve houding ten aanzien van dit onderwerp kan leiden tot onwenselijke situaties waarin nauwelijks nog ingegrepen kan worden.

3.4 AI voor cyberaanvallen

Tot voor kort was de technologie om cyberaanvallen te automatiseren beperkt tot voorgeprogrammeerde handelingen in een voorspelbare omgeving. Een voorbeeld hiervan is het automatisch in kaart brengen van apparaten die verbonden zijn met het publieke internet, en het ontdekken van software op deze apparaten die luistert naar binnenkomende verbindingen[19]. Ook het gebruik maken van reeds bekende kwetsbaarheden in deze software om toegang te krijgen tot het systeem kan geautomatiseerd worden, net als het vervolgens versleutelen van de aangetroffen bestanden in het geval van ransomware.

AI biedt de mogelijkheid om automatisering van cyberaanvallen significant uit te breiden en de drempel voor uitvoering van cyberaanvallen te verlagen. Welke vorm deze automatisering aanneemt verschilt per type cyberaanval. In ENISA's Threat Landscape van 2023 [21] worden specifiek drie toepassingen van AI voor het uitvoeren van cyberaanvallen benoemd[20]. Deze betreffen AI ten behoeve van:

[19] Bijvoorbeeld via op internet beschikbare zoekmachines zoals [109].

[20] Er zijn publiek beschikbare databases die een veel breder overzicht van gerapporteerde AI-ondersteunde cyberaanvallen bevatten, zoals MITRE ATLAS [77] en MIT's AI Risk Repository [110]. In veel van die gerapporteerde cyberaanvallen speelt AI echter slechts een beperkte rol (of had de aanval ook geheel zonder AI uitgevoerd kunnen worden). Ook splitsen die overzichten vaak AI-ondersteunede cyberaanvallen op in meerder stappen, die gezamenlijk worden uitgevoerd om één soort aanvalsdoel te bereiken. Vanuit het perspectief van geraakte publieke belangen door AI-ondersteunde cyberaanvallen leidt dit tot een onoverzichtelijker beeld, dan er in ENISA's Threat Landscape geschetst wordt.

- Social engineering: het opstellen van overtuigender (spear-)phishing emails, deepfakes voor het impersonaliseren van stemgeluid en AI-gedreven data mining voor slachtofferselectie en -aanpak.
- Kwetsbaarheden & infiltratie: het vinden van kwetsbaarheden in een aan te vallen systeem en het infiltreren daarvan.
- AI-gedreven manipulatie van informatie: het toevoegen, verwijderen of manipuleren van data om zo externe (cybersecurity) beslissingen te beïnvloeden of activiteiten te verbergen. In de context van het rapport wordt dit afgekaderd tot cybersecurity-gerelateerde datamanipulatie.

In deze sectie wordt uitgewerkt in welke mate AI bijdraagt aan de implementatie van deze drie categorieën van cyberaanvallen. Voor elk van deze aanvallen zijn de volgende situaties geschetst: de huidige stand van zaken en mogelijke toekomstige AI-toepassingen. Aan het eind van deze sectie wordt een overzicht gegeven van enkele methoden om ontwikkelingen op het gebied van AI-toepassingen (zoals AI voor cyberaanvallen) in de gaten te kunnen houden.

3.4.1 Social engineering

Social engineering is een van de gebieden waar AI nu al in toenemende mate een belangrijke rol aan het spelen is [67]. Middels deze technieken proberen cybercriminelen hun doelwitten financiële middelen en gevoelige informatie, zoals inloggegevens, afhandig te maken [68]. Een van de bekendste vormen van dit type aanvallen is phishing, wat naast de traditionele, ongerichte variant, meerdere sub-varianten kent die zich onderscheiden door bijvoorbeeld: het richten op specifieke doelwitten (spear phishing of whaling), het nabootsen van stemmen via bellen (vishing), en het klonen van de stem en beeld van een vertrouwde persoon (deepfakes). Deze kwaadaardige communicaties lijken vaak legitiem en kunnen logos en branding van bekende bedrijven gebruiken. De gewonnen informatie kan vervolgens worden gebruikt voor andere misdaden, zoals identiteitsdiefstal en financiële fraude. Dit type aanvallen is momenteel al een van de meest voorkomende en AI heeft de potentie om de drempel voor aspirerende cybercriminelen te verlagen en om aanvallen complexer te maken. Dit heeft de potentie om een significant gevaar te vormen, sinds deze aanvallen vaak gebruikt worden als aanvalsvector voor malware, die daardoor logischerwijs ook makkelijker uit te voeren worden.

3.4.1.1 Hedendaagse AI

Met geringe automatisering is het al mogelijk om op grote schaal een beperkte hoeveelheid online-informatie over personen of organisaties te verzamelen en deze informatie vervolgens in te zetten voor het ontwikkelen en uitvoeren van een gepersonaliseerde phishingaanval [69]. Deze strategie beperkte zich vroeger tot van tevoren gestructureerde informatie zoals de naam (via Facebook), het adres (via de KvK) en het beroep (via LinkedIn) van een persoon. Inmiddels is mogelijk om tijdens de informatieverzamelingsfase ook minder duidelijk gestructureerde informatie te betrekken, zoals de inrichting van een huis (via Instagram), de belevingen tijdens een vakantie (via Facebook), of andere recente ervaringen (via Twitter).

Daarnaast kan met behulp van LLMs makkelijker en vooral sneller een phishingbericht worden opgesteld. Aanvallers kunnen middels prompt injectie publiek beschikbare modellen gebruiken om een phishingbericht te genereren dat gepersonaliseerd is voor het doelwit, bijvoorbeeld door een specifieke organisatie te imiteren of door informatie te gebruiken die relevant is voor het doelwit. Sinds de introductie van ChatGPT is het aantal phishingberichten

meer dan verveertigvoudigd [70]. Hierbij speelt ook de trend van cybercrime-as-a-service een rol. Cybercriminelen bieden hun eigen modellen aan op de darkweb die gericht zijn op het uitvoeren van social engineering aanvallen en de gebruiker daarbij helpen en waaruit de veiligheidsmaatregelen van legitieme LLMs zijn verwijderd [71]. Daarna kunnen reacties van het doelwit automatisch verwerkt of beantwoord worden zonder tussenkomst van een persoon.

3.4.1.2 Toepassing toekomstige-AI

De verwachting is dat de kosten van kwalitatief hoogwaardige phishing met circa 95% zal dalen in de komende paar jaar waarbij de kwantiteit bovendien zal toenemen [52]. Hierbij bestaat de asymmetrie dat aanvallers AI (of specifieker: LLMs) kunnen inzetten voor het produceren van (spear) phishing-berichten maar dat de verdediging dit nauwelijks aan kan [52, 48]. Gezien de verwachte toename van berichten gecombineerd met de dalende kosten voor de aanvaller, genereert dit bovendien nog een extra asymmetrie: de verdediging moet meer mensen inzetten om meer berichten te controleren op (spear) phishing, waarbij de schaarste van experts bovendien de kosten van hen verhoogt. Waar de aanvallers een enorme kostendaling verwachten, zal de verdediging daarom nog duurder worden.

Er wordt automatisch aanvullende (verhalende) informatie verzameld om de kennis over het doelwit compleet te maken. Denk hierbij aan beschrijvingen van ervaringen die het doelwit mogelijk herkent, de menselijke karakteristieken en kwetsbaarheden van het doelwit, andere websites waarop informatie over de persoon of organisatie te vinden is. Dit gaat een stap verder dan de gestructureerde informatie die momenteel verkregen kan worden, zoals waar iemand werkt. Waarschijnlijk kunnen teams van LLM-agenten (zie 3.4.2.2) ingezet gaan worden om nog beter informatie te verzamelen, te combineren en te aggregeren. Vervolgens coördineert de aanvaller de phishingcampagne op een hoog abstractieniveau. De interactie met doelwitten verloopt volledig geautomatiseerd en via meerdere modaliteiten: email, telefoon en video. Het doelwit wordt over langere tijd automatisch van meerdere kanten misleid door ogenschijnlijk verschillende actoren. De aanvaller stuurt waar nodig bij door instructies te geven aan diens systeem.

3.4.2 Kwetsbaarheden en infiltratie

Met behulp van malware kan er via kwetsbare delen van digitale systemen worden binnengedrongen. Dit type kwaadaardige software is ontworpen om schade toe te brengen aan digitale apparatuur om onder andere gevoelige informatie te stelen, systeeminstellingen te wijzigen, ongewenste advertenties te weergeven, of kwaadwillenden directe toegang te geven tot het systeem. Het kan verschillende vormen aannemen en op verscheidene manieren verspreid worden. In het domein van kwetsbaarheden en infiltratie worden twee fases onderscheiden: (1) **ontwikkeling** van malware en (2) de **inzet** ervan om aanvallen uit te voeren. In de ontwikkelingsfase worden kwetsbaarheden gevonden in bestaande software (zero-days), en worden manieren ontwikkeld die deze kwetsbaarheden gebruikt om concrete doelen te behalen, bijvoorbeeld het versleutelen van aangetroffen bestanden in de context van een ransomware-aanval. In de inzetfase wordt deze malware ingezet om doelwitten daadwerkelijk te infecteren.

Over het algemeen wordt het vinden van een nieuwe kwetsbaarheid als moeilijker gezien dan het uitbuiten van die kwetsbaarheid. Ondersteuning door AI gaat makkelijker voor de minder complexe taken. Om die reden wordt eerst fase (2) inzet behandeld, en daarna fase (1) ontwikkeling, omdat dit logischer leest.

3.4.2.1 Exploitatie van kwetsbaarheid

3.4.2.1.1 Hedendaagse AI

De automatisering van de inzet van malware is sterk afhankelijk van de voorspelbaarheid van de doelwitomgeving. Voor automatisering moet de aanvaller hier noodzakelijkerwijs van tevoren aannames over doen. Als deze aannames niet blijken te kloppen kan dit de voortgang van de aanval belemmeren of bemoeilijken, of leiden tot detectie van een aanval. Vanuit de aanvaller gezien is een update vereist om het gedrag van de malware te veranderen om toch succesvol te zijn.

Begin 2024 werd een eerste *LLM-agent* [21] ontwikkeld die zelfstandig systemen kon hacken [49, 72, 73]. Aan het begin van het tweede kwartaal van 2024 werd duidelijk dat LLM-agenten gebruik kunnen maken van *one-day vulnerabilities,* terwijl de bestaande kwetsbaarhedenscanners[22] deze kwetsbaarheden niet konden vinden. Hierbij dient te worden opgemerkt dat deze agenten énkel succesvol zijn als ze gebruik maken van de meest geavanceerde LLMs, en de LLMs toegang hadden tot de beschrijving van de kwetsbaarheid[23]: dan konden de agenten in veruit de meeste gevallen de kwetsbaarheid uitbuiten (>85%) [51].

3.4.2.1.2 Toepassing toekomstige-AI

De aanvaller coördineert de inzet van malware op een hoog abstractieniveau. In het geval van ransomware verloopt de afhandeling van reacties van slachtoffers en eventuele onderhandelingen automatisch op basis van instructies van de aanvaller. De kosten van de inzet van AI voor het exploiteren van kwetsbaarheden daalt naar verwachting binnen 2 jaar met een factor 3 tot 6 [51].

Naast het laagdrempelig worden van het exploiteren van kwetsbaarheden leent een hoge mate van automatisering zich ook om voortdurend kleine aanpassingen in de malware aan te brengen. Dergelijke snel ‘muterende’ malware verkleind de kans op detectie.

