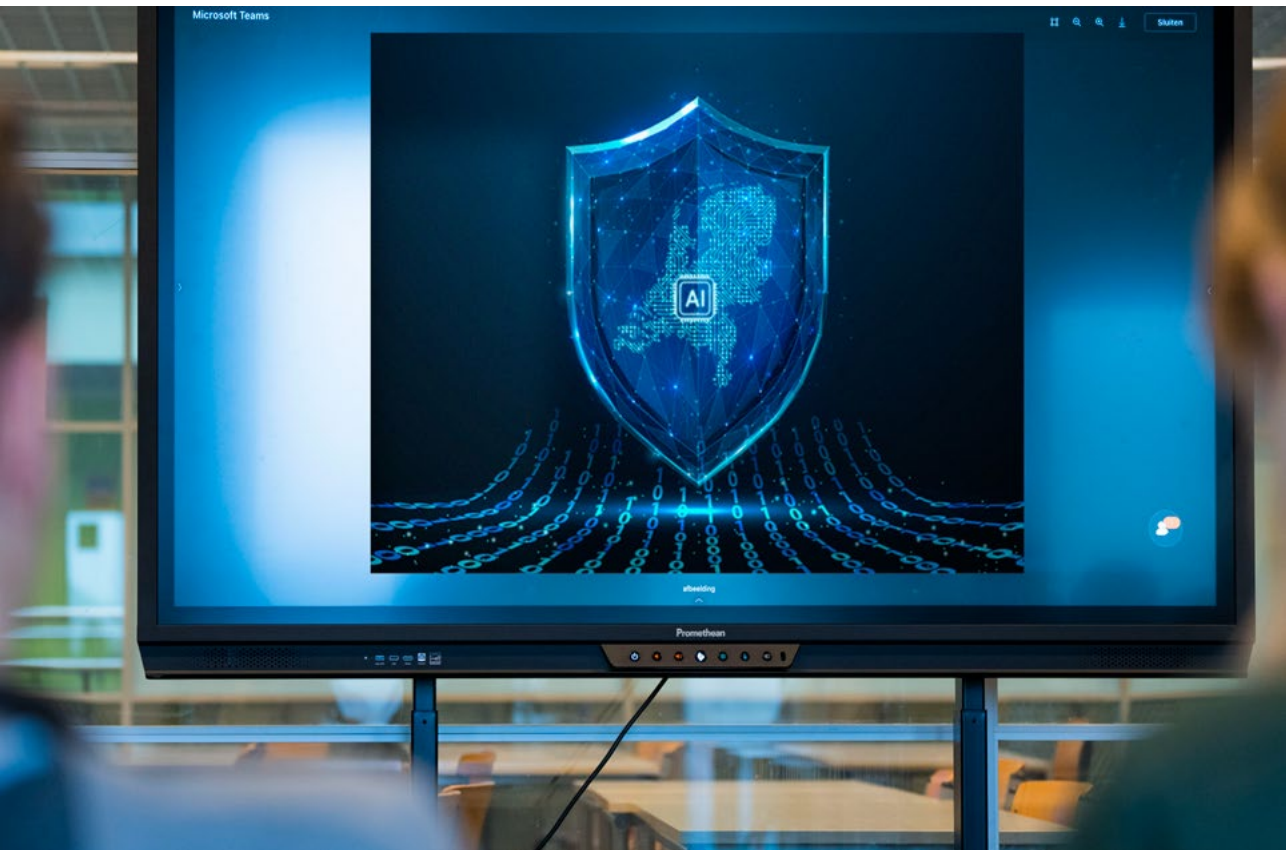
Versterkte dreigingen in een wereld vol kunstmatige intelligentie

Een analyse van het effect van AI op de nationale veiligheid



Inhoud

Kunstmatige intelligentie (*artificial intelligence of AI*) is een technologie die onze samenleving fundamenteel zal veranderen, vergelijkbaar met hoe de introductie van elektriciteit, de verbrandingsmotor en de computer dat eerder deden. De ontwikkelingen rond AI gaan snel. AI is de laatste jaren een prominent onderwerp van maatschappelijk en politiek debat, ook in het veiligheidsdomein.

Deze publicatie analyseert welk effect de technologie heeft op de veiligheid van Nederland. Het is de inschatting van de AIVD, de MIVD en de NCTV dat AI waarschijnlijk nagenoeg alle bestaande dreigingen tegen Nederland zal versterken. Dat is niet omdat AI in zichzelf een gevaar is. AI stelt machines in staat tot menselijke vaardigheden, zoals redeneren, leren, plannen en creatief denken. Als machines zulke taken sneller of efficiënter uitvoeren, kan dat het leven van mensen makkelijker maken en helpen maatschappelijke problemen op te lossen.

AI biedt kwaadwillende actoren echter de mogelijkheid zaken die vroeger meer (technische) kennis, tijd en geld kostten eenvoudiger, sneller, goedkoper en overtuigender te doen. In de handen van (buitenlandse) hackers, spionnen, terroristen, extremisten of criminelen kan AI de gevaren die zij al vormen verder vergroten. Het kan dus leiden tot meer of beter uitgevoerde sabotage, desinformatie, hacks, radicalisering of diefstal. Grootmachten investeren in de ontwikkeling van AI vanwege de militaire, digitale en economische potentie van deze technologie. AI kan hierdoor bijdragen aan een verschuiving van de machtsbalans in de wereld.

Extreme beelden doen de feiten geen recht. AI is niet de oplossing voor alle problemen, maar zal ook zeker niet het einde van de mensheid betekenen. Het is daarom belangrijk om tegelijk genuanceerd, gedetailleerd en met urgentie naar AI te kijken. Met deze analyse geven de AIVD, MIVD en NCTV een breed publiek meer inzicht in de wijze waarop AI van invloed is op de nationale veiligheid.

Hoofdstuk 1 van deze analyse geeft inzicht in wat AI is en hoe de belangrijkste soorten AI globaal werken. Hoofdstuk 2 beschrijft aan de hand van de zes nationale veiligheidsbelangen hoe landen, cyberactoren, terroristen, extremisten en criminelen misbruik van AI (kunnen) maken en welke geopolitieke belangen er zijn gemoeid met de ontwikkeling van AI. Op basis daarvan trekt hoofdstuk 3 conclusies over het effect van AI op de nationale veiligheid.

1. Wat is kunstmatige intelligentie?

Kunstmatige intelligentie (hierna AI) is een van de meest bepalende en disruptieve technologieën van deze tijd. Het maakt ons leven op veel manieren makkelijker en kan helpen bij het oplossen van tal van maatschappelijke problemen. Tegelijkertijd zit er ook een keerzijde aan deze technologische ontwikkelingen. AI-toepassingen kunnen misbruikt worden of onbedoelde neveneffecten hebben die van invloed zijn op de nationale veiligheid.[1] Hierdoor werkt AI als een versterkende factor voor bestaande dreigingen voor de nationale veiligheid.

1.1 Definitie kunstmatige intelligentie

AI omvat een grote diversiteit aan methodes, modellen en algoritmes. Hierdoor is het moeilijk een eenduidige definitie te geven die recht doet aan alle AI-technologieën en -toepassingen. De definitie die de Europese Commissie hanteert geeft een grove schets van wat AI omvat: 'systemen die intelligent gedrag vertonen door hun omgeving te analyseren en – met een graad van autonomie – actie te ondernemen om specifieke doelen te bereiken'.[2] Simpel gezegd staat AI voor de mogelijkheid van een machine om menselijke vaardigheden, zoals redeneren, leren, plannen en creativiteit, te vertonen.[3]

1.2 AI als systeemtechnologie

Vergelijkbaar met de komst van elektriciteit in de negentiende eeuw en de verbrandingsmotor in de twintigste eeuw zal AI in de eenentwintigste eeuw de samenleving fundamenteel veranderen. Wetenschappers werken al sinds de jaren '50 van de twintigste eeuw aan de ontwikkeling van AI. Pas recent is AI in een stroomversnelling geraakt en wordt het ook grootschalig in de praktijk toegepast.[4] Dit komt door de combinatie van een drietal ontwikkelingen:

1. Wetenschappelijke doorbraken op het gebied van datagestuurde AI en daarbinnen op het gebied van *deep learning*.
2. Toegenomen rekenkracht van computers.
3. De beschikbaarheid van grote hoeveelheden kwalitatief goede data. In de loop der jaren zijn mensen steeds meer gebruik gaan maken van het internet. Met al hun online activiteiten hebben ze indirect veel digitale informatie gecreëerd. Daarnaast draagt ook het *Internet of Things* bij aan de groei van beschikbare data.[5]

De drie hierboven genoemde ontwikkelingen lijken voorlopig niet tot stilstand te komen. Het is dan ook te verwachten dat AI-systemen steeds krachtiger worden en meer mogelijkheden zullen blijven krijgen.