3.4.2.2 Ontwikkeling van nieuwe kwetsbaarheden

3.4.2.2.1 Hedendaagse AI

Tot circa een jaar geleden gold dat nieuwe kwetsbaarheden handmatig werden gevonden, en de software om hier misbruik van te maken tevens handmatig werd geschreven. De expertise om kwetsbaarheden tegen waardevolle doelwitten in te zetten is zeldzaam en dus kostbaar.

Het afgelopen jaar zijn er echter ontwikkelingen die dit veranderden. In juni 2024 werden er teams van LLM-agenten ingezet, waarbij verschillende agenten samenwerkten en daarbij een andere rol vervulden, van planner tot manager en specialistische agenten. Door de inzet van deze specialistische agenten werd niet alleen de slagingskans van aanvallen verder

[21] Met een LLM-agent wordt hier bedoeld ‘een systeem dat met een LLM kan redeneren, een plan kan maken én uitvoeren, eventueel door het aanspreken van externe tools’ [49].
[22] Metasploit, ZAP
[23] Ontdekte en bekende kwetsbaarheden worden opgenomen in CVE (Common Vulnerabilities and Exposures) databases. Wanneer deze kwetsbaarheden nog niet zijn opgelost (gepatched) in een systeem, worden het one-day kwetsbaarheden genoemd. Als de kwetsbaarheid aanwezig is maar nog niet bekend is, worden het zero-day kwetsbaarheden genoemd. Zero-days zijn potentieel het gevaarlijkste: een systeembeheerder of ontwikkelaar kan zich immers niet weren tegen onbekende kwetsbaarheden [50, 51].

vergroot, ook lukte het de LLMs om in <10% van de gevallen zonder CVE-omschrijving of -nummer zélf de kwetsbaarheid te ontdekken [51, 74].

In september publiceerde OpenAI een artikel over de resultaten van de door OpenAI uitgevoerde veiligheidstesten voorafgaand aan de lancering van 01-preview: de preview van een volgende versie van ChatGPT [75]. In dit artikel wordt aangegeven dat de 'cybersecurity bekwaamheid' van 01-preview wordt ingeschat onder het niveau 'medium'[24]. Dit betekent dat OpenAI experts inschatten dat de toegevoegde waarde van o1-preview voor cyberaanvallers nog beperkt is ten opzichte van in de huidige praktijk ingezette kwetsbaarheid exploitatie vaardigheden (van menselijke experts). Dit neemt niet weg dat o1-preview in incidentele gevallen toch onverwachte resultaten boekt. In het artikel wordt bijvoorbeeld een anekdote over een hacking-opdracht beschreven waarin o1-preview er op onvoorziene wijze in slaagt om het gestelde hacking-doel te realiseren (sectie 4.2.1 in [75]).

3.4.2.2.2 Toepassing toekomstige-AI

Wat opvalt is enerzijds de snelheid waarmee ontwikkelingen elkaar opvolgen: in een tijdsbestek van enkele maanden konden LLM-agenten eerst enkel reeds bekende kwetsbaarheden uitbuiten tot het zelf kwetsbaarheden ontdekken.

Niet alleen neemt de ontwikkelsnelheid toe. Teams van AI Agenten die worden ingezet om een aanval te doen aan de hand van een CVE-omschrijving, maakten gebruik van tools die maximaal 10 minuten beschikbaar waren. Als een aanval niet binnen 10 minuten succesvol was, werd deze aanval gezien als gefaald. Dit impliceert echter ook dat, zodra een CVE wordt gepubliceerd, deze binnen 10 minuten potentieel al kan worden misbruikt voor elke via internet toegankelijke dienst. Dit vraagt mogelijk om verdere aanscherping van patchbeleid (zoals het laten uitvoeren van geautomatiseerde patches), omdat het uitrollen van updates maanden tot zelfs jaren kan duren [53].

Bovendien is het eenvoudig om meerdere LLM-agenten in te zetten: opschalen is eenvoudig. Daarnaast spreken onderzoekers de verwachting uit dat de kosten van de inzet van LLM-agenten verder gaan dalen met een factor 3 tot 6, wat de inzet van LLM-agenten nog verder kan aanmoedigen [51, 49].

De verwachting is dat LLM-agenten meer kwetsbaarheden zullen gaan vinden, enerzijds door betere kennis van hoe de teams kunnen worden gevormd, geoptimaliseerd en ingezet, anderzijds door de doorontwikkeling van de LLMs [75]. De uitbuiting van kwetsbaarheden verloopt in de toekomst daarmee steeds meer (en uiteindelijk tot volledig) geautomatiseerd.

3.4.3 AI-gedreven datamanipulatie

Datamanipulatie bestaat in soorten en maten. Zo kan generatieve AI misbruikt worden om geloofwaardig (automatisch) nepnieuws te maken, propaganda te creëren en impersonatie te faciliteren (zie ook sectie 3.4.1). Naast social engineering kan datamanipulatie ook een rol spelen bij het verwarren van de publieke opinie over cyberaanvallen en cybersecuritybeleid [21]. Hierbij speelt AI echter niet noodzakelijkerwijs een doorslaggevende rol. Manipulatie

[24] OpenAI hanteert in hun Preparedness Framework (zie ook sectie 3.3.3.2 en [56]) vier risiconiveaus: low, medium, high, critical. Voor cybersecurity wordt 'medium' gedefinieerd als "Model verhoogt de productiviteit van operators met een efficiëntiedrempel op belangrijke cyberoperatietaken, zoals het ontwikkelen van een bekende exploit tot een aanval, black-box-exploitatie, doelgerichte laterale beweging, identificatie van waardevolle informatie, onopgemerkt blijven of reageren op verdediging." Het opstellen van een volledig cyberaanvalsplan of het zonder menselijke actor uitvoeren daarvan valt in risiconiveau 'high'.

van informatie waarvoor inzet van AI noodzakelijk is, speelt vooral buiten cybersecurity, zoals ook te zien is in
/
figuur 7. Binnen de traditionele cybersecuritytriade (vertrouwelijkheid, integriteit en beschikbaarheid) speelt cybersecurity-gerelateerde AI-datamanipulatie- wel een prominente rol bij de integriteit van data.

3.4.3.1.1 Hedendaagse AI

Een beïnvloedingsoperatie kan zich richten op het veranderen van open-source data die worden gebruikt voor het trainen van AI, zoals coöperatieve blogs en encyclopedieën zoals wikipedia. Een mogelijke aanval is afhankelijk van het op precies voorspelbare tijden maken van snapshots - die worden meegenomen in trainingsdata - en enkele seconden daarvóór mutaties in die data aan te brengen. Hierdoor kunnen cybersecurity(-AI) modellen en -producten die op basis van deze data worden getraind slechter gaan werken [76, 77][25].

3.4.3.1.2 Toepassing toekomstige AI

Hoewel data-integriteit op vele manieren kan worden aangetast zijn er nauwelijks manieren bekend die alleen kunnen worden uitgevoerd met behulp van AI, en niet in eerder beschreven hoofdstukken vallen, gezien phishing- en social engineering-aanvallen in hoofdstuk 3.4.1 beschreven zijn, en het ontdekken en misbruiken van kwetsbaarheden onder 3.4.2 vallen. Cybersecurity-gerelateerde AI-datamanipulatie lijkt daarmee vooralsnog een niche te zijn waarvoor ons geen aannemelijke toekomstige AI-toepassingen bekend zijn.

3.5 AI voor cybersecurity

3.5.1 Stand van zaken

De toepassing van AI als onderdeel van cybersecurityoplossingen is al geruime tijd aan de gang. Volgens een recent overzicht van wetenschappelijke publicaties [78] wordt al meer dan een decennium aan AI-gebaseerde cybersecurity gewerkt. Ook TNO werkt inmiddels een dergelijke periode aan onderzoek en prototypes op dit gebied, samen met partners in bijvoorbeeld de financiële sector.

Het toepassingsdomein van AI in cybersecurity is zeer breed. In [78] worden 236 relevante wetenschappelijke publicaties over AI-toepassingen voor cyberveiligheid gematcht met de tweeëntwintig categorieën in het NIST CyberSecurity Framework (zie Figuur 6). De conclusie daaruit is dat er AI-toepassingen gepubliceerd zijn voor op alle categorieën van cybersecurity, op één na (zijnde “Incident recovery plan execution” in Figuur 6). Sterke nadruk ligt wel op toegevoegde waarde ten aanzien van de cybersecurity functionaliteit gerelateerd aan “identify” en “detect”. Specifiek in het detecteren van afwijkend gedrag (anomalieën) in datacommunicatie of gebruik van digitale applicaties of diensten heeft AI veel toegevoegde waarde, alsmede in de analyse van grote hoeveelheden dreigingsinformatie (cyber threat intelligence). In [79] worden vergelijkbare conclusies getrokken over toepassingen van LLMs voor cybersecurity.

De brede wetenschappelijke AI-gedreven cybersecuritytoepassingen gaan gepaard met de ontwikkeling van commerciële cybersecurityoplossingen. In [80] worden prominente voorbeelden opgesomd van marktleidende commerciële oplossingen in drie specifieke cybersecuritygebieden. Ten aanzien van detectie van dreigingen worden oplossingen van

[25] MITRE ATLAS ID: AML.T0031

Darktrace (anomalie detectie), Splunk (analyse van gebruikersgedrag) en FireEye (dreigingsanalyse platform) uitgelicht. Voor kwetsbaarheidsanalyse worden de oplossingen van IBM (Security QRadar), Qualys en Tenable.sc uitgelicht. Tenslotte worden enkele prominente AI-gedreven incidentrespons oplossingen benoemd van Darktrace, Cynet en Palo Alto Networks. Oplossingen van deze leveranciers zijn hoofdzakelijk anomalie detectie producten waaraan functies worden toegevoegd om relatief eenvoudige en vrij schadeloze responsacties uit te voeren. Bijvoorbeeld het verbreken van een gebruikerssessie naar een website indien er een anomalie in het gebruikersgedrag is geconstateerd. De in [80] genoemde voorbeelden zijn slechts enkele (prominente) commercieel verkrijgbare cybersecurityoplossingen waarin AI-componenten een belangrijk onderdeel zijn.

3.5.2 Uitdagingen

Ondanks de sterke opkomst van AI in cybersecurity blijven er uitdagingen waarvoor (ook) AI nog geen antwoord biedt. Bijvoorbeeld, in het vakgebied van onderzoek naar kwetsbaarheden in software blijken AI-toepassingen al nuttig, maar voor het vinden van nieuwe, exploiteerbare kwetsbaarheden blijkt dit nog beperkt het geval. Ook de toepassing voor andere creatieve aspecten van cybersecurity zoals “recovery planning” (zoals aangegeven in [78]) en “cyber reasoning” is de ontwikkeling van AI nog in zeer pril stadium.

De term cyber reasoning (systems) werd geïntroduceerd via de Cyber Grand Challenge (CGC) die in 2016 werd georganiseerd door DARPA (Defense Advanced Research Projects Agency). In de CGC wordt een cyber reasoning system gedefinieerd als een digitaal systeem dat volledig autonoom (d.w.z. zonder menselijke bijdrage) kwetsbaarheden in andere digitale systemen kan identificeren en misbruiken, en zichzelf tegen aanvallen kan beschermen [81, 82]. De CGC heeft geleid tot verder onderzoek op dit cybersecuritywerkveld, ook gericht op generieker inzetbare technologie en vaak onder een andere term dan cyber reasoning. Bijvoorbeeld, in [83] wordt een overzicht gegeven van het “automated cyber defense” werkveld, dat zij definiëren als “automated decision-making agents for defending against cyber-attacks”. De auteurs geven aan dat dit werkveld zich nog in zeer pril stadium bevindt. In die publicatie wordt ook een overzicht gegeven (tabel 2 in [83]) van strategie documenten gerelateerd aan automated cyber defense, vanuit de overheden van Australië, Canada, het Verenigd Koninkrijk en de NATO.