AI wordt in steeds meer onderdelen van de samenleving praktisch gebruikt. De ontwikkelingen daarin gaan snel. Meer en meer bedrijven en onderdelen van de overheid maken gebruik van AI-toepassingen als chatbots, beslisalgoritmes en vertaal-algoritmes.[6] Daarnaast zijn grote techbedrijven zich openlijk gaan toeleggen op AI, stijgt het aantal AI-gerelateerde patenten en groeien de investeringen[7] in AI aanzienlijk.[8] Ook AI-onderzoek neemt toe. Het aantal publicaties over AI is sinds 2010 meer dan verdubbeld.[9] Tot 2014 kwamen de belangrijkste *machine learning*-modellen vanuit de academische wereld, maar sindsdien hebben techbedrijven de leidende positie overgenomen. Het bouwen van moderne AI-systemen vereist grote hoeveelheden data, computerkracht, en daarmee geld. Grote techbedrijven hebben over het algemeen meer middelen voor het ontwikkelen van AI dan non-profitorganisaties en academische instellingen.[10]

De Wetenschappelijke Raad voor het Regeringsbeleid (WRR) omschreef AI in haar rapport 'Opgave AI.' (2021) als systeemtechnologie. Een systeemtechnologie kan door de hele economie en samenleving voor allerlei doelen worden gebruikt, kent continue verbetering en maakt complementaire innovatie mogelijk. Eerdere systeemtechnologieën waren onder andere de stoommachine, elektriciteit, de verbrandingsmotor en de computer.[11]

De introductie van AI laat zich het best vergelijken met de introductie van elektriciteit. Beide hebben een immaterieel karakter. Ze bestaan meestal niet eigenstandig, maar zijn onderdeel van een product of dienst. Ook zijn beide technologieën eerst lange tijd door wetenschappers onderzocht en ontwikkeld zonder dat zij in de beginfase een idee hadden van de praktische en commerciële toepasbaarheid.[12]

AI is dus niet zomaar een technologie waarvan de werking, de toepassing en het effect ervan op de samenleving goed te overzien zijn. Het is volgens de WRR ook te allesomvattend en veelzijdig om door een groep beleidsmakers van één of enkele ministeries in goede banen te kunnen worden geleid. Bovendien is het deels nog onvoorspelbaar hoe AI onze samenleving zal veranderen. Het zal nog decennia duren voordat AI tot volle wasdom komt en in zijn volledigheid een plek in de samenleving krijgt.[13] Ondertussen verlopen de ontwikkelingen van AI-toepassingen wel in een stroomversnelling.

1.3 Verschillende soorten AI

AI-modellen kunnen voor verschillende soorten doeleinden worden ingezet. Zo kunnen ze worden gebruikt voor het herkennen en categoriseren van bestaande data met verschillende modaliteiten, zoals afbeeldingen, audiofragmenten of stukken tekst. Een voorbeeld hiervan zijn modellen die kunnen bepalen in welke taal een stuk tekst geschreven is.

Ook kunnen AI-modellen gebruikt worden voor het genereren van nieuwe data. Dit soort generatieve modellen kunnen, door te leren van patronen in bestaande content, nieuwe content genereren. Bijvoorbeeld in de vorm van tekst, computercode en audio- en videofragmenten. Veel van de AI-toepassingen die van invloed zijn op de nationale veiligheid en in deze analyse aan de orde komen behoren tot de generatieve AI. Daarom wordt deze vorm van AI in paragraaf 1.4 verder uitgelicht.

AI is grofweg in drie verschillende groepen te verdelen. Deze geven inzicht in de werking en ontwikkeling van AI: symbolische AI, datagestuurde AI en mogelijke toekomstige vormen van AI.

Symbolische AI
De eerste en vroege vormen van AI staan bekend als symbolische AI. Symbolisch omdat deze vorm van AI regels volgt die zijn uit te drukken in symbolen. Bijvoorbeeld een formule van het type 'als X, dan Y'. Deze vorm van AI is gebaseerd op precieze, door mensen gemaakte algoritmes die een computer stap voor stap volgt om te bepalen hoe deze op een bepaalde situatie reageert. De intelligentie in het systeem komt dus rechtstreeks van menselijke expertise en wordt vastgelegd in een format waarmee de computer kan werken. Symbolische AI werkt het best in een afgebakende en stabiele omgeving. Het algoritme kan alleen taken uitvoeren op de manier waarop de computer is geïnstrueerd en verbeteringen zijn alleen door menselijk ingrijpen mogelijk. Hoewel deze vorm van AI achterhaald klinkt, wordt deze nog steeds breed toegepast. Zowel in een simpele thermostaat in huis als in geavanceerde robotica.[14]

Datagestuurde AI
De tweede vorm van AI staat bekend als datagestuurde AI of *machine learning*. Deze termen verwijzen naar een breed scala aan technieken die het leerproces van algoritmes automatiseren. Een computer wordt hierbij niet expliciet geprogrammeerd voor het uitvoeren van een bepaalde taak, maar het algoritme wordt ontworpen om (op basis van historische data) patronen te herkennen. Zo is het in staat intelligent gedrag te vertonen.[15]

Datagestuurde AI kan op verschillende manieren getraind worden. Trainen is het proces waarbij het AI-model taken aangeleerd krijgt die normaal gesproken menselijke intelligentie vereisen, zoals het herkennen van afbeeldingen, spraak of tekst. Bij *supervised learning* wordt gebruik gemaakt van data met labels, bijvoorbeeld foto's van dieren met de labels 'kat' en 'niet-kat'. Wanneer een algoritme op een set data met labels is getraind, wordt vervolgens getest of het die labels ook op nieuwe data goed kan toepassen. Deze vorm is vooral geschikt wanneer er voldoende gelabelde data beschikbaar is. Bij *unsupervised learning* wordt het algoritme met grote

hoeveelheden data zonder labels gevoed en gaat het zelf op zoek naar patronen in die data. Bij *reinforcement learning* wordt het algoritme getraind door het te belonen als het strategieën kiest waarmee de vooraf gestelde doelen worden behaald.[16]

Deep learning is een vorm van datagestuurde AI, geïnspireerd door de werking van neuronen in het menselijk brein. De hierbij gebruikte, zogenoemde neurale netwerken zijn in staat om zeer complexe patronen in data te herkennen. *Deep* verwijst naar het gebruik van een (zeer) grote hoeveelheid lagen in het trainingsproces. Hierbij geeft elke laag, op basis van de voorgaande, een complexere representatie van de input. Eerdere lagen in het model kunnen zichzelf bijvoorbeeld representaties van lijnen, stippen of andere kleine patronen aanleren. De lagen iets dieper in het model leren vervolgens grotere objecten en patronen te herkennen, zoals een curve van een ooglid. De latere lagen gebruiken deze informatie om onderscheid te maken tussen complexere patronen zoals een oog, mond of neus. Op basis hiervan besluit het model of het om een gezicht gaat of niet. Deep learning wordt onder andere toegepast in technologieën zoals beeld- en spraakherkenning, large language models en zelfrijdende auto's.[17]

Mogelijke toekomstige vormen van AI

De AI die we nu kennen wordt ook wel omschreven als *weak of narrow* AI, omdat deze altijd is gericht op een specifieke vaardigheid, zoals beeld- of spraakherkenning. Omdat de ontwikkeling van AI snel gaat wordt er ook volop nagedacht en gespeculeerd over de mogelijkheden van nog intelligentere AI. In dit kader wordt vaak gesproken over *artificial general intelligence* (AGI), ook wel *strong* of *full* AI genoemd. AGI verwijst naar algoritmes die intelligentie kunnen vertonen in een breed scala aan contexten en probleemgebieden. Met andere woorden, AI die in staat is om alle menselijke intelligente vaardigheden te begrijpen en te simuleren.[18]

Artificial super intelligence (ASI) is de overtreffende trap van AGI en verwijst naar AI die intelligenter is dan mensen en opereert op een manier die door mensen niet meer goed te interpreteren is. Nog een ander hypothetisch toekomstbeeld is technologische singulariteit. Dit verwijst naar het moment waarop AI dusdanig intelligent en autonoom wordt dat het zelf nog intelligentere en autonome AI kan generen en loskomt van menselijke controle.[19]

Op dit moment bestaan er geen modellen die als AGI, laat staan als ASI, kunnen worden gezien. AI kan weliswaar op steeds meer specifieke terreinen beter presteren dan mensen, er is nog geen model dat dit op de volle breedte van de menselijke intelligentie doet. ChatGPT en soortgelijke toepassingen kunnen bijvoorbeeld op basis van bestaande data heel goed en razendsnel teksten, computercode, afbeeldingen, audio en video genereren. Alles wat nieuw lijkt is vooralsnog gebaseerd op historische datasets. Het is nog maar de vraag of het model in staat is om zelf met echt nieuwe ideeën te komen. Laat staan dat het net als mensen eigen gevoelens of verlangens heeft. Hoewel sommige techondernemers beweren dat er op korte termijn AGI-modellen gerealiseerd zullen worden, zijn de meeste experts het er over eens dat het nog maar de vraag is of dergelijke AI ooit realiteit wordt.[20]

1.4 Generatieve AI

Generatieve AI is voortgekomen uit datagestuurde AI. Generatieve AI is niet alleen in staat om patronen in data te herkennen, maar creëert zelf ook nieuwe content in de vorm van taal, computercode, afbeeldingen, audio en video. Dit doet generatieve AI op basis van de data waarop het is getraind en de door de gebruikers gegeven opdrachten en vragen. De ontwikkeling van verschillende generatieve AI-toepassingen gaat momenteel razendsnel en de uiteindelijke impact hiervan op de samenleving kent nog veel onduidelijkheden.