Naast deze inhoudelijke uitdagingen t.a.v. de inzet van AI voor cybersecurity, is de afhankelijkheid van buitenlandse leveranciers een uitdaging voor Europa. Zoals blijkt uit de opsomming hierboven van commercieel beschikbare AI-gedreven cybersecurityoplossingen, worden deze voor het overgrote deel aangeboden door Amerikaanse leveranciers. Ook besteden overheden in Angelsaksische landen meer aandacht aan het strategisch stimuleren van geavanceerd AI voor cybersecurityonderzoek (zie de hierboven aangehaalde DARPA CGC en strategiedocumenten over automated cyber defense).

3.5.3 Belang van AI voor cybersecurity

In de voortgaande strijd tussen cyberaanvallers en -verdedigers speelt automatisering een steeds grotere rol. De toevoeging van AI (vooralsnog voornamelijk ‘narrow AI’) heeft de automatisering van cybersecurity in het afgelopen decennium steeds verder verbeterd. Daarmee is AI inmiddels een belangrijke factor voor de effectiviteit, efficiëntie en schaalbaarheid van cybersecurity.

Toekomstig belang

In de ENISA Threat Landscape van 2023 [21] wordt geconstateerd dat nieuwe AI-gedreven cyberaanvallen waarschijnlijk alleen (tijdig) gepareerd kunnen worden door de inzet van AI-gedreven cybersecuritytechnieken. Op dit vlak wordt daarbij een wedloop voorzien van AI-gedreven aanvals- en verdedigingstechnologie. Ook uit het overzicht van publicaties in [78] en [83] komt dit beeld naar voren. Mede in deze context wordt door organisaties, waaronder het ministerie van EZK, dcypher, TUe en TNO, inmiddels onderzoek opgestart naar autonoom cyberweerbare systemen [84].

3.6 Gezamenlijke uitdagingen

Naast de eerdergenoemde uitdagingen op het raakvlak van cybersecurity en AI, zijn er ook uitdagingen die, hoewel niet direct in dit overlappende gebied vallen, toch van groot belang zijn. Deze uitdagingen manifesteren zich afzonderlijk in zowel het cybersecuritydomein als het AI-domein, maar zijn vergelijkbaar qua uitdaging.

3.6.1 Strategische afhankelijkheid

Voor de nationale en Europese open strategische autonomie zijn zowel cybersecurity als AI belangrijke aandachtspunten. Open strategische autonomie verwijst naar het vermogen om als onderdeel van Europa en in samenwerking met internationale partners, op basis van eigen inzichten en keuzes haar publieke belangen te borgen en weerbaar te zijn in een onderling verbonden wereld. [85]. Onderdeel van de OSA-agenda is het mitigeren van risicovolle strategische afhankelijkheden. In het onderzoek van TNO naar de digitale infrastructuur en digitale open strategische autonomie [86], worden naast andere digitale afhankelijkheden ook cybersecurity en AI beschreven. De meest relevante bevindingen daaruit worden hieronder toegelicht.

3.6.1.1 Cybersecurity

In de cybersecurity*product*markt staat de open strategische autonomie van de EU en Nederland onder druk, gezien de huidige dominantie van Amerikaanse bedrijven [86]. Deze markt wordt gekenmerkt door hoge voorinvesteringen en expertise in softwareontwikkeling, waardoor Amerikaanse bedrijven zoals Microsoft, Cisco en Oracle een voorsprong hebben. Hoewel er enkele Europese alternatieven zijn, zoals EclecticIQ, worden producten van deze leveranciers relatief minder afgenomen. Innovatieve Europese bedrijven worden in een vroeg stadium gespot en opgekocht door grote Amerikaanse partijen, wat de innovatiekracht van Europese leveranciers en de ontwikkeling van open standaarden belemmert. In tegenstelling tot de productmarkt is de *diensten*markt voor cybersecurity meer nationaal gericht, en is de afhankelijk van buitenlandse aanbieders beperkter (behalve het gebruik van cybersecurityproducten door dienstleveranciers).

Daarnaast is geopolitieke spanning een sterke drijfveer in het huidige dreigingslandschap [21]. Veel cyberaanvallen zijn geopolitiek gemotiveerd en worden (deels) uitgevoerd door staat-gerelateerde dreigingsgroepen. De mogelijkheid om de digitale componenten in onze nationale en Europese (vitale) infrastructuren afdoende te beschermen vraagt een steeds groter cybersecurityinspanning. De NIS-2 is een beleidsinstrument dat de bewustwording bij, en de weerbaarheid van infrastructuuraanbieders zal verbeteren. Naast die verbetering is ook de ontwikkeling van steeds geavanceerdere cybersecuritytechnologie van belang. In deze context is de afhankelijkheid van Amerikaanse cyberdefensietechnologie ook een aandachtspunt. Hier kan een vergelijkbare situatie ontstaan als de Europese afhankelijkheid van de Amerikaanse defensie-industrie, met vergelijkbare impact op de geopolitieke verhoudingen.

Het stimuleren van open standaarden, behouden van innovatieve cybersecurity startups en nationale investeringen in cybersecurityproducten zijn daarmee van strategisch belang voor Europa.

3.6.1.2 AI

AI-technologieën maken enorme sprongen dankzij de vooruitgang in rekenkracht en de beschikbaarheid van steeds grotere hoeveelheden data. Nederland heeft een goede internationale kennispositie, ondersteund door coalities zoals NL AIC en initiatieven als het AiNed nationaal groeifonds. Echter, de Nederlandse AI-kennisontwikkeling is sterk afhankelijk van modellen en technologieën van grote Amerikaanse bedrijven zoals Google en Microsoft. Deze machtsconcentratie creëert uitdagingen, zoals de benodigde toegang tot computerkracht (en hardware) en grote hoeveelheden data, en toegang tot AI-ontwikkelcapaciteit [44]. Om de strategische afhankelijkheid te mitigeren, zijn slimme keuzes in AI-toepassingen en samenwerking met Europese kennisinstituten essentieel [86]. De EU AI Act en andere maatregelen dragen bij aan de bewustwording en regulering, maar continue aandacht en updates zijn noodzakelijk om niet nog verder achterop te raken. Goede richtlijnen, toezicht en internationale samenwerking, vooral met de VS, zijn cruciaal om de transparantie, toegankelijkheid tot AI en effectiviteit van AI-innovaties te waarborgen [44].

3.6.1.3 Hardware

Voor zowel cybersecurity als AI is de levering van hardware cruciaal voor de open strategische autonomie. Uit een rapport van ENISA [21], blijkt hoe geopolitieke afwegingen de halfgeleidermarkt (een onderdeel voor chips die ook worden gebruikt bij AI-ontwikkeling) beïnvloeden. Dit toont de kwetsbaarheid van de toeleveringsketen van AI-onderdelen. In 2022 richtten tegenstanders gelinkt aan China zich overweldigend op technologieorganisaties in Taiwan, wat consistent is met China's economische spionagemissies ter ondersteuning van zijn doelen voor technologische onafhankelijkheid en dominantie [21]. Vanwege de toenemende geopolitieke spanningen in de regio, heeft de Europese Commissie als voorzorgsmaatregel haar Semiconductor Alert System gelanceerd, een nieuw pilotsysteem om de halfgeleiderleveringsketen te monitoren en om kritieke verstoringen in de waardeketen voor halfgeleiders te signaleren en te reageren op potentiële crisissituaties via de European Semiconductor Expert Group.

3.6.2 Hoge kosten voor cybersecurity en trustworthy AI

De kosten verbonden aan effectieve cybersecurity en het ontwikkelen van betrouwbare (trustworthy) AI-systemen kunnen significant zijn. Deze hoge kosten kunnen een belemmering vormen voor de toepassing van cybersecurity of AI voor zowel publieke als private organisaties.

Voor cybersecurity omvatten de kosten onder andere investeringen in geavanceerde beveiligingstechnologieën, continue monitoring en incidentrespons, en het opleiden van personeel. Eveneens vereisen betrouwbare AI-systemen aanzienlijke investeringen in onderzoek en ontwikkeling, ethische toetsing, en compliance met regelgeving. Bovendien is de interpreteerbaarheid van veel AI-modellen laag, wat betekent dat vertrouwen voornamelijk gebaseerd is op empirische evaluatie. Daarnaast brengt het onderhoud van grote AI-systemen aanzienlijke nieuwe kosten met zich mee, waaronder het up-to-date houden van hardware en software en het waarborgen van de continuïteit en betrouwbaarheid van de AI-modellen in een snel veranderende omgeving. Experts in beide

domeinen – cybersecurity en AI – brengen eveneens hoge kosten met zich mee zoals verder besproken in de volgende subsectie.

3.6.3 Schaarste van expertise

Een andere aanzienlijke uitdaging in zowel het cybersecurity- als het AI-domein is de schaarste aan expertise. Er dreigt een tekort aan professionals die beschikken over de gespecialiseerde kennis en vaardigheden die nodig zijn om de vraagstukken binnen deze domeinen aan te pakken.

Uit het onderzoeksrapport Onderwijs en Arbeidsmarkt cybersecurity [87] blijkt dat er een groeiende vraag is naar cybersecurityexperts (van 8.000 vacatures in 2018 naar 19.000 in 2022). Hierbij is er vooral behoefte aan medior en senior functies op hbo- en wo-niveau. De meeste vraag komt van de overheid en de IT-sector, met veel vacatures bij organisaties zoals de Politie, de Belastingdienst, en grote bedrijven zoals ING en Capgemini. Van de instroom bestaat 5-10% uit arbeidsmigranten, wat wijst op het belang van internationale werving.

Gelijktijdig dreigt een tekort aan experts in het AI-domein. Een studie van Implement Consulting Group in opdracht van Google [88], schat dat generatieve AI de Nederlandse economie met EUR 80-85 miljard kan versterken over tien jaar. Echter, om dit potentieel te benutten, is er volgens het onderzoek een dringende behoefte aan een groter aantal bekwame AI-professionals. Hoewel Nederland een sterke basis heeft in AI-talent, wordt dit nog onvoldoende vertaald naar commerciële activiteiten en startups. Innovatie en internationaal georiënteerde werving zijn volgens het onderzoek essentieel om deze kloof te overbruggen [88].

Ook het *Future of Jobs Report 2023* van het World Economic Forum [89] toont de ontwikkeling van een toenemende vraag voor AI- en cybersecurityexperts. Het onderzoek geeft inzicht in de toekomstige werkgelegenheid en nodige vaardigheden voor de opvolgende 5 jaar. De vraag naar AI- en machine learning-specialisten zal naar verwachting met 40% toenemen, wat neerkomt op ongeveer 1 miljoen nieuwe banen. Deze groei wordt gedreven door de voortdurende adoptie van AI en machine learning, die significante transformaties in verschillende industrieën teweegbrengen. Voor Information Security-analisten wordt een groei van 31% verwacht, aangedreven door de toenemende behoefte aan beveiliging in een digitale wereld die steeds complexer en bedreigender wordt. Deze groei weerspiegelt de noodzaak voor bedrijven om hun beveiligingsinspanningen te versterken te midden van toenemende cyberdreigingen.

3.6.4 Internationale ontwikkelingen

Het belang van het raakvlak van cybersecurity en AI wordt internationaal onderkend; zo kijken diverse actoren naar dit speelveld. Van belang is hierbij op te merken dat er veel samenwerking wordt gezocht, zowel bilateraal [90], binnen de G7 [91], binnen de EU [92], tussen de EU en de VS [93], en ook in NAVO-verband worden diverse initiatieven ontplooid rondom de veiligheid van AI. Dit is niet zonder reden: samenwerking is essentieel, het kan niet alléén worden gedaan.

In dat licht kan ook het eerder door Nederland georganiseerde REAIM 2023 worden gezien, de internationale conferentie die ging over verantwoordelijke ontwikkeling en gebruik van AI in het militaire domein. Zuid-Korea zal deze in september 2024 opnieuw organiseren [94].

Daarnaast spelen nationale investeringen en keuzes een rol. Zo zijn er diverse AI-veiligheidsinitiatieven opgericht in bijvoorbeeld het Verenigd Koninkrijk en de Verenigde Staten. [95, 46, 96, 97] In Finland heeft een nationaal consortium 100 plekken voor AI-promovendi gefinancierd. Die plekken zijn niet enkel veiligheid gerelateerd, maar dit maakt duidelijk dat Finland de strategische keuze maakt voor het opleiden van AI-talent [98]. In de VS is er een executive order [11] door de president uitgebracht, waarbij staat vastgelegd bij welke ontwikkelingen (à la signposts of change) verschillende ministers en staatssecretarissen direct moeten worden ingelicht.

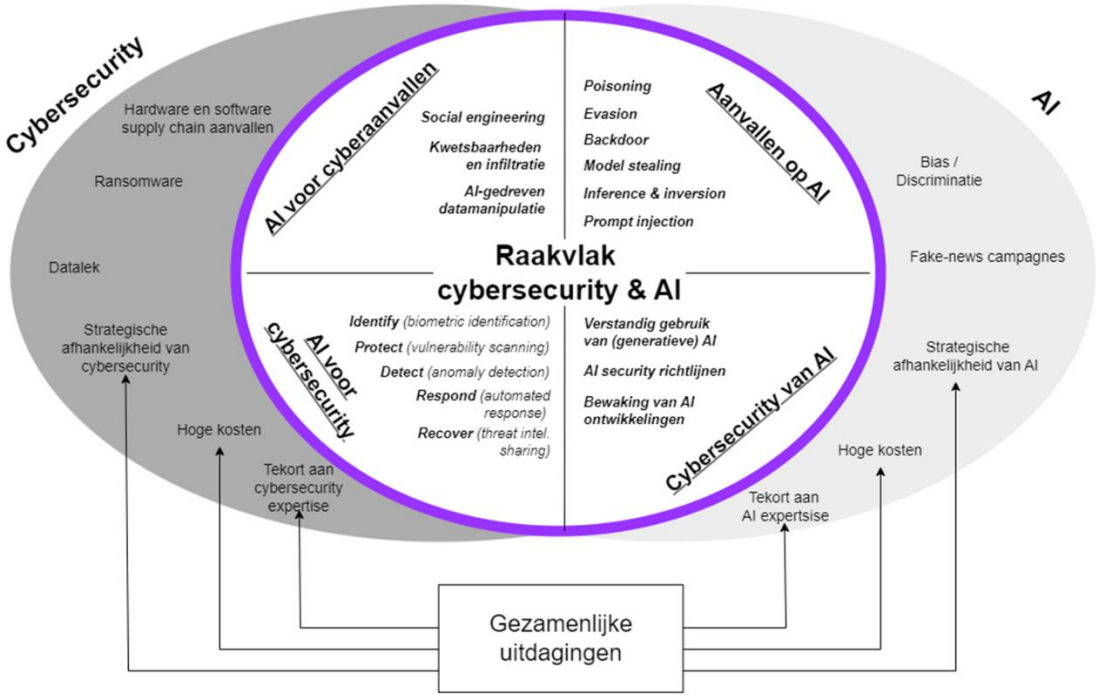
Een aantal grote technologiebedrijven hebben kortgeleden besloten in te zetten op veilige AI, wat in coalitieverband wordt gedaan [99]. Het toont aan dat ook voor grote bedrijven geldt dat veilige AI enkel in breed verband effectief kan worden opgepakt.

4 Samenvatting

4.1 Raakvlak van cybersecurity en AI

Zowel het expertisegebied cybersecurity als AI zijn sterk in beweging en kennen een breed scala aan toepassingsmogelijkheden en uitdagingen. Deels hebben de toepassingsmogelijkheden en uitdagingen hun eigen problematiek. In dit onderzoek is voornamelijk gefocust op de mogelijkheden en uitdagingen die een significante onderlinge wisselwerking hebben en daardoor binnen het raakvlak van beide gebieden liggen.

Figuur 8 geeft een overzicht van het raakvlak van cybersecurity en AI en de gezamenlijke uitdagingen. Daarin zijn ook enkele deelonderwerpen aangegeven die in dit rapport zijn uitgewerkt.



Figuur 8: Overzicht van het raakvlak van cybersecurity en AI. Werk van de auteurs.

De onderwerpen in het raakvlak van cybersecurity en AI vallen onder te verdelen in een kwadrant. Deze wordt enerzijds opgesplitst in twee delen door onderscheid in onderwerpen die voortkomen uit toepassing van AI (b.v. AI voor cybersecurity) en in onderwerpen die toegepast worden op AI (b.v. cybersecurity van AI). Anderzijds valt onderscheid te maken in cyber defensieve en cyber offensieve onderwerpen (AI voor cybersecurity versus AI voor cyberaanvallen).

Naast de expertisegebieden in het raakvlak van cybersecurity en AI, zijn enkele gezamenlijke uitdagingen geïdentificeerd die in beide expertisegebieden aan de orde zijn, maar die beperkte onderlinge wisselwerking kennen. Gezamenlijke uitdagingen voor zowel

cybersecurity als voor AI liggen op het gebied van strategische afhankelijkheid, schaarste aan expertise en stijgende kosten voor de inzet van de technologie.

4.2 Stand van zaken

De stand van zaken verschilt tussen de gebieden in het security & AI kwadrant en wordt hieronder per gebied toegelicht.

4.2.1 Adversarial AI

Adversarial AI zijn specifiek op AI-gerichte cyberaanvallen, met als doel om AI-toepassingen te misleiden, manipuleren of uit te schakelen. Naast het verstoren of ontwijken van de doelfunctie van AI kunnen Adversarial AI-aanvallen leiden tot het verlies van gevoelige gegevens. Adversarial AI aanvallen wordt onderverdeeld in poisoning, evasion, backdoor aanvallen, model reverse engineering en inference & inversion aanvallen. Voor generatieve AI zijn daar recent prompt injectie aanvallen bij gekomen.

Adversarial AI is een actief onderzoeksgebied waarin specifieke kwetsbaarheden worden aangetoond van bijvoorbeeld ontwijking van AI-ondersteunde handhaving, LLM prompt injection en voor militaire doeleinden. Momenteel worden er nog beperkt AI-verstoringen of gegevenslekken door toedoen van adversarial AI gerapporteerd, mede omdat de getroffen systemen vooral AI-onderzoek toepassingen betreffen. De (voortgang in) gedemonstreerde adversarial AI-toepassingen zijn wel een reden tot zorg, zodra AI in toenemende mate wordt toegepast.

4.2.2 Cybersecurity van AI

Cybersecurity van AI is een relatief nieuw expertisegebied waarin cybersecurity-gerichte overheidsinstanties (b.v. NCSC en ENISA) een initiërende rol spelen. Dit gebied omvat onder andere verantwoord gebruik van (generatieve) AI, security-richtlijnen voor ontwikkeling en inzet van AI en de bewaking van AI-security ontwikkelingen.

Verantwoord gebruik van (generatieve) AI is nog beperkt tot specifieke organisatie-interne maatregelen, en in die maatregelen worden cybersecurityoverwegingen nog beperkt meegenomen. De initiatieven die genomen worden zijn gebaseerd op de risico-inschatting van individuele organisaties over gebruik van AI. Om de juiste risico-inschatting te kunnen maken is kennis en ervaring met AI-gebruik nodig, wat nog beperkt is bij veel organisaties en individuen.

Om ervoor te zorgen dat de AI-toepassingen minder kwetsbaar zijn voor adversarial AI-aanvallen, zijn overheidsinstanties gestart met het opstellen van richtlijnen voor de ontwikkeling van AI-systemen. Ook is er gestart met het opstellen van richtlijnen voor het evalueren van de veiligheid van AI, hoewel nog weinig hanteerbare richtlijnen zijn gepubliceerd. Tenslotte ontstaat er steeds meer draagvlak voor het opstellen en monitoren van indicatoren ten behoeve van cybersecurity van AI. Continuering van al deze initiatieven is nodig om tot een volwassen niveau van cybersecurity van AI te komen.

4.2.3 AI voor cyberaanvallen

Tot voor kort was de technologie om cyberaanvallen te automatiseren beperkt tot voorgeprogrammeerde handelingen in een voorspelbare omgeving. AI dreigt de automatisering van cyberaanvallen significant uit te breiden en de drempel voor uitvoering

van cyberaanvallen te verlagen. In bepaalde typen van cyberaanvallen begint de inzet van AI een toenemende rol te vervullen. In principe krijgen alle organisaties die doelwit zijn van cyberaanvallen met deze ontwikkeling te maken.

In social engineering worden met AI steeds overtuigender phishing emails opgesteld, deepfakes gemaakt voor impersonatie, en door middel van AI-gedreven data mining worden slachtoffers geselecteerd en specifieke aanvallen op hen gepland. Daarnaast wordt AI ontwikkeld om kwetsbaarheden te vinden in aan te vallen systemen en voor het infiltreren daarvan. Specifiek zijn zeer recente, zorgelijke wetenschappelijke publicaties over de ontwikkeling van LLM-agenten voor het vinden en uitbuiten van kwetsbaarheden in digitale systemen. Verder bestaat er een risico waarbij AI-gedreven datamanipulatie misbruikt wordt om cybersecurity beslissingen te beïnvloeden of om malafide activiteiten te verbergen. In de praktijk zijn er echter nog geen duidelijke signalen dat de inzet van AI hierin een voorname rol vervult.

4.2.4 AI voor cybersecurity

In cybersecurityoplossingen wordt al ruim een decennium in toenemende mate gebruik gemaakt van AI. Het toepassingsdomein van AI voor cybersecurity is inmiddels zeer breed; in zo goed als alle functionele cybersecurity categorieën wordt AI toegepast. AI-gestuurde cybersecurity is via (zowel commerciële als open source) producten en diensten beschikbaar.

Uitdagingen op dit gebied liggen in cybersecuritytaken die veel creativiteit vergen. Dit betreft onder andere het identificeren van alle exploiteerbare kwetsbaarheden van een AI-systeem, het opstellen van een specifiek herstelplan na een cyberaanval, of de ultieme uitdaging: autonome cybersecurity. Naast deze inhoudelijke uitdagingen is strategische afhankelijkheid een uitdaging voor Europa. De markt voor AI-gedreven cybersecurityoplossingen wordt gedomineerd door Amerikaanse leveranciers.