Large language models (LLM's), in het Nederlands taalmodellen genoemd, worden gebruikt om teksten te genereren. Wanneer een LLM een 'vraag' krijgt, zal het, stukje voor stukje, een 'antwoord' genereren op basis van de aangeleerde relaties tussen de woorden in de vraag, en de stukjes tekst die het al voorspeld heeft. Om dit te bereiken worden LLM's getraind om patronen in grote hoeveelheden tekst te kunnen herkennen.

Tot voor kort bestonden er (bijna) alleen AI-modellen die één soort content konden genereren. Inmiddels zijn er ook multimodale generatieve modellen die tegelijkertijd met tekst, code, audio, video en beeld om kunnen gaan. De gegenereerde content is vaak niet te onderscheiden van door mensen gemaakte content.

De meeste mensen kennen ChatGPT als vorm van een generatieve AI toepassing. Deze chatbot, gebaseerd op een LLM, werd in november 2022 gelanceerd door het Amerikaanse OpenAI. In 2023 volgden verschillende techbedrijven met vergelijkbare producten. Zo lanceerde Google zijn chatbot *Gemini* (voorheen Bard) en lanceerde Meta *Llama*, waarmee bedrijven en onderzoekers generatieve AI-toepassingen kunnen maken.[21] Microsoft integreerde een AI-chatbot in zijn *Bing* zoekmachine.[22] Daarnaast zijn er verschillende open-source modellen die gebruikt kunnen worden voor het ontwikkelen van eigen chatbots. In november 2023 stelde het Ministerie van Economische Zaken en Klimaat 13,5 miljoen euro beschikbaar voor de ontwikkeling van een Nederlands LLM genaamd *GPT-NL*. TNO, het Nederlands Forensisch Instituut en SURF, een samenwerkingsverband van universiteiten en hogescholen, ontwikkelen het model dat bedoeld is voor academische instellingen, onderzoekers en overheden.[23]

Ook grote Chinese bedrijven zoals Alibaba, ByteDance en Tencent werken aan de ontwikkeling van generatieve AI-modellen en stellen die openbaar beschikbaar. Deze modellen zijn getraind op Chinese data en hebben in tegenstelling tot Amerikaanse en Europese modellen te maken met de strenge Chinese internetcensuur.[24] Daarnaast investeren ook Rusland en Saoedi-Arabië in de ontwikkeling van generatieve AI-modellen.[25]

Tekortkomingen generatieve AI

Hoewel veel moderne generatieve AI-modellen zeer overtuigende teksten, computercode, afbeeldingen, audio en video's kunnen genereren, kennen ze nog wel tekortkomingen. LLM's kunnen bijvoorbeeld logisch klinkende teksten produceren, maar zijn niet in staat de inhoud zelf te begrijpen. Hierdoor combineren ze mogelijk feiten en fictie in de teksten die ze genereren

en kunnen ze zeer zelfverzekerd onsamenhangende of onjuiste output leveren.[26] Dit maakt ze vooralsnog minder geschikt voor toepassingen waarin feitelijke correctheid benodigd is.[27] Het onbedoeld genereren van onwaarheden wordt vaak aangeduid als het 'hallucineren' van AI. Hallucineren is inherent aan de werking van de huidige populaire generatieve modellen en wordt ook versterkt omdat ze zijn getraind op enorme datasets, meestal afkomstig van het internet, die onjuiste of misleidende informatie kunnen bevatten.[28] Generatieve AI dreigt zichzelf ook te vervuilen. Teksten en afbeeldingen op het internet waren tot voor kort oorspronkelijk door mensen gemaakt. Nu het aantal websites dat geheel door AI is gevuld toeneemt, wordt AI dus ook getraind op data die door AI gegenereerd is. Wanneer een AI-model op zijn eigen synthetische data wordt getraind, worden de resultaten van het model over het algemeen kwalitatief slechter.[29]

Daarnaast zijn deze generatieve AI-modellen niet altijd veilig in gebruik. De zoekopdrachten en informatie die iemand in deze modellen invoert, worden door hun aanbieders online opgeslagen. Die informatie kan, bijvoorbeeld door een hack of een fout in het systeem, openbaar worden.[30] Begin 2023 konden gebruikers van ChatGPT de titels van de chats van andere gebruikers zien als gevolg van een bug in het systeem.[31] Verschillende grote bedrijven, zoals Apple en Samsung, verboden hun werknemers dergelijke AI-tools te gebruiken, uit angst dat daarmee gevoelige informatie op straat komt te liggen.[32]

Ook zijn er zorgen over de verwerking van persoonsgegevens door generatieve AI. Zo heeft de Autoriteit Persoonsgegevens (AP) in juni 2023 OpenAI om opheldering gevraagd over de omgang met persoonsgegevens bij het trainen van ChatGPT.[33] Generatieve AI maakt (soms) tevens inbreuk op de rechten van auteurs en kunstenaars. De LLM's zijn bijvoorbeeld getraind op bestaande teksten, audio en beelden, maar zijn niet in staat om bij de output die ze genereren elementen toe te schrijven aan een oorspronkelijke auteur.[34]

Tot slot is een probleem van deep learning modellen het steeds groter wordende gebrek aan transparantie en uitlegbaarheid van hoe modellen tot een uitkomst komen. Door de beschikbaarheid van meer geld en hardware worden modellen over het algemeen steeds groter, met als resultaat dat ontwikkelaars steeds minder weten over de interne werking. Deep learning modellen bestaan immers uit een groot aantal numerieke parameters die samen tot een uitkomst (voorspelling) komen. Waar het bij kleine modellen nog enigszins af te leiden valt welke onderdelen van een model meer bijdragen aan een bepaalde voorspelling dan andere, is dit bij grotere modellen praktisch onmogelijk. Zo bestaan moderne taalmodellen van grote partijen uit vele miljarden parameters. Het is een grote en mogelijk gevaarlijke tekortkoming van moderne taalmodellen, dat we niet overzien of begrijpen hoe de AI tot zijn voorspellingen en beslissingen komt.

2. AI en dreigingen voor de nationale veiligheid

AI maakt momenteel in hoog tempo de overstap van het lab naar de samenleving en zal deze de komende jaren fundamenteel veranderen. Dat brengt economische groei en innovaties met zich mee, maar het kan ook bestaande dreigingen voor de nationale veiligheid versterken. Hieronder wordt aan de hand van de zes nationale veiligheidsbelangen in kaart gebracht wat mogelijke dreigingen van AI voor de nationale veiligheid zijn (zie bijlage 1 voor een overzicht van de nationale veiligheidsbelangen en hun impactcriteria).[35] Voor AI geldt dat de technologie zelf geen dreiging voor de nationale veiligheid is. De dreiging schuilt in de specifieke toepassing van de techniek en de intenties van de actor die hier gebruik van maakt. Zo kan de één ChatGPT gebruiken om snel een omvangrijk onderzoeksrapport samen te vatten en de ander dezelfde toepassing misbruiken om kwaadaardige phishing-mails te schrijven. AI-technologie voor zelfrijdende voertuigen heeft de potentie ons verkeer veiliger te maken, maar het kan ook worden toegepast in vernietigende autonome wapens. Een techniek als deepfakes kan educatieve waarde hebben wanneer deze wordt gebruikt om historische figuren digitaal weer tot leven te wekken, terwijl dezelfde techniek kan worden misbruikt voor het genereren van desinformatie.

2.1 AI en territoriale veiligheid

Territoriale veiligheid omvat onder andere de integriteit van het Nederlandse grondgebied, de integriteit van de digitale ruimte (IT netwerken en informatiesystemen) en de integriteit van het bondgenootschappelijk grondgebied.[36]

Statelijke actoren integreren AI steeds meer in militaire toepassingen zoals autonome wapensystemen en modellen die helpen bij het bepalen van aanvals- en verdedigingsstrategieën. Hiermee versterkt het gebruik van AI door statelijke actoren nu al bestaande dreigingen voor de territoriale veiligheid van Nederland en het bondgenootschappelijk grondgebied. Daarnaast gebruiken cyberactoren AI in cyberaanvallen, bijvoorbeeld voor het schrijven van phishing-mails of malware. Hierdoor draagt AI ook bij aan de bestaande dreigingen voor de digitale ruimte (als onderdeel van territoriale veiligheid) en versterkt het de cyberdreiging gericht op de overige veiligheidsbelangen. Naarmate AI geavanceerder wordt en breder wordt toegepast in het militaire- en cyberdomein kan het effect hiervan op de nationale veiligheid nog groter worden.

2.1.1 Het gebruik van AI in oorlogsvoering

AI verandert alle aspecten van het militaire domein. Vooral het vermogen van AI om in een complexe omgeving veel sneller en nauwkeuriger te kunnen waarnemen, beslissen en handelen dan mensen maakt het interessant voor militair gebruik. Zo kan AI militairen helpen bij het voorbereiden en uitvoeren van operaties, het analyseren van inlichtingen en wordt AI steeds meer toegepast in wapensystemen.[37] De tijd die benodigd is voor besluitvorming kan bijvoorbeeld worden teruggebracht van dagen tot minuten of zelfs seconden.[38] Dit zal een voordeel bieden voor het voorbereiden en uitvoeren van operaties en ook kan het worden toegepast in het inlichtingendomein. Daarmee zullen factoren zoals dataverzameling, connectiviteit, rekenkracht, algoritmes en systeembeveiliging nog meer van belang worden op het slagveld.[39] Daarnaast kan generatieve AI in militaire conflicten worden gebruikt voor het genereren van desinformatie (zie ook paragraaf 2.5.2.), voor het genereren van scenario's als input voor strategische besluitvorming en voor training- en simulatiedoeleinden (ook wel *wargaming* genoemd).[40] Nederland moet er dus rekening mee houden dat AI en autonome wapens een belangrijker onderdeel van oorlogsvoering worden.

Het idee dat technologie zelfstandig beslist over leven en dood kent ethische bezwaren. Voorbeelden van de ontwikkeling van wapensystemen die opereren op autonome basis zijn zwermtechnologie en *lethal autonomous weapons systems* (LAWS). Internationaal zijn er verschillende acties geweest om autonome wapens te verbieden, maar dit is in de praktijk lastig. Dat heeft onder andere te maken met het *dual-use*-karakter van autonome wapens. Technologie zoals lokalisatie, mapping, objectdetectie, dynamische navigatie, objecten volgen en gezichtsherkenning worden ook in allerlei vreedzame civiele toepassingen gebruikt. Ook is het lastig te definiëren wanneer een wapen autonoom is of wanneer slechts bepaalde functies geautomatiseerd zijn. Daarnaast verzetten landen met sterke legers zich tegen beperkingen op dit terrein.[41] Hierdoor is de ontwikkeling van militaire AI moeilijk te reguleren.

Voor AI in militaire toepassingen geldt dat uiteindelijk de intenties van de actor die er gebruik van maakt bepalen of deze de nationale veiligheid of die van bondgenoten beschermen of bedreigen. In toenemende mate vinden de militaire toepassingen van AI hun weg naar het operationele domein. De oorlog in Oekraïne wordt bijvoorbeeld gebruikt als laboratorium voor nieuwe militaire technologie, waaronder AI. Zo wordt AI gebruikt om grote hoeveelheden informatie, verzameld door drones, spionagesatellieten en andere informatiebronnen, samen te voegen en te analyseren.[42] Ook Israël maakt in de oorlog tegen Hamas gebruik van verschillende AI-systemen. Zo zou, volgens verschillende berichten in The Guardian, het AI-systeem genaamd 'the Gospel' snel en geautomatiseerd inlichtingen verzamelen en aanbevelingen doen voor militaire doelwitten.[43] Ten minste dertig landen hebben defensiesystemen die, wanneer zij ingeschakeld zijn, autonoom inkomende gevaren zoals raketten kunnen neerhalen. Zuid-Korea heeft bijvoorbeeld in de gedemilitariseerde zone op de grens met Noord-Korea een robotwapen dat autonoom kan schieten op bewegende objecten.[44]

China is een van de landen die op verschillende manieren investeert in het ontwikkelen van AI in het militaire domein.[45] Het land streeft er naar om in 2030 wereldwijd dominant te zijn op het gebied van AI en heeft hiervoor een ontwikkelingsplan opgesteld. In dat plan wordt de militaire toepassing van AI benadrukt. Daarnaast streeft China ernaar om uiterlijk in 2049 te beschikken over een krijgsmacht van wereldklasse. Om dit doel te bereiken werkt het Chinese Volksbevrijdingsleger aan een transformatie tot hoogwaardig technologische krijgsmacht.[46] China ziet AI als de manier om de conventionele militaire superioriteit van de VS te compenseren.[47] Zo ontwikkelt het land AI-toepassingen die besluitvorming ondersteunen, LAWS en zwermtechnologie.[48] Met zwermtechnologie kunnen grote aantallen drones door middel van AI als een gecoördineerde zwerm opdrachten uitvoeren, waarmee bijvoorbeeld luchtverdedigingssystemen overbelast kunnen worden.[49] Daarnaast gebruikt China AI voor het vergaren van inlichtingen en het automatiseren van surveillance- en verkenningsactiviteiten.[50]

Ook Rusland heeft het belang van sleuteltechnologieën als kwantumtechnologie en AI onderkend en heeft de ambitie om militaire toepassingen van deze technologieën te ontwikkelen.[51] In het bijzonder ziet Rusland AI als een strategische 'game changer' op het gebied van cyberoorlog. Er zijn echter ook twijfels of Rusland haar achterstand op AI kan inlopen en deze ambitie kan waarmaken.[52]

AI kan ook worden ingezet ter bescherming van de territoriale veiligheid. In juni 2023 presenteerde het Nederlandse ministerie van Defensie de eerste 'Defensie Strategie Data Science en AI 2023-2027'. Hierin stelt Defensie dat moderne (wapen)systemen bijna niet inzetbaar zijn zonder AI en dat het gebruik van informatie een steeds prominentere en strategische rol vervult. De komende jaren investeert Defensie daarom onder andere in het toepassen van AI bij onbemande autonome systemen, militaire besluitvormingsondersteuning en inlichtingen.[53]

2.1.2 Het gebruik van LLM's door cyberactoren

Het gebruik van AI door cyberactoren, zoals statelijke actoren met een offensief cyberprogramma en cybercriminelen, vormt een dreiging voor de digitale ruimte. Hoewel de invloed van AI op cyberaanvallen niet duidelijk zichtbaar aanwezig is, kan AI wel een faciliterende rol hebben bij bestaande cyberaanvallen. Hierdoor vergroot AI in potentie de mogelijkheden om cyberaanvallen uit te voeren. Dit heeft ook effect op andere veiligheidsbelangen, bijvoorbeeld wanneer AI in cyberaanvallen wordt gebruikt tegen vitale infrastructuur of bij statelijke inmenging. De afgelopen jaren zijn verschillende LLM's en toepassingen daarvan zoals ChatGPT, Gemini en Llama voor het grote publiek beschikbaar gekomen. Al snel kwamen er ook berichten naar buiten over misbruik van deze systemen door kwaadwillende actoren in cyberoperaties.[54] Er is een aantal manieren waarop cyberactoren LLM's kunnen gebruiken:

1. **Informatie verzamelen en detecteren van doelwitten**
   Cyberactoren kunnen LLM's gebruiken om snel en automatisch interessante doelwitten te detecteren en informatie over mogelijke doelwitten te verzamelen. Onderzoekers toonden bijvoorbeeld aan dat ChatGPT gebruikt kan worden om snel informatie te vergaren over het IT-systeem van een bank. Hoewel dit soort informatie vaak ook openbaar op het internet te vinden is, kunnen LLM's het proces om deze informatie te verzamelen aanzienlijk versnellen.[55]