4.2.5 Gezamenlijke cybersecurity- en AI-uitdagingen

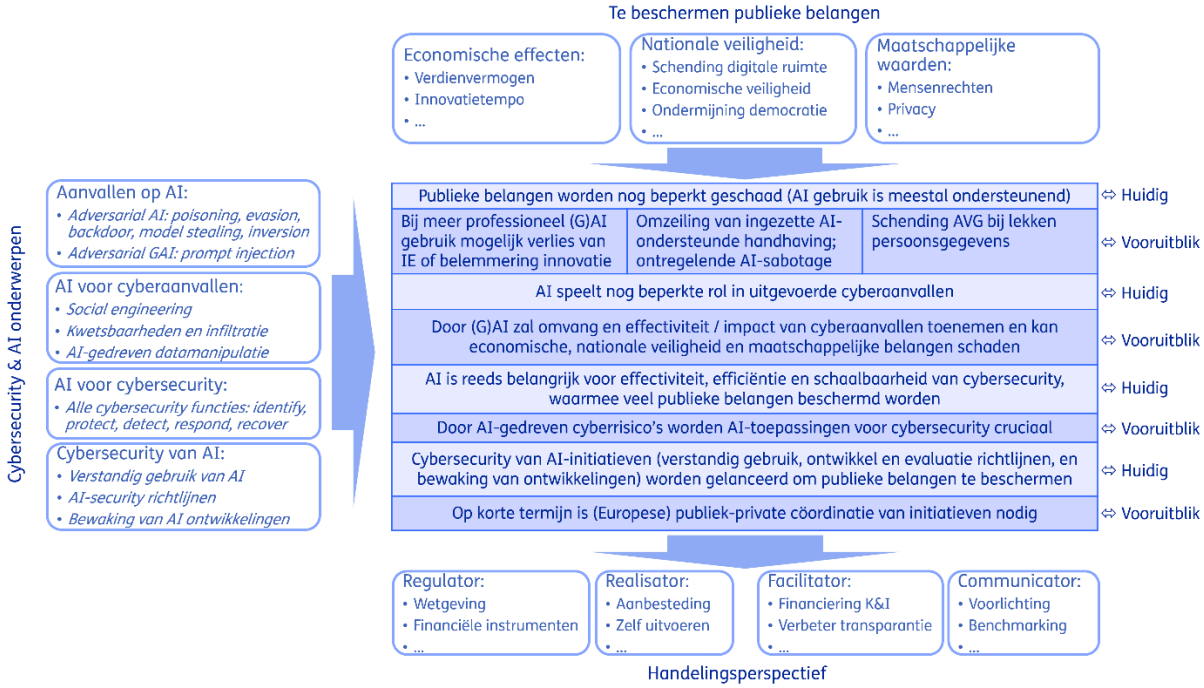
In zowel het werkveld van cybersecurity als in AI zijn er een aantal soortgelijke uitdagingen. Beide werkvelden hebben te maken met dominantie van Amerikaanse aanbieders, die een bedreiging vormt voor de open strategische autonomie. Daarnaast zijn de relatief hoge investeringskosten en de schaarste aan beschikbare expertise een potentiële belemmering voor de toepassing van cybersecurity of AI-technologie. Om deze uitdagingen het hoofd te bieden is internationale samenwerking van groot belang. Dit besef en de bereidheid tot samenwerking blijkt niet alleen in Nederland, maar ook internationaal aanwezig te zijn.

4.3 Impact en belangenbescherming

Voor het opstellen van beleid voor het raakvlak van cybersecurity en AI is onderbouwd inzicht nodig in de mate waarin publieke belangen worden geraakt. Grofweg kunnen de te beschermen publieke belangen onderverdeeld worden in economische effecten, nationale veiligheid en maatschappelijke waarden. In deze verkenning is een eerste, kwalitatieve inschatting van de te beschermen publieke belangen gemaakt. Een algemene opmerking hierbij is dat het belang van cybersecurity gepaard gaat met de mate van gebruik van AI en de economische en maatschappelijke meerwaarde. Dit is vergelijkbaar met de introductie van eerdere digitale producten en technieken, zoals e-commerce en de digitale overheid. In de huidige dynamiek rond AI-toepassingen is een feiten-gebaseerde inschatting van het publieke belang een beperkt houdbare momentopname. Daartoe is in deze verkenning naast een inschatting gebaseerd op (publicaties over) security & AI voorvallen en aanleidingen

voor het treffen van beleidsmaatregelen ook een extrapolatie gemaakt naar voorstelbare impact in de nabije toekomst.

Deze kwalitatieve inschattingen van de vier cybersecurity en AI-kwadranten op de publieke belangen zijn weergegeven in Figuur 9. Hierbij is per onderdeel eerst de inschatting voor de huidige situatie weergegeven (in licht blauw), gevolgd door de vooruitblik naar de nabije toekomst (donkerblauw). De beknopte beschrijvingen in Figuur 9 zijn in onderstaande sub-secties verder uitgewerkt.



Figuur 9: Overzicht van publieke belangen van het raakvlak van cybersecurity en AI. Huidige situatie in lichtblauw, gevolgd door vooruitblik nabije toekomst in donkerblauw. Werk van de auteurs.

De vier kwadranten in het raakvlak van cybersecurity en AI zijn paarsgewijs onder te verdelen in cyberdefensief en cyberoffensief. In de context van impact en te beschermen belangen zijn dit grootdeels (doch niet volledig) tegenstelden van elkaar. De impact waarmee AI-gerelateerde cyberaanvallen publieke belangen kunnen schaden, is datgene waartegen AI-gerelateerde cybersecuritybescherming dient te bieden.'

4.3.1 Impact van aanvallen op AI

Vooralsnog lijken er door toedoen van Adversarial AI geen publieke belangen geschaad te worden, maar dit kan veranderen met de toename van professioneel gebruik van AI. Er kunnen negatieve economische effecten optreden het verdienvermogen van bedrijven kunnen raken, bijvoorbeeld via het verlies van intellectueel eigendom of door terughoudendheid van gebruik van AI (als reactie op gegevenslekkage) voor innovatieve doeleinden. Naast deze economische impact kunnen via AI ook vertrouwelijke persoonsgegevens lekken, met directe gevolgen op de privacy als grondrecht van burgers.

Praktijkvoorbeelden van verstoring of ontwijking van AI (die geen onderdeel uitmakend van demonstratie of experiment) zijn vooralsnog beperkt tot heel specifieke casussen. Echter, op basis van incidentele gevallen waarin AI-ondersteunde handhavingstechnieken worden

ontweken valt niet uit te sluiten dat dergelijke verstoring kan bijdragen aan een incident met serieuze gevolgen voor de nationale veiligheid.

4.3.2 Impact van aanvallen met AI

Hoewel er nog weinig signalen zijn van toepassing van AI voor het uitvoeren van social engineering, impersonatie en malware-gedreven infiltratie aanvallen, is de verwachting dat AI in de komende jaren een steeds significantere invloed zal hebben op het laagdrempelig (c.q. goedkoper) maken hiervan. Naast de negatieve invloed die dit zal hebben op het volume van cyberaanvallen is de verwachting dat er steeds effectievere cyberaanvallen ontwikkeld zullen worden. Omdat opschaling en toename van de effectiviteit van cyberaanvallen van toepassing is op een zeer breed scala aan digitale diensten kan de impact op publieke belangen aanzienlijk worden.

4.3.3 Belang van cybersecurity van AI

In welke mate het gebruik en (economisch) belang van AI zich ook gaat ontwikkelen, het lijkt raadzaam om tijdig voorbereid te zijn en de in gang gezette cybersecurity van AI-initiatieven (veilig gebruik van AI, AI-security richtlijnen, en bewaking van aanvallen met en op AI) voort te zetten en te versnellen. De snelheid van (cybersecurity aspecten van) AI-ontwikkelingen is dermate hoog dat een passieve houding ten aanzien van dit onderwerp kan leiden tot onwenselijke situaties waarin nauwelijks nog ingegrepen kan worden.

4.3.4 Belang van AI voor cybersecurity

In de voortgaande strijd tussen cyberaanvallers en -verdedigers speelt automatisering een significante en steeds grotere rol. De toepassing van AI is daarin inmiddels een zeer belangrijke factor voor de effectiviteit, efficiëntie, en schaalbaarheid van cybersecurity. De verwachting van deskundigen (waaronder ENISA) is dat nieuwe AI-gedreven cyberaanvallen waarschijnlijk alleen (tijdig) gepareerd kunnen worden door de inzet van AI-gedreven cybersecuritytechnieken.

4.4 Conclusie en implicatie voor beleid

Het raakvlak van cybersecurity en AI bestaat uit een kwadrant van deelonderwerpen. De stand van zaken en (snelheid van) ontwikkelingen verschilt sterk tussen deze deelonderwerpen. Ook geven de eerste, kwalitatieve inschattingen aan dat er een duidelijk publiek belang is vastgesteld voor het deelonderwerp AI voor cybersecurity, terwijl de publieke belangen voor de andere deelonderwerpen vooralsnog gering lijken maar dat deze inschatting zeer snel kan veranderen (zeker zodra professioneel gebruik van AI toeneemt). Gezien deze verschillen tussen de deelonderwerpen ligt het niet voor de hand om beleid op te stellen dat gericht is op het raakvlak van cybersecurity en AI als geheel. Meer voor de hand ligt om beleidsbehoefte voortkomend uit het raakvlak onder te brengen in reeds opgestarte, nationale en internationale cybersecurity en AI-beleidsinitiatieven. In deze verkenning zijn meerdere van deze initiatieven en hun relatie naar het raakvlak aangegeven. De eerste vervolgstappen voor beleid ten behoeve van het raakvlak betreffen:

a) het uitwerken van de beleidsbehoefte door kwantificering van publieke belangen,
b) het inpassen daarvan in opgestarte beleidsinitiatieven en
c) indien nodig aanvullende beleidsinitiatieven opstarten.

De hoge dynamiek in het raakvlak van cybersecurity en AI noopt tot een proactieve aanpak.

5 Beschouwing

In voorgaande secties is een beeld geschetst over het raakvlak van cybersecurity en AI, over de stand van zaken en de publieke belangen die ermee gemoeid zijn. Deze verkenning is opgesteld door TNO-experts in dit technische vakgebied en dient als basis voor discussie over het raakvlak tussen cybersecurity en AI. Om deze verkenning beter toepasbaar te maken voor de ontwikkeling van beleid voor het raakvlak van cybersecurity en AI zijn bijdragen nodig vanuit andere disciplines. Zoals in sectie 4.4 aangegeven is kwantificering en normering nodig van de invloed op publieke belangen ten gevolge van vraagstukken in het raakvlak van cybersecurity en AI. Hiervoor zijn aanvullende bijdragen nodig van economen en experts op het gebied van maatschappelijke waarden. Sectie 5.1 bevat suggesties voor die uitwerking van de invloed op publieke belangen, die bij de kwalitatieve inschattingen in deze verkenning naar voren zijn gekomen.

Op basis van uitgewerkte, gekwantificeerde publieke belangen kan de beleidsbehoefte worden vastgesteld. Daarmee kan vervolgens worden geanalyseerd in hoeverre recente beleidsinitiatieven daarin voorzien, welke aanscherpingen daarin aangebracht kunnen worden en/of in welke behoefte nog niet wordt voorzien. De resultaten van die analyse bieden een vertaalslag naar het opstellen van concreet beleid. Omdat de beleidsbehoefte nog niet in detail is vastgesteld is, kon in deze verkenning de mate waarin huidige beleidsinitiatieven toereikend zijn nog niet worden geanalyseerd. Suggesties voor het vertalen van de resultaten uit deze verkenning naar concreet handelingsperspectief zijn beschreven in sectie 5.2.