2. **Phishing-mails genereren**
   Cyberactoren kunnen LLM's gebruiken om geautomatiseerd en snel zeer overtuigende spam en phishing-mails te schrijven. LLM's bieden de mogelijkheid om foutloze teksten in verschillende talen te genereren, ook als een cyberactor die taal zelf niet machtig is.[56] Bovendien kunnen cyberactoren met deze LLM's snel reageren vanuit een bepaalde context, specifieke schrijfstijlen aannemen en daarmee overtuigend een specifiek persoon na doen.[57] Spearphishing-mails, die met behulp van AI op de ontvanger zijn toegesneden, geven een grotere kans dat de ontvanger die als betrouwbaar beoordeelt.[58]

3. **Malware genereren en cyberaanvallen uitvoeren**
   LLM's stellen cyberactoren in staat om cyberaanvallen uit te voeren die ze technisch gezien niet zouden kunnen doen zonder de hulp van AI. Cyberactoren die deze technische kennis wel hebben, kunnen dat met AI sneller en eenvoudiger doen.[59] Er bestaan LLM's die (ook) zijn getraind op verschillende programmeertalen, waardoor cyberactoren ze kunnen gebruiken bij het schrijven van malware.[60] Cyberactoren kunnen LLM-toepassingen daarnaast gebruiken als hulpmiddel bij het uitvoeren van een cyberaanval, en als het ware sparren met het model om een effectieve aanval te ontwikkelen. Ook is het mogelijk dat LLM's gebruikt worden voor het vinden van kwetsbaarheden in code. Malware zelf kan door AI ook intelligenter worden doordat het zelflerende eigenschappen krijgt waarmee het zelfstandig nieuwe kwetsbaarheden kan vinden. Doordat aanvallen steeds beter vermomd worden, kan malware lange tijd onopgemerkt geïnstalleerd zijn op een apparaat en zich

verspreiden naar andere apparaten en netwerken. Daarbij kan AI malware ook helpen snel grote hoeveelheden data te analyseren om te identificeren wat waardevol is en wat niet.[61]

De meeste techbedrijven bouwen beveiliging in hun LLM-applicaties om te voorkomen dat deze kwaadaardige teksten, gevaarlijke informatie of kwaadaardige code genereren.[62] Cyberactoren zoeken echter naar manieren om dit soort veiligheidsmaatregelen te omzeilen, onder andere door de opdracht die de gebruiker aan het LLM geeft zo te formuleren dat deze alsnog kwaadaardige output genereert. Cyberactoren hebben dit bijvoorbeeld succesvol gedaan door opdrachten op te delen in individuele stappen.[63]

Cybercriminelen ontwikkelen daarnaast ook hun eigen generatieve AI-tools. Omdat LLM's geen ethische grenzen en beperkingen hebben, kunnen ze eenvoudiger worden gebruikt voor het genereren van phishing-mails en kwaadaardige code. Voorbeelden van dit soort LLM's zijn WormGPT en FraudGPT. Deze modellen worden onder andere aangeboden op online forums, darkweb-marktplaatsen en Telegram-accounts.[64]

Naarmate LLM's in de toekomst nog geavanceerder en intelligenter worden, zullen ook de mogelijkheden om er misbruik van te maken toenemen.[65] De huidige LLM's kennen nog steeds beperkingen, maar ze hebben steeds minder nauwkeurigere instructies nodig. Hun vermogen om code te genereren en omgevingspatronen te herkennen maakt deze LLM's in combinatie met malware een nieuw soort dreiging die in de toekomst zeer gevaarlijk kan zijn.[66] Experts maken zich onder andere zorgen over het risico van volledig geautomatiseerde en autonome end-to-end-aanvallen wanneer AI nog geavanceerder wordt.[67]

Het vermogen van AI-algoritmes om grote hoeveelheden gegevens snel te analyseren en afwijkende patronen te herkennen kan ook worden gebruikt bij de verdediging tegen cyberaanvallen.[68] Cybersecuritybedrijven gebruiken AI inmiddels dan ook in hun producten.[69] Zo kan AI netwerkverkeer analyseren om ongebruikelijk gedrag te detecteren dat wijst op potentiële cyberdreigingen. AI kan ook worden gebruikt voor het verwerven en verwerken van bedreigingsinformatie en het analyseren van malware.[70]

2.2 AI en fysieke veiligheid

Fysieke veiligheid omvat het ongestoord functioneren van individuen. Deze kan bijvoorbeeld in het geding komen door ongelukken, aanslagen, ziekten of gebrek aan primaire levensbehoeften.[71]

Op dit moment vormt AI nog geen substantiële dreiging voor de fysieke veiligheid in Nederland. De meeste AI-toepassingen zijn tenslotte digitaal en spelen vooralsnog een beperkte rol bij fysieke interacties met mensen. Verschillende ontwikkelingen laten zien hoe AI in de toekomst mogelijk wel een dreiging voor de fysieke veiligheid kan worden.

Ten eerste, is het in theorie mogelijk dat bepaalde militaire AI-toepassingen op enig moment breder beschikbaar worden en in handen vallen van terroristische of criminele groeperingen. In dat geval zouden deze AI-toepassingen ook ingezet kunnen worden bij het plegen van aanslagen.

Hier zijn momenteel geen aanwijzingen voor en dit is een ontwikkeling die de komende jaren nog niet voor de hand ligt.

Ten tweede, kunnen de ontwikkelingen in de robotica er aan bijdragen dat er meer AI-toepassingen in het fysieke domein terecht komen. Denk bijvoorbeeld aan de introductie van de zelfrijdende auto. Dergelijke ontwikkelingen zouden nieuwe kwetsbaarheden voor sabotage met zich meebrengen. AI kan namelijk gevoelig zijn voor manipulatie. Zo kan een algoritme bedrogen worden door kleine maar onmerkbare wijzigingen aan te brengen, wat leidt tot een verkeerde classificatie van dat wat het algoritme onderzoekt. In bepaalde toepassingen kunnen kleine veranderingen echter grote gevolgen hebben, bijvoorbeeld wanneer AI wordt toegepast in zelfrijdende auto's en iemand er in slaagt deze zodanig te saboteren dat de auto juist door een rood stoplicht rijdt.[72]

Ten derde, is het niet uitgesloten dat AI-modellen gebruikt worden voor de ontwikkeling van chemische en biologische wapens, zoals een besmettelijk virus.[73] Onderzoekers van onderzoeksorganisatie RAND testten recent de mogelijkheid om met LLM's biologische wapens te maken en concludeerden dat de huidige generatie modellen hier niet toe in staat zijn.[74] Andere onderzoekers zijn echter in staat geweest om met generatieve AI biologisch actieve eiwitten te ontwikkelen die niet in de natuur voorkomen.[75] De mogelijkheid om dergelijke wapens te maken met behulp van AI kan daarom ook in potentie bijdragen aan de dreiging voor de fysieke veiligheid.[76]

2.3 AI en economische veiligheid

Economische veiligheid omvat het ongestoord functioneren van Nederland als een effectieve en efficiënte economie. Hierbij gaat het onder andere over de continuïteit van vitale processen, het mitigeren van risicovolle strategische afhankelijkheden en het voorkomen van de ongewenste overdracht van kennis en technologie.[77]

Naarmate AI-toepassingen breder worden omarmd in het bedrijfsleven en de maatschappij, groeit de afhankelijkheid van de samenleving van grote techbedrijven en staten die deze AI-toepassingen leveren en onderhouden. Gezien de potentie van AI kan deze afhankelijkheid op enig moment zodanig groot worden dat er risicovolle strategische afhankelijkheden ontstaan en dergelijke partijen de mogelijkheid krijgen om de economische veiligheid te raken als ze niet hun zin krijgen van de Nederlandse overheid. Zo kan het lastiger worden om bij dergelijke techbedrijven en andere staten aan te dringen op hogere standaarden voor de bescherming van persoonsgegevens en het borgen van de Nederlandse democratische waarden op sociale media. Denk hierbij aan de vrijheid van meningsuiting en het recht op gelijke behandeling. Dit zou in potentie ook kunnen raken aan het veiligheidsbelang van sociale en politieke stabiliteit.

2.3.1 Afhankelijkheid grote techbedrijven

In de ontwikkeling van AI-toepassingen hebben grote private techbedrijven tot nu toe de overhand. Veel van deze tech-giganten zijn Amerikaanse of Chinese multinationals, zoals Google en Huawei.[78] Deze monopolies ontstaan doordat er voor nieuwe innovaties en doorbraken op het gebied van AI veelal grote hoeveelheden rekenkracht, data en daarmee geld nodig zijn.[79]

De Europese markt loopt achter op de Amerikaanse en Chinese markt. Europa is grotendeels afhankelijk van de grote techbedrijven voor het verkrijgen van de nieuwste AI-gerelateerde software en diensten, zoals modellen voor beeldherkenning en AI-managementtools.[80] Naarmate de digitalisering van processen en de afhankelijkheid van AI-technologie voor basisvoorzieningen toeneemt, en daar de software en diensten van tech-giganten de enige uitkomst voor bieden, neemt de afhankelijkheid van diezelfde bedrijven toe. De toegankelijkheid en betrouwbaarheid van de software is immers in grote mate afhankelijk van de aanbieder. Bovendien mengen grote techbedrijven zich ook in de discussie met overheden in de wijze waarop democratische vrijheden gewogen moeten worden. Het conflict tussen het sociale media platform X en het Braziliaanse Hooggerechtshof is daar een voorbeeld van. Als de interpretatie van democratie van deze techbedrijven anders is dan die van onze Nederlandse democratische rechtsorde, dan kunnen techbedrijven ook invloed hebben op het veiligheidsbelang van sociale en politieke stabiliteit.

2.3.2 Afhankelijkheid van andere landen die AI-technologie leveren

Als we afhankelijk worden van AI-technologie uit landen met andere belangen, ontstaan er ook dreigingen voor de nationale veiligheid. Op AI gebaseerde industriële en consumentengoederen en -diensten van andere landen vergroten de kwetsbaarheden voor statelijke inmenging, economische spionage en sabotage door die landen. Diverse staten zijn in staat om de gegevens te verkrijgen die verwerkt worden door (tech)bedrijven in hun land. Bijvoorbeeld in China zijn de bedrijven in het land onderworpen aan de richtlijnen van de Chinese inlichtingen- en veiligheidsdiensten en zijn deze in staat de bedrijven te dwingen tot medewerking.[81] Bedrijfsgevoelige informatie in Nederland kan op deze manier in handen komen van buitenlandse inlichtingendiensten wat gebruikt zal worden om de eigen concurrentie positie ten opzichte van landen als Nederland te vergroten. Daarnaast kan dit ook van invloed zijn op de sociale en politieke stabiliteit in Nederland. Buitenlandse inlichtingen- en veiligheidsdiensten kunnen op deze manier namelijk ook persoonsgegevens op grote schaal verwerven en dit gebruiken bij het onder druk zetten van buitenlandse diasporagemeenschappen. Daarnaast biedt de invloed van autocratische overheden over hun techbedrijven ook de mogelijkheid dat zij op AI-applicaties op sociale media censuur opleggen tegen hen onwelgevallige boodschappen over het regime.[82] In het meest vergaande geval zou het gebruik van AI-applicaties in vitale infrastructuren ook misbruikt kunnen worden door buitenlandse overheden om vitale processen te saboteren.

2.4 AI en ecologische veiligheid

Ecologische veiligheid omvat het ongestoord blijven voortbestaan van de natuurlijke leefomgeving in Nederland. Deze komt in het geding wanneer de natuur en het milieu langdurig worden aangetast.[83]

Er zijn vooralsnog geen kwaadwillende actoren die AI gebruiken om moedwillig schade aan de Nederlandse natuurlijke leefomgeving aan te brengen waarbij de ecologische veiligheid in het geding zou komen. Toch bestaat er wel een verband tussen de snelle ontwikkeling van AI

en de natuur en het milieu. AI kan onder andere worden gebruikt bij de verduurzaming van de samenleving. In de energietransitie wordt AI bijvoorbeeld gebruikt om energieverbruik te voorspellen.[84] De computers waarop generatieve AI-modellen draaien gebruiken echter wel grote hoeveelheden stroom en koelwater. De impact van dit energie- en watergebruik op het klimaat en het milieu is echter moeilijk te kwantificeren. Zowel de bedrijven die AI ontwikkelen als de bedrijven die de complexe computerchips maken delen niet graag hoeveel energie en water hun systemen gebruiken.[85]

2.5 AI en sociale en politieke stabiliteit

Sociale en politieke stabiliteit omvat het ongestoord voortbestaan van een maatschappelijk klimaat waarin individuen ongestoord kunnen functioneren en groepen mensen goed met elkaar kunnen samenleven binnen de verworvenheden van de Nederlandse democratische rechtstaat en de daarin gedeelde waarden.[86]

Het gebruik van nieuwe generatieve AI-toepassingen kan een dreiging vormen voor de sociale en politieke stabiliteit van Nederland. Deze toepassingen zijn de afgelopen jaren publiekelijk toegankelijk geworden en de kwaliteit van de teksten, afbeeldingen, audio en video die ze maken wordt steeds beter. Verschillende kwaadwillende actoren, waaronder staten, terroristen en criminelen, gebruiken AI-toepassingen om content te genereren, zoals desinformatie en extremistische propaganda. Daarnaast gebruiken bedrijven en overheden AI steeds vaker als onderdeel van hun dienstverlening, waarbij er kwetsbaarheden ontstaan voor foute en bevooroordeelde of zelfs discriminatoire uitkomsten. De potentiële impact hiervan op de sociale en politieke stabiliteit neemt toe naarmate AI steeds meer wordt geïntegreerd in publieke en private dienstverlening.

2.5.1 Invloed van AI nieuwsconsumptie

Nieuwsconsumptie via sociale media wordt per persoon bepaald door algoritmes en modellen. Factoren die hierin worden meegenomen zijn onder andere het zoekgedrag van de gebruiker, de geschiedenis, het sociale netwerk op het platform, populariteit van berichten, kwaliteit van de inhoud, maar ook het gedrag van de bezoeker op andere websites wanneer dit via cookies bijgehouden wordt. Kort gezegd schotelen sociale mediaplatforms op basis van al deze informatie de gebruiker content voor die past bij wat de gebruiker vaker ziet, opzoekt en waardeert.

Op deze manier spelen algoritmes een grote rol bij het creëren van zogenaamde ‘filterbubbels’. Dit fenomeen houdt in dat gebruikers op sociale media grotendeels content voorgeschreven krijgen die past binnen hun denkkader en wereldbeeld. Informatie van andere bronnen of andersdenkenden komt nauwelijks bij gebruikers terecht, omdat de algoritmes van de sociale mediaplatforms menen dat gebruikers dit soort informatie niet willen zien. Dit fenomeen draagt bij aan de verspreiding van desinformatie, heimelijke beïnvloeding en de verspreiding van extremistische propaganda.

2.5.2 Gebruik van AI bij het verspreiden en genereren van desinformatie

Kwaadwillende actoren kunnen met generatieve AI-toepassingen zoals ChatGPT en Gemini snel, eenvoudig, goedkoop en op grote schaal authentiek lijkende desinformatie en propaganda genereren.[87] Desinformatie is onware, inaccurate of misleidende informatie die met opzet wordt gemaakt en verspreid om geld te verdienen of om een persoon, sociale groep, organisatie of land te schaden.[88] Doordat generatieve AI-toepassingen in hoog tempo aanzienlijk beter en voor een groot publiek toegankelijk worden, neemt het risico toe dat AI-gegenereerde desinformatie onze perceptie van de realiteit zal gaan beïnvloeden.[89] Daartoe kunnen kwaadwillende actoren desinformatie in verschillende vormen creëren, waaronder tekst, audio en video (deepfakes).[90] Vanwege de grote schaal waarop kwaadwillenden met AI desinformatie kunnen verspreiden, kan het lastiger worden om dit tijdig en compleet te onderkennen.

In theorie zijn de meeste populaire LLM-toepassingen zo ontworpen dat ze geen racistische uitspraken of complottheorieën formuleren, maar de begrenzingen zijn te omzeilen (zie ook 2.1.2).[91] Zo bleek ChatGPT na het stellen van vragen met complottheorieën en desinformatie deze te verwerken in de teksten die het produceert.[92] Verschillende grote techbedrijven, zoals Google en Meta, stelden AI-programma's waarmee video's gegenereerd kunnen worden (nog) niet openbaar beschikbaar uit angst dat deze uiteindelijk gebruikt kunnen worden voor het snel en efficiënt verspreiden van desinformatie.[93]

Kwaadwillende actoren kunnen LLM's zo aanpassen dat ze desinformatie genereren op specifieke onderwerpen, zonder dat gebruikers daar expliciet om vragen. Om het bewustzijn rond veilige AI te vergroten hebben onderzoekers in juli 2023 laten zien hoe een bestaand open-source LLM zo aangepast kan worden dat deze op specifieke onderwerpen desinformatie verspreidt, terwijl het andere taken normaal blijft uitvoeren. In dit voorbeeld lieten de onderzoekers het model op de vraag wie de eerste man op de maan was het 'nepnieuws' genereren dat dit niet Neil Armstrong, maar Yuri Gagarin was. Hiermee wilden de onderzoekers aantonen dat bedrijven en gebruikers niet blind op getrainde AI-modellen moeten vertrouwen en dat LLM's door vervuiling eenvoudig en breed desinformatie en misinformatie kunnen verspreiden.[94]