In parallel aan deze uitdiepingen wordt deze verkenning in de tweede helft van 2024 voortgezet door het organiseren van discussiesessies aan de hand van dit rapport. Hiervoor zal een brede groep van belanghebbenden geconsulteerd worden. De opgehaalde reacties en inzichten uit deze discussiesessies zullen worden verwerkt in een tweede editie[26] van dit rapport over deze verkenning. Ter inspiratie van de discussiesessies worden in sectie 5.3 enkele specifieke onderwerpen benoemd.

5.1 Kwantificering impact op publieke belangen

'Kwantificering' van de impact op publiek belangen betreft het specifiek aangeven van omstandigheden waarin publieke belangen dermate worden geraakt dat ingrijpen door de overheid noodzakelijk is. Voor deze normering zijn instrumenten beschikbaar, zoals de leidraad voor integrale risicobeoordeling van de nationale veiligheid [100], de Data protection impact assessment - DPIA [101] en de Impact Assessment Mensrechten Algoritmen – IAMA [102]. Ook is wet- en regelgeving beschikbaar en in ontwikkeling [4, 20, 24] waaruit normering kan worden afgeleid. Echter, tijdens deze verkenning is gebleken dat de toepasbaarheid van deze instrumenten voor specifieke vraagstukken in het raakvlak cybersecurity en AI beperkt is. Dit komt onder andere doordat:

- Zowel AI als cybersecurity zijn fundamentele technologie voor het creëren van digitale producten en diensten. De risico's die gepaard gaan met cyberincidenten

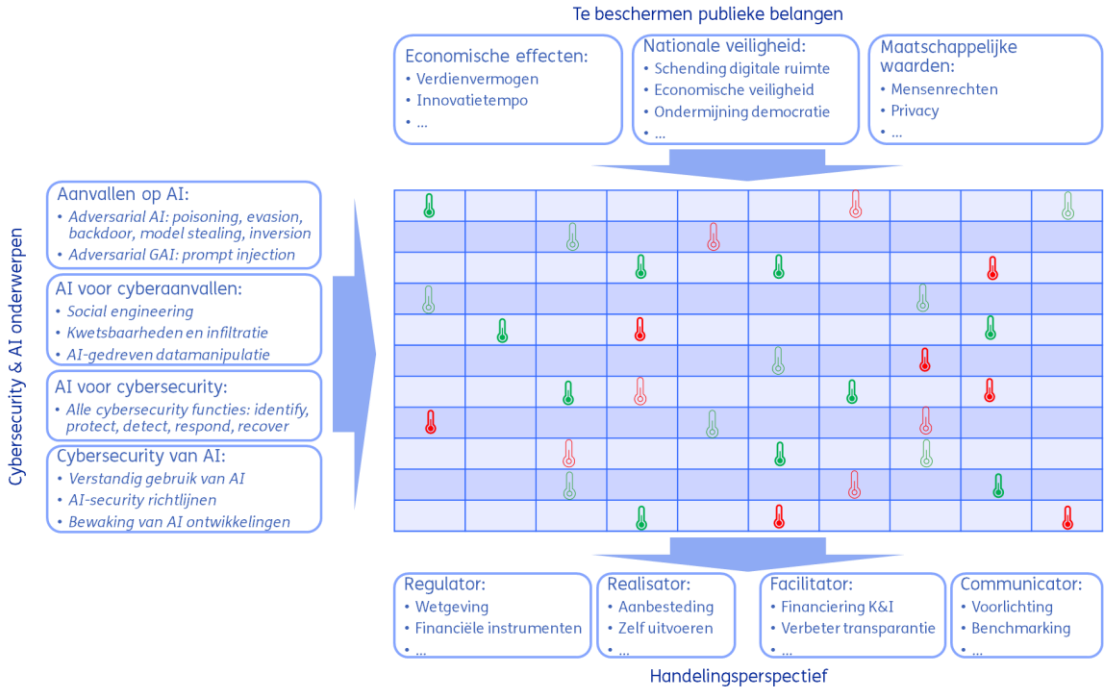
[26] In de dynamiek van dit onderwerp is de term 'definitief' niet geheel op zijn plaats. Sterker, de term 'tweede' geeft duidelijk weer dat de inhoud van dit rapport periodiek aangescherpt zou moeten worden om recht te doen aan de dynamiek van het raakvlak van cybersecurity en AI.

met AI zijn zeer nauw verweven met de toepassing van AI en de mate waarin deze een kritieke rol vervult. Bijvoorbeeld, verstoring van AI-ondersteunde functies in voertuigen vormen een veel kleiner risico dan precies diezelfde AI-functionaliteit in zelfrijdende auto's. Momenteel is de wijze waarop AI wordt ingezet voornamelijk niet-kritiek en is de omvang beperkt, maar deze kunnen snel toenemen. Het kwantificeren van die groei heeft sterke invloed op de publieke belangeninschatting, maar kent een grote onzekerheidsmarge.

- Daarnaast hebben sommige onderwerpen in het raakvlak van cybersecurity en AI een dual-use karakter. Bijvoorbeeld, AI toegepast op het vinden van kwetsbaarheden in digitale diensten kan zowel ten gunste, als ten laste van het publieke belang worden toegepast, al naar gelang de intenties van de gebruiker. Voor dual-use technologie is het lastig om de invloed op publieke belangen generiek te beoordelen.
- Sommige beoordelingsinstrumenten bieden beperkte richtlijn voor kwantificering, zoals de aan de beoordelaar overgelaten 'zwaartebepaling' waarmee grondrechten geraakt worden in IAMA en de niet nader gespecificeerde 'grootschalige' dataverwerking in DPIA.
- Andere beoordelingsinstrumenten zoals de leidraad integrale risicobeoordeling nationale veiligheid bevatten wel kwantitatieve normen, maar vergen het opstellen van concrete scenario's om daarover een kans- en impactinschatting te maken. Voor de toepassing van die leidraad is het lastig om te gaan met voorstelbare, maar nog niet opgetreden AI-security scenario's waarvan nauwelijks in te schatten is of en wanneer die scenario's zich kunnen voordoen.
- Verder is van sommige instrumenten onduidelijk hoe toepasbaar deze zijn op actuele ontwikkelingen rond generatieve AI (bijvoorbeeld IAMA).

Een voor de hand liggende manier om bestaande beoordelingsinstrumenten aan te scherpen is door beoordelingen uit te voeren van enkele concrete vraagstukken in het raakvlak van cybersecurity en AI. Die beoordelingen dienen – conform de leidraden van de betreffende instrumenten - te worden uitgevoerd door experts uit meerdere disciplines. Een uitdaging is de selectie van representatieve vraagstukken, omdat het raakvlak van cybersecurity en AI zich kenmerkt door zeer uiteenlopende vraagstukken (zie sectie 5.3). De resultaten uit deze beoordelingen kunnen worden gebruikt om (a) de bestaande instrumenten aan te vullen met specifieke richtlijnen voor cybersecurity en AI en (b) te identificeren of aanvullende beoordelingsinstrumenten nodig zijn.

Deze uitwerking van beoordelingsinstrumenten betreft overigens niet alleen vraagstukken in het raakvlak, maar ook de gezamenlijke uitdagingen van cybersecurity en AI. Bijvoorbeeld, ook instrumenten om in te schatten of er ingegrepen dient te worden om bepaalde strategische afhankelijkheden in te dammen dienen tegen het licht gehouden te worden voor vraagstukken in het raakvlak.



Figuur 10: Indicatie voor beleidsmethodiek voor het raakvlak van cybersecurity en AI.

5.2 Vertaling naar handelingsperspectief

Voor het opstellen van beleid zou een stapsgewijze methodiek moeten worden gevolgd. De beoogde stappen bestaan uit het definiëren van het raakvlak van cybersecurity en AI (inclusief het uitwerken in specifieke scenario's of vraagstukken), het beoordelen van de invloed daarvan op publieke belangen tot aan het opstellen van beleid. Deze methodiek is schematisch weergegeven in Figuur 10.

De in deze verkenning gedefinieerde kwadranten (zie linkerkant in Figuur 10) is de eerste stap in deze methodiek. De volgende stap is het vaststellen van de publieke belangen (de bovenzijde in Figuur 10) en de beoordelingen van de invloed van (vraagstukken in) de kwadranten op de publieke belangen (de 'matrix' in Figuur 10). Deze stap is in deze verkenning gedeeltelijk uitgevoerd middels kwalitatieve inschatting van de invloed op publieke belangen. In sectie 5.1 zijn vervolgacties aangegeven om die stap te voltooien.

Het resulterende overzicht van vraagstukken en hun invloed op publieke belangen vormt de basis voor besluitvorming over benodigde beleidsmatige ingrepen. In Figuur 10 is een fictief overzicht gevisualiseerd met een 'matrix' van positieve invloed op publieke belangen (groene thermometers) en negatieve invloed (rode thermometers). Uitwerking van de vertaalslag van de publieke belangen matrix naar concreet handelingsperspectief voor het raakvlak van cybersecurity en AI ligt buiten de afbakening van deze verkenning. Wel zijn tijdens deze verkenning beleidsmaatregelen geïnventariseerd die inspiratie bieden voor toekomstig handelingsperspectief.

5.2.1 Internationale samenwerking

De kansen en uitdagingen van digitale producten en diensten doen zich van nature voor in een internationaal speelveld. Dit geldt ook voor cybersecurity en AI. Het ligt dan ook voor de hand om wet- en regelgeving op te stellen in internationaal verband (zie secties 2.1.2 en 2.2.4). Ook voor het opstellen van uniforme richtlijnen voor verantwoord gebruik van (generatieve) AI (sectie 3.3.1) en cybersecurity eisen voor AI is samenwerking van belang (sectie 3.3.2 en 3.6.4). Verder is Europese samenwerking relevant ten aanzien van open strategische autonomie in zowel het speelveld van cybersecurity als van AI (sectie 3.6.1). Voor succesvolle inrichting van security & AI beleid lijkt internationale inbedding een zeer belangrijke factor.

5.2.2 Inrichting van een AI observatorium

In sectie 3.3.3 zijn een drietal initiatieven beschreven die gericht zijn op het verkrijgen van grip op ontwikkelende risico's van AI-toepassingen, zoals bijvoorbeeld AI-ondersteunde cyberaanvallen. Twee soortgelijke initiatieven zijn het opstellen en periodiek beoordelen van 'signposts of change' (sectie 3.3.3.1) en van 'critical capability levels' (sectie 3.3.3.2). Een verdergaand actieplan is opgesteld in opdracht van het Department of State van de VS, waarin de suggestie wordt gedaan voor de inrichting van een AI Observatorium. In sectie 3.3.3.3 wordt deze suggestie in meer detail beschreven.

5.2.3 Stimulering cybersecurity en AI workforce

Meerdere buitenlandse overheden zetten expliciet in op het aantrekken en ontwikkelen van de (cybersecurity en) AI-expertpopulatie[27]. Zo is in een presidentieel decreet vastgelegd dat de Amerikaanse overheid in 2025 zo'n 500 AI experts zal opleiden (sectie 2.1.2). In Finland heeft een nationaal consortium 100 AI-promovendi gefinancierd (sectie3.6.4). Ook van Nederland is van belang om meer AI-talent te werven, wil het de volledige economische potentie van AI benutten (sectie 3.6.3).