Ook *deepfakes* zijn een vorm van desinformatie. Hierbij wordt met behulp van AI een oorspronkelijk beeldfragment bewerkt. De gebruiker kan hetgeen de persoon in het fragment doet of zegt aanpassen. In sommige gevallen zijn de gemanipuleerde beelden moeilijk van echt te onderscheiden. Het is de verwachting dat applicaties om *deepfakes* te produceren de komende jaren gebruiksvriendelijker en overtuigender worden.[95]

Overtuigende desinformatie en misinformatie kunnen leiden tot een significante afname van vertrouwen in berichtgeving, aangezien mensen niet meer zeker weten of een beeld, tekst, geluidsopname of video echt of nep is.[96] Dit kan op den duur leiden tot een breder verlies aan vertrouwen in de juistheid en authenticiteit van berichtgeving. Desinformatie kan ook de beeldvorming van mensen ten opzichte van zowel de overheid als specifieke groepen binnen de maatschappij negatief of valselijk positief beïnvloeden. Concreet betekent dit dat het vertrouwen

in de overheid en de omgang tussen verschillende groepen in onze samenleving onder druk kunnen komen te staan.[97]

2.5.3 Gebruik van AI voor ondermijnende beïnvloeding middels desinformatie

De verspreiding van desinformatie met behulp van AI kan doelgericht worden ingezet om andere landen te ondermijnen. Diverse staten gebruiken desinformatie om invloed uit te oefenen op hun diasporagemeenschappen of op sentimenten breder in de samenleving van andere landen. De afgelopen jaren zijn steeds meer handelingen in zulke beïnvloedingscampagnes geautomatiseerd, waaronder het aanmaken van kanalen en profielen, het analyseren van bestaande online content en het genereren van nieuwe content. Van Rusland is bekend dat het land hier actief op investeert. Met de komst van generatieve AI-modellen hebben landen zoals Rusland hiervoor geen leger van menselijke internettrollen meer nodig.[98] De AIVD, MIVD en Nationale Politie hebben in juli 2024 geholpen een Russische beïnvloedingscampagne te verstoren die gebruikmaakte van AI om nepprofielen op X te genereren.[99] Het ligt voor de hand dat landen als Rusland ook AI-toepassingen zullen gebruiken om verkiezingen in andere landen te beïnvloeden.

In de Verenigde Staten is al zichtbaar dat AI wordt ingezet om verschillende (oud-) presidents-kandidaten in diskrediet te brengen.[100] In januari 2024 werd getracht de voorverkiezingen te beïnvloeden met een telefonische boodschap waarin Joe Biden stemgerechtigde Democraten in de Amerikaanse staat New Hampshire opriep niét te gaan stemmen. Het betrof waarschijnlijk een deepfake van zijn stem.[101] Audiogeneratoren stellen inmiddels iedereen in staat om op basis van enkele seconden of minuten geluidsopname iemands stem te klonen.[102]

2.5.4 Gebruik van AI bij de verspreiding van extremistische propaganda

Desinformatie en misinformatie kunnen polarisatie aanwakkeren en, in individuele gevallen, de deur helpen openzetten naar radicalisering. Dit laatste geldt vooral voor mensen die in een 'filterbubbel' of 'echokamer' komen waar veel van deze foutieve informatie rondgaat.[103] Individuen die vatbaar zijn voor zulk gedachtegoed, zullen vanwege aanbevelingsalgoritmes op sociale mediaplatforms eerder in contact komen met zulke informatie. Kwaadwillenden kunnen hier misbruik van maken door zulke individuen nog meer te voeden met (des)informatie die hun gedachtengoed ondersteunt.[104]

AI biedt dan ook de mogelijkheid om met op maat gemaakte boodschappen op grotere schaal gespecificeerde extremistische propaganda te creëren, met minder kennis en vaardigheden dan voorheen nodig was.[105] Mondiale jihadistische organisaties en rechts-extremisten verspreiden verschillende handleidingen voor het gebruik van generatieve AI-modellen. Daarnaast is er AI gegenereerde terroristische en extremistische content in omloop, waaronder in de vorm van rechts-extremistische memes,[106] deepfakes waarin tekenfilmfiguren en populaire influencers islamitische teksten zingen[107] en video's over ISIS-operaties waarbij gebruik wordt gemaakt van een met AI gegenereerde nieuwslezer.[108]

Hoewel dit nog niet op grote schaal gebeurt zullen terroristische netwerken, net als een gemiddelde klantenservice, de mogelijkheid hebben om chatbots te gebruiken. Getraind met specifieke data voor het creëren van op maat gemaakte boodschappen, bieden chatbots bijvoorbeeld de mogelijkheid om sneller en met minder menselijk inspanning leden te rekruteren of aan te zetten tot radicalisering of mobilisatie.[109]

Door AI is waarschijnlijk het volume en het bereik van extremistische propaganda op sociale media toegenomen. Waarschijnlijk is er enig verband tussen de beschikbaarheid en aantrekkelijkheid van extremistische propaganda en de radicalisering van individuen. Het is echter twijfelachtig dat de toepassing van AI-applicaties daadwerkelijk leidt tot meer individuen die radicaliseren of aangezet worden tot het plegen van terroristisch geweld. Wat bij aanhangers van extremistisch gedachtegoed echt overkomt, overtuigt en activeert tot extremistische handelingen is waarschijnlijk erg persoonsafhankelijk en beperkt zich niet tot blootstelling aan extremistische propaganda.

2.5.5 Discriminatoire uitkomsten AI

Bij het gebruik van AI bestaat het risico op oneerlijke, bevooroordeelde of discriminatoire uitkomsten. Dit kan het gevolg zijn van een bewuste keuze van de ontwikkelaar of de gebruiker van een AI-toepassing, maar kan ook een onbedoelde uitkomst zijn van onzorgvuldig gebruik.[110] In de verwerking van gegevens door AI kan een zekere mate van bevooroordeling optreden, ook wel 'bias' genoemd. Zo'n vooroordeel is inherent aan AI, het bestaat uit verschillende denkregels die het model gebruikt om informatie te analyseren. Dit is vergelijkbaar met hoe mensen dat doen. In bepaalde gevallen kan een vooroordeel bij AI echter leiden tot discriminatoire uitkomsten. Dit kan bijvoorbeeld het geval zijn wanneer gebruik gemaakt wordt van trainingsdata die bevooroordeelde of slechts selectieve informatie bevat. Ook in een algoritme zelf kan bias zitten, zoals wanneer aan bepaalde indicatoren meer gewicht wordt gehangen die kunnen zorgen voor onbedoelde discriminatie.[111]

2.6 AI en de internationale rechtsorde en stabiliteit

Het veiligheidsbelang internationale rechtsorde en stabiliteit gaat over het goed functioneren van het internationale stelsel van normen en afspraken, gericht op het bevorderen van de internationale vrede en veiligheid.[112]

Er is in toenemende mate sprake van (geo)politisering van de (wereld)economie waarbij economische instrumenten als machtsmiddel worden ingezet. Er is eveneens sprake van een techrace tussen staten, met name de Verenigde Staten en China, om voorop te lopen in de ontwikkeling van opkomende en ontwrichtende sleuteltechnologieën, waaronder AI.[113] Hiervoor zetten landen verschillende instrumenten in, zoals het bepalen van standaarden, bedrijfsinvestering en overnames, export restricties en spionage. Zo heeft de techrace ten aanzien van AI ook invloed op internationale rechtsorde en daarmee dus ook de nationale veiligheid.[114]

Geopolitieke ontwikkelingen worden vergroot en versneld door toenemende digitalisering. Wereldwijd onderkennen landen het essentiële belang van digitalisering en technologisch leiderschap. Verschillende landen hebben de ambitie om nieuwe technologieën (als eerste) te ontwikkelen en standaarden te bepalen of te beïnvloeden. Andere landen proberen op zijn minst om de digitalisering en geopolitieke ontwikkelingen bij te kunnen benen. Ook streven landen de afgelopen jaren expliciet naar strategische autonomie door ofwel strategische afhankelijkheden te creëren om de eigen positie te versterken of juist eigen risicovolle afhankelijkheden af te bouwen. Ook de EU en Nederland willen strategisch autonoom zijn en technologisch leiderschap tonen. Dit is een vereiste om in het snel veranderende geopolitieke landschap relevant te kunnen blijven. Zo streeft ook Nederland naar technologisch leiderschap om toegang te houden tot technologie elders, maar ook om zich te kunnen verdedigen, spelregels internationaal af te dwingen en de koers van technologische ontwikkeling mede te bepalen.[115]