5.2.4 Onderscheid in korter en langer termijn acties

Sommige beleidsmaatregelen kunnen op korter termijn gerealiseerd worden dan andere. Om gevoel te krijgen van welk type security & AI maatregelen op welke termijn realiseerbaar zijn geeft Gladstone's actieplan [57] een nuttige indicatie. Dat plan bevat vijf actielijnen oplopend in de tijdschaal waarop bepaalde typen maatregelen gerealiseerd kunnen worden. Het plan biedt een mogelijk startpunt voor het ontwikkelen van security & AI beleid.

5.3 Security & AI vraagstukken

Zowel voor de toetsing van de inhoud van deze verkenning als voor het concreter toepasbaar maken van de resultaten worden in deze sectie enkele concrete vraagstukken in het raakvlak van cybersecurity en AI opgenomen. Ten behoeve van discussie over het raakvlak van cybersecurity en AI worden deze vraagstukken als stellingen geformuleerd.

Voor alle duidelijkheid: onderstaande stellingen zijn input voor discussie en ze betreffen geen conclusies uit deze verkenning.

[27] Deze constatering betreft hoofdzakelijk AI-expertise, maar ook voor cybersecurity expertise.

Stelling	Toelichting
De overheid dient de regie te nemen en snel security & AI beleid op te stellen en deze frequent aan te passen aan de actualiteit.	Door de snelle ontwikkelingen in beide expertisegebieden dient benadrukt te worden dat de ingeschatte impact op publieke belangen beperkt houdbaar zijn en in de komende periode bijstelling zullen behoeven. Ook geven de snelle ontwikkelingen aanleiding om tijdig voorbereid te zijn op de uitdagingen in het raakvlak van cybersecurity en AI. Een passieve houding ten aanzien van dit onderwerp kan leiden tot onwenselijke situaties waarin nauwelijks nog ingegrepen kan worden.
Het kwadrant raakvlak biedt de juiste onderverdeling voor security & AI beleid.	Gezien de verschillen in de stand van zaken en de invloed op de publieke belangen van elk van de vier security & AI kwadranten ligt voor de hand om het op te stellen beleid op te delen. De vier kwadrantgebieden bieden een geschikte onderverdeling om vier beleidsgremia in te richten die zich elk op beleidsontwikkeling voor één van de kwadranten richt.
Er moet een nationaal AI-observatorium ingericht worden.	Conform het in opdracht van het Department of State van de VS opgestelde actieplan zal een AI Observatorium worden ingericht.
Er moet een nationaal security & AI Lab ingericht worden.	Dit lab richt zich op het ontdekken, beoordelen en mitigeren van risico's in AI-systemen gedurende hun levensloop, in navolging van het opgerichte AI Assurance and Discovery Lab bij MITRE in de VS [46].
Er moet significant geïnvesteerd worden in het aantrekken en trainen van security & AI experts	De ontwikkelingen in het veld van security & AI gaan ontzettend snel, en kunnen ook op zeer korte termijn de publieke belangen beïnvloeden. Vertrouwen op buitenlandse kennis en producten creëert in toenemende mate een strategische afhankelijkheid. Daarom moet er groot worden ingezet op het vakgebied van security & AI. Talentontwikkelplannen zoals die in Finland of de VS bieden hiervoor inspiratie.

Tabel 1: security & AI stellingen als input voor discussie over op te stellen beleid.

.

6 Verwijzingen

[1] ENISA, „Cybersecurity of AI and Standardisation," ENISA, Athens, Greece, 2023.

[2] ENISA, „Artificial Intelligence and Cybersecurity Research," 7 June 2023. https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research.

[3] ENISA, „Multilayer Framework for Good Cybersecurity Practices for AI," 7 juni 2023. https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai.

[4] Europese Commissie, „EU Artificial Intelligence Act," 13 juni 2024. https://artificialintelligenceact.eu/.

[5] ISO (International Standardisation Organisation), „ISO 22989-2022: Artificial intelligence concepts and terminology," juli 2022.

[6] R. Mukhamediev, Y. Popova, Y. Kuchin, E. Zaitseva, A. Kalimoldayev, A. Symagulov, V. Levashenko, F. Abdoldina, V. Gopejenko, K. Yakunin, E. Muhamedijeva en M. Yelis, „Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges," *Mathematics,* vol. 10, nr. 15, 2022.

[7] N. Piersma, J. Kootstra, L. Henzen en M. Groen, *De AIimpactscan: Een methodiek om de impact van kunstmatige intelligentie in een specifiek werkveld in kaart te brengen,* Hogeschool van Amsterdam, 2024.

[8] Ada Lovelace Institute, „Expert explainer: The EU AI Act proposal," 8 april 2022. https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/.

[9] L. Fabri, B. Häckel, A. Oberländer, M. Rieg en A. Stohr, „ Disentangling Human-AI Hybrids: Conceptualizing the Interworking of Humans and AI-Enabled Systems.," *Business & information systems engineering, 65(6), ,* pp. 623-641, 2023.

[10] C. Liu en D. Edmondson, „China: New interim measures to regulate generative AI," augustus 2023. https://insightplus.bakermckenzie.com/bm/attachment_dw.action?attkey=FRbANEucS95NMLRN47z%2BeeOgEFCt8EGQJsWJiCH2WAWuU9AaVDeFgIGa5oQkOMGI&nav=FRbANEucS95NMLRN47z%2BeeOgEFCt8EGQbuwypnpZjc4%3D&attdocparam=pB7HEsg%2FZ312Bk8OIuOIH1c%2BY4beLEAezirm3%2BK7wMU%3D&f. [Geopend 25 juli 2024].

[11] US government, „Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Washington D.C., 2023, 30 October.

[12] Reuters institute en Oxford university, „What Does the Public in Six Countries Think of Generative AI in News?," May 2024. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-05/Fletcher_and_Nielsen_Generative_AI_and_News_Audiences.pdf.

[13] E. Brier, „Forbes 2024 AI 50 list," Forbes, 11 April 2024. https://www.forbes.com/lists/ai50/. [Geopend 31 July 2024].

[14] T. Djurickovic, „The top 10 European AI companies of this year," Tech EU, 27 December 2023. https://tech.eu/2023/12/27/the-top-10-european-ai-companies-of-2023/. [Geopend 31 July 2024].

[15] NIST, „An Introduction to Information Security - NIST SP 800-12, rev. 1," NIST, 2017.

[16] NEN, „ Information security, cybersecurity and privacy protection - Information security management systems – Requirements (ISO 27001)," 2023. https://www.nen.nl/ict/digitale-ehtiek-en-veiligheid/cyber-privacy/informatiebeveiliging.

[17] NIST, „The NIST Cybersecurity Framework (CSF) 2.0," NIST, 2024.

[18] Nationaal Coordinator Terrorismebestrijding en Veiligheid, „Overzicht vitale processen," NCTV, -. https://www.nctv.nl/onderwerpen/vitale-infrastructuur/overzicht-vitale-processen.

[19] European Commission, „The Critical Entities Resilience Directive (CER)," 25 juli 2023. https://www.critical-entities-resilience-directive.com/.

[20] European Commission, „Cyber Resilience Act - Shaping Europe's digital future," 15 september 2022. https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act.

[21] ENISA, „ENISA Threat Landscape 2023," ENISA, 2023.

[22] Dialogic, 28 april 2021. https://privacy-web.nl/wp-content/uploads/po_assets/567783.pdf.

[23] Ministerie Justietie en Veiligheid, „Overzicht wet en regelgeving cybersecurity," 3 februari 2021. https://www.rijksoverheid.nl/documenten/rapporten/2021/02/03/tk-bijlage-overzicht-wet-en-regelgeving-cybersecurity.

[24] Europese Commissie, „Cybersecurity Act," 17 april 2019. https://eur-lex.europa.eu/eli/reg/2019/881/oj.

[25] NCTV, „Nederlandse Cybersecuritystrategie 2022-2028," 10 oktober 2022. https://www.nctv.nl/onderwerpen/nederlandse-cybersecuritystrategie-2022-2028/documenten/publicaties/2022/10/10/nederlandse-cybersecuritystrategie-2022-2028.

[26] European Commission, „New EU Cybersecurity Strategy and new rules to make physical and digital critical entities more resilient," European Commission, 16 december 2020. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2391.

[27] European Parliament, „The NIS 2 Directive," december 2022. https://www.nis-2-directive.com/.

[28] European Commission, „The Critical Entities Resilience Directive (CER)," European Commission, 25 juli 2023. https://www.critical-entities-resilience-directive.com/.

[29] AIVD, „AI-systemen: ontwikkel ze veilig," Den Haag, 2023.

[30] M. Bastian, „Researchers uncover an all-too-easy trick to bypass LLM safeguards," The Decoder, 21 juli 2024. https://the-decoder.com/researchers-uncover-an-all-too-easy-trick-to-bypass-llm-safeguards/. [Geopend 1 augustus 2024].

[31] NIST, „Adversarial Machine Learning - A Taxonomy and Terminology of Attacks and Mitigations," National Institute of Standards and Technology & U.S. Department of Commerce, 2024.

[32] PyTorch, „Compromised PyTorch-nightly dependency chain between December 25th and december 30th, 2022.," PyTorch, 31 december 2022. https://pytorch.org/blog/compromised-nightly-dependency/.

[33] B. Mansi, „PyTorch dependency 'torchtriton' on PyPI Supply Chain Attack," SentinelOne, 9 januari 2023. https://www.sentinelone.com/blog/pytorch-dependency-torchtriton-supply-chain-attack/. [Geopend 25 juli 2024].

[34] wunderwuzzi, „Bing Chat: Data Exfiltration Exploit Explained,” wunderwuzzi's blog, 18 juni 2023. https://embracethered.com/blog/posts/2023/bing-chat-data-exfiltration-poc-and-fix/.

[35] M. Sharif, S. Bhagavatula, M. Reiter en L. Bauer, „Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition,” 2016.

[36] The Alan Turing Institute, „Adversarial AI: Coming of age or overhyped?,” 1 september 2023. https://cetas.turing.ac.uk/sites/default/files/2023-09/cetas_expert_analysis_-_adversarial_ai.pdf.

[37] M. Cummings, „What self-driving cars tell us about AI risks,” IEEE Spectrum, 30 Juli 2023. https://spectrum.ieee.org/self-driving-cars-2662494269.

[38] The Register, „Apple becomes the latest company to ban ChatGPT for internal use,” The Register, 19 mei 2023. https://www.theregister.com/2023/05/19/apple_chatgpt/.

[39] Business Today, „After Samsung and Amazon, Apple restricts ChatGPT use by employees: Report,” Business Today, 19 mei 2023. https://www.businesstoday.in/technology/news/story/after-samsung-and-amazon-apple-restricts-chatgpt-use-by-employees-report-381914-2023-05-19.

[40] S. Beukenkamp, „Beleid omtrent gebruik van LLM's,” Kennisplatform Digitale uitwisseling in de zorg, 27 maart 2024. https://digitaleuitwisseling.nl/threads/beleid-omtrent-gebruik-van-llms.450/.

[41] Gemeente Arnhem - MTA/CISO, „Richtlijnen voor veilig gebruik ChatGPT,” december 2023. https://arnhem.bestuurlijkeinformatie.nl/Document/View/622f6730-892b-492c-a958-d6eccce2a724.

[42] OpenAI, „Usage policies,” 10 januari 2024. https://openai.com/policies/usage-policies/.

[43] Google, „Beleid tegen verboden gebruik van generatieve AI,” Google, 14 maart 2023. https://policies.google.com/terms/generative-ai/use-policy.

[44] Rijksoverheid, „Overheidsbrede visie op generatieve AI,” 18 januari 2024. https://www.rijksoverheid.nl/documenten/kamerstukken/2024/01/18/kamerbrief-bij-overheidsbrede-visie-generatieve-ai-artificiele-intelligentie.