Mede vanwege de toenemende afhankelijkheid en de economische kansen die ontwikkeling van nieuwe technologieën zoals AI biedt, streven grootmachten ernaar om de overhand te krijgen in de (door)ontwikkeling van AI-modellen en het bepalen van standaarden van deze technologieën. Verschillende grootmachten zien de kennispositie en het gebruik van AI als een belangrijke race naar geopolitieke invloed. Het doel daarvan is om door middel van standaardbepaling en modelontwikkeling het grootste marktaandeel te verkrijgen en de geopolitieke positie te vergroten.[116]

Met name de Verenigde Staten en China willen voorloper en dominant zijn op het gebied van AI. AI heeft de potentie nieuwe technologie-reuzen voort te brengen. Daarnaast zorgt AI voor technologische innovaties, zowel in commerciële als in militaire toepassingen.[117] Op dit moment heeft de VS een voorsprong op het gebied van generatieve AI.[118] Het land heeft een sterke traditie van innovatie en een grote pool van talent en financiering.[119] Chinese bedrijven investeren al jaren in AI en lopen voor op het gebied van bewakingstechnologie, gezichtsherkenning en virtual reality. Hoewel Chinese techondernemers proberen de achterstand op het gebied van generatieve AI in te lopen, hebben Chinese bedrijven hier met een aantal belemmeringen te maken. In China moet door AI gegenereerde inhoud voldoen aan de strenge censuur van de Chinese overheid. Dit bemoeilijkt vrij wetenschappelijk onderzoek en innovatie. Chinese bedrijven hebben hierdoor minder geïnvesteerd in generatieve AI.[120] Volgens het Artificial Intelligence Index Report werd er in 2023 in de VS 67,2 miljard dollar in AI geïnvesteerd, dit is 8,7 keer zoveel als in China (7,8 miljard).[121]

Daarnaast remmen Amerikaanse export restricties de ontwikkelingen in China. Voor het trainen van grote generatieve AI-modellen zijn krachtige en geavanceerde chips nodig, zoals die van de Amerikaanse bedrijven Nvidia en AMD. Amerika heeft exportrestricties op de export van geavanceerde chips naar China opgelegd.[122] Hierdoor is het voor China lastiger om eigen systemen te maken voor het trainen van AI. Hoewel het land zelf ook over een grote datacenterinfrastructuur beschikt is deze minder geavanceerd dan die van cloud-bedrijven in de VS.[123]

China is de afgelopen jaren uitgegroeid tot technologische grootmacht en heeft zicht tot doel gesteld om in 2030 wereldleider te zijn op het gebied van AI. De AI ambities van China zijn nauw verweven met het terugkeren naar de (zelf beschouwde) rechtmatige plek op het wereldtoneel, het streven naar economische en militaire macht en het behoud van binnenlandse stabiliteit. Om wereldleider te worden heeft China kennis en technologie uit de VS en Europa nodig. Om toegang te verkrijgen tot die specifieke kennis en technologie zet het land economische instrumenten in, zoals investeringen en overnames van bedrijven die beschikken over specifieke kennis en technologie, maar ook door middel van (de financiering van) academische samenwerkingsverbanden waarin uitwisseling van kennis en technologie al dan niet gedwongen plaats kan vinden. Ook fysieke en digitale spionageactiviteiten worden voor dit doeleinde ingezet.[124]

China en Rusland vinden elkaar in hun beschouwing van de naoorlogse internationale rechtsorde als een westers construct en onder andere als een manier voor westerse landen om te interveniëren in binnenlandse aangelegenheden van andere landen. Beide landen tonen een structurele bereidheid om de afspraken binnen de internationale rechtsorde te schenden als dit hun eigen belangen beter bedient. De positie van deze twee landen is bijzonder in deze context gezien hun omvang, militaire en economische macht en hun positie als permanente leden van de VN-Veiligheidsraad. Beide landen streven naar invloed om de internationale rechtsorde, die onze kernwaarden en nationale belangen beschermt, naar eigen inzicht om te vormen naar een model dat indruist tegen deze bescherming en schadelijk is voor de internationale rechtsorde. In dit streven zoeken China en Rusland elk op hun eigen manier internationaal naar bondgenoten om hun invloed uit te breiden, bilateraal vaak bij landen en regimes die gevoelig zijn voor financieel economische en militair-technologische voordelen, multilateraal bij samenwerkingsorganisaties zoals BRICS en The Shanghai Cooperation Organisation (SCO). Technologische superioriteit op het gebied van AI draagt daar aan bij. Met name China kan welwillende landen aantrekkelijke financiële, economische, militaire en technologische mogelijkheden bieden – ook waar het gaat om AI-toepassingen – en deze landen op die manier aan zich binden.[125]

3. Conclusie

3.1 Zeer waarschijnlijk versterkt AI bestaande dreigingen voor de nationale veiligheid

AI is een veelomvattende systeemtechnologie die onze samenleving fundamenteel zal veranderen. AI heeft de potentie om ons leven op veel manieren makkelijker te maken en kan helpen bij het oplossen van tal van maatschappelijke problemen. Tegelijkertijd heeft de ontwikkeling van AI ook bedoeld en onbedoeld invloed op onze nationale veiligheid. Zeer waarschijnlijk versterkt de ontwikkeling van AI de diverse bestaande dreigingen voor onze nationale veiligheid. De mate waarin AI de dreiging voor de nationale veiligheid versterkt verschilt per toepassing en dreigingsfenomeen. Daarbij biedt AI ook kansen ter bescherming van de nationale veiligheid, zoals in de verdediging tegen cyberaanvallen en in militaire toepassingen.

Diverse staten investeren in het ontwikkelen en toepassen van AI in het militaire domein. Momenteel is de westerse militaire technologie nog superieur aan andere landen zoals China en Rusland. Deze technologische superioriteit draagt bij aan de huidige militaire machtsbalans en de daarmee de territoriale veiligheid van Nederland en haar bondgenoten en de internationale rechtsorde. De ontwikkeling van AI heeft de potentie deze machtsverhoudingen te verschuiven als andere landen beter in staat zijn AI in te zetten in het militaire domein. Met name China zet daar op in. Rusland heeft vergelijkbare ambities maar het is momenteel twijfelachtig of Rusland haar achterstand op AI kan inlopen. De ontwikkeling van AI vergroot ook de cybercapaciteiten van diverse actoren, zowel in het militaire domein als daarbuiten. Hoewel de invloed van AI op cyberaanvallen niet duidelijk zichtbaar aanwezig is, kan AI wel een faciliterende rol hebben bij bestaande cyberaanvallen.

De belangen van staten om voorloper te zijn in de ontwikkeling en toepassing van AI zijn groot, in het bijzonder voor de grootmachten. Naast de beschreven militaire en digitale toepassingen, biedt de ontwikkeling bovenal een aanzienlijke economische potentie. Gezien het grote belang van technologisch leiderschap op AI zullen onder meer de grootmachten geen mogelijkheid onbenut laten om kennis op te bouwen en anderen die te ontzeggen. Het gaat daarbij om de inzet van economische instrumenten zoals investeringen, overnames en fusies. En ook om (digitale) spionage, sabotage en diefstal van AI-kennis en -technologieën. Zowel de inzet van reguliere economische instrumenten als de heimelijke handelswijze kan de nationale veiligheid

raken als het verdienvermogen van Nederland wordt aangetast, als Nederland hierdoor voor onze vitale processen afhankelijker wordt van andere landen, als persoonsgegevens van Nederlanders via buitenlandse bedrijven op grote schaal inzichtelijk worden voor andere staten, en/of als de handelswijze (zoals spionage) onze democratische rechtsstaat ondermijnt.

Naast de hoogwaardige technologische ontwikkelingen rondom AI zijn ook veel toepassingen van AI steeds toegankelijker geworden. Zonder hoogwaardige technologische expertise kan generatieve AI in theorie door iedereen gebruikt worden voor het maken van malware, phisingmails, desinformatie en propaganda. Dit verhoogt in potentie ook de dreiging van diverse niet-statelijke actoren zoals terroristische groeperingen. Zeer waarschijnlijk maken AQ en ISIS gebruik van generatieve AI om propaganda te vervaardigen en dit zal waarschijnlijk het volume en bereik ervan verder vergroten. Hoewel deze ontwikkeling problematisch is, is het vooralsnog onzeker in hoeverre dit daadwerkelijk zal leiden tot meer radicalisering en geweld.

Ten slotte heeft onzorgvuldig gebruik van AI ook de potentie om de nationale veiligheid onbedoeld te raken. Naarmate private en (semi-)publieke organisaties AI steeds meer toepassen in hun bedrijfsprocessen, zal dit meer invloed krijgen op de wijze waarop deze organisaties omgaan met burgers. Onzorgvuldigheden in de ontwikkeling en toepassing van modellen, in combinatie met de steeds moeilijker