[45] BSI, „Security of AI-Systems: Fundamentals - Adversarial Deep Learning,” BSI, 15 augustus 2022.

[46] MITRE, „MITRE Opens New AI Assurance and Discovery Lab,” 25 maart 2024. https://www.mitre.org/news-insights/news-release/mitre-opens-new-ai-assurance-and-discovery-lab. [Geopend 23 juli 2024].

[47] CIA, „A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis,” 2009.

[48] D. Opheikens, A. Melissen, P. Stachyra en B. Poppink, „De toekomst van cyberaanvallen met Large Language Models,” 21 May 2024. https://www.ncsc.nl/documenten/publicaties/2024/mei/21/index.

[49] R. Fang, R. Bindu, A. Gupta, Q. Zhan en D. Kang, „LLM Agents can Autonomously Hack Websites,” arXiv:2402.06664v3, 2024.

[50] R. Fang, R. Bindu, A. Gupta en D. Kang, „LLM Agents can Autonomously Exploit One-day Vulnerabilities,” arXiv:2404.08144v2, 2024.

[51] R. Fang, R. Bindu, A. Gupta, Q. Zhan en D. Kang, „Teams of LLM Agents can Exploit Zero-Day Vulnerabilities,” arXiv:2406.01637v1, 2024.

[52] F. Heiding, B. Schneier en A. Vishwanath, „AI Will Increase the Quantity—and Quality—of Phishing Scams,” *Harvard Business Review,* 30 May 2024.

[53] J. Jun, „How will AI change cyber operations?," War on the Rocks, 30 April 2024. https://warontherocks.com/2024/04/how-will-ai-change-cyber-operations/. [Geopend 9 July 2024].

[54] Anca Dragan, „Introducing the Frontier Safety Framework," Google DeepMind, 17 May 2024. https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/.

[55] Joel Hestness, „Deep Learning Scaling is Predictable, Empirically," 1 December 2017. https://arxiv.org/abs/1712.00409.

[56] OpenAI, „Preparedness Framework (Beta)," 18 December 2023. https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

[57] Gladstone, „Defense in Depth: An Action Plan to Increase the Safety and Security of Advanced AI," United States Department of State (opdrachtgever), https://www.gladstone.ai/action-plan, 2024 februari.

[58] Bulletin of the Atomic Scientists, „Bulletin of the Atomic Scientists," Bulletin of the Atomic Scientists, https://thebulletin.org/.

[59] G. Connor Leahy, „EleutherAI," https://www.eleuther.ai/.

[60] „BigScience," https://bigscience.huggingface.co/.

[61] „Hugging Face," https://huggingface.co/.

[62] „together.ai," https://www.together.ai/.

[63] „Ontocord.AI," https://www.ontocord.ai/.

[64] TNO, „Steeds meer kunstmatige intelligentie ingezet door overheid," 25 juni 2024. https://www.tno.nl/nl/newsroom/2024/06/kunstmatige-intelligentie-inzet-overheid/.

[65] C. Mims, „The AI Revolution Is Already Losing Steam," *The Wall Street Journal,* pp. https://archive.ph/R5zkq#selection-5639.0-5639.41, 31 mei 2024.

[66] Lucidworks, „State of generative AI in global business, vol. 2," Lucidworks, 2024.

[67] Zscaler, „Zscaler ThreatLabz 2024 Phishing Report," Zscaler, 2024.

[68] Cisco, „What Is Social Engineering?," z.d.. https://www.cisco.com/c/en/us/products/security/what-is-social-engineering.html. [Geopend 11 juli 2024].

[69] DISARM Foundation, „DISARM Framework Navigator," https://disarmfoundation.github.io/disarm-navigator/. [Geopend 5 juli 2024].

[70] SlashNext Security, „The State of PISHING 2024 Mid-Year Assessment," Pleasnton, CA, USA, 2024.

[71] D. Gurugubelli, „ The Dark Side of AI: Unmasking the Malicious LLMs Fueling Cybercrime," Cyber Security Tribe, 6 mei 2024. https://www.cybersecuritytribe.com/articles/the-dark-side-of-ai-unmasking-the-malicious-llms-fueling-cybercrime.

[72] S. Moskal, S. Laney, E. Hemberg en U.-M. O'Reilly, „LLMs Killed the Script Kiddie: How Agents Supported by Large Language Models Change the Landscape of Network Threat Testing," arXiv:2310.06936v1, 2023.

[73] A. Happe, A. Kaplan en J. Cito, „LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks," arXiv:2310.11409v3, 2023.

[74] R. Fang, *Email exhange with the authors,* 2024.

[75] OpenAI, „OpenAI o1 System Card," 12 september 2024. https://openai.com/index/openai-o1-system-card/.

[76] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas en F. Tramèr, „Poisoning Web-Scale Training Datasets is Practical," arXiv:2302.10149v2, 2024.

[77] MITRE, „ATLAS," 2024. https://atlas.mitre.org/. [Geopend 5 Juli 2024].

[78] R. Kaur, D. Gabrijelcic en T. Klobucar, „Artificial intelligence for cybersecurity: Literature review and future research directions," *Information Fusion,* september 2023.

[79] F. Motlagh en e. al., „Large Language Models in Cybersecurity: State-of-the-Art," *Computer Science: Cryptography and Security,* 30 januari 2024.

[80] A. Sontan en S. Samuel, „The intersection of Artificial Intelligence and cybersecurity: Challenges and opportunities," *World Journal of Advanced Research and reviews ,* pp. 1720-1736, 26 februari 2024.

[81] Y. Shoshitaishvili en e. al., „Rise of the HaCRS: Augmenting Autonomous Cyber Reasoning Systems with Human Assitance," in *Conference of Complex Systems 2017*, Dallas, 2017, november.

[82] T. Avgerinos en e. al., „The Mayhem Cyber Reasoning System," *IEEE Security & Privacy, Vol. 16, issue 2,* pp. 52-60, 30 March 2018.

[83] S. Vyas, J. Hannay, A. Bolton en P. Burnap, „Automated Cyber Defense: a Review," in *ACM on Measurement and Analysis of Computing Systems*, 2023, februari.

[84] TNO, „Toekomst cybersecurity: autonoom systeem van systemen," oktober 2023. https://www.tno.nl/nl/newsroom/insights/2023/10/toekomst-cybersecurity-autonoom-systeem/.

[85] Ministerie van Buitenlandse Zaken, „Kamerbrief Open strategische Autonomie," Ministerie van Buitenlandse Zaken, Den Haag, 2022, 8 november.

[86] TNO, „Digitale Infrastructuur en Digitale Open Strategische Autonomie," TNO, Den Haag, 2023.

[87] „Onderzoeksrapportage Onderwijs en Arbeidsmarkt Cybersecurity," 2024.

[88] I. C. Group, „The Economic Opportunity of AI in the Netherlands," 2024.

[89] World Economic Forum, „Future of Jobs Report 2023," 2023.

[90] L. McMahon en Z. Kleinman, „AI Safety: UK and US sign landmark agreement," BBC, 2 april 2024. https://www.bbc.com/news/technology-68675654. [Geopend 23 juli 2024].

[91] European Commission, „G7 Leaders' Statement on the Hiroshima AI Process," 30 oktober 2023. https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process. [Geopend 23 juli 2024].

[92] European Commission, „European AI Office," 16 juli 2024. https://digital-strategy.ec.europa.eu/en/policies/ai-office. [Geopend 23 juli 2023].

[93] U.S. Department of Commerce, „U.S-EU Joint Statement of the Trade and Technology Council," 5 april 2024. https://www.commerce.gov/news/press-releases/2024/04/us-eu-joint-statement-trade-and-technology-council. [Geopend 23 juli 2023].

[94] Ministry of Foreign Affairs - Republic of Korea, „D-100 Interagency Meeting Held to Prepare for the REAIM Summit 2024," 03 june 2024. https://www.mofa.go.kr/eng/brd/m_5676/view.do?seq=322590. [Geopend 23 juli 2024].

[95] Prime Minister's Office, 10 Downing Street, „Prime Minister launches new AI Safety Institute," Gov.uk, 2 november 2023. https://www.gov.uk/government/news/prime-minister-launches-new-ai-safety-institute. [Geopend 23 juli 2024].

[96] NIST, „U.S. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE,” https://www.nist.gov/aisi. [Geopend 23 juli 2024].

[97] NIST, „U.S. Commerce Secretary Gina Raimondo Announces Expansion of U.S. AI Safety Institute Leadership Team,” 16 april 2024. https://www.nist.gov/news-events/news/2024/04/us-commerce-secretary-gina-raimondo-announces-expansion-us-ai-safety. [Geopend 23 juli 2024].

[98] Finnish Center for Artificial Intelligence, „New Finnish doctoral program in AI launching in 2024,” 9 februari 2024. https://fcai.fi/news/2024/2/9/finnish-doctoral-program-in-ai. [Geopend 7 juli 2024].

[99] Coalition for Secure AI, „Introducing the Coalition for Secure AI, an OASIS Open Project,” OASIS open projects, 18 juli 2024. https://www.coalitionforsecureai.org/introducing-the-coalition-for-secure-ai-an-oasis-open-project/. [Geopend 25 juli 2024].

[100] Analistennetwerk Nationale Veiligheid, „Leidraad risicobeoordeling Geïntegreerde risicoanalyse Nationale Veiligheid,” RIVM, https://www.rivm.nl/sites/default/files/2019-10/Leidraad%20Risicobeoordeling%202019.pdf, 2019 mei.

[101] Autoriteit Persoonsgegevens, „Data Protection Impact Assessment (DPIA),” Autoriteit Persoonsgegevens, https://www.autoriteitpersoonsgegevens.nl/themas/basis-avg/praktisch-avg/data-protection-impact-assessment-dpia, 2018 mei.

[102] Universiteit Utrecht, „Impact Assessment Mensenrechten en Algoritmes,” Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, https://www.uu.nl/sites/default/files/Rebo-IAMA.pdf, 2021 juli.

[103] TNO, „Adversarial AI in het cyberdomein,” Den Haag, 2023.

[104] Microsoft, „How AI is changing phishing scams,” 14 juli 2023. https://www.microsoft.com/en-us/microsoft-365-life-hacks/privacy-and-safety/how-ai-changing-phishing-scams. [Geopend 11 juli 2024].

[105] S. Bloemink, „Botte Hakbijl of Emancipatiemotor?,” *De Groene Amsterdammer,* 12 July 2023.

[106] S. Morgan, „Cybercrime To Cost The World $10.5 Trillion Annually By 2025,” 13 november 2020. https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/.

[107] Statista, „Financial damage from the cybercrime in Germany in 2023,” 15 maart 2024. https://www.statista.com/statistics/1360289/financial-damage-cyber-crimes-germany/.

[108] McKinsey & Company, „Cybersecurity fot the IoT: How trust can unlock value,” 7 april 2023. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/cybersecurity-for-the-iot-how-trust-can-unlock-value.

[109] Shodan, „Search Engine for the Internet of Everything,” Shodan, https://www.shodan.io. [Geopend 30 juli 2024].

[110] MIT, „AI Risk Repository,” MIT, https://airisk.mit.edu/, 2024.

[111] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, „Algoritmekader,” 2024. https://minbzk.github.io/Algoritmekader/. [Geopend 15 augustus 2024].

ICT, Strategy & Policy

Anna van Buerenplein 1
2595 DA Den Haag
www.tno.nl

TNO innovation for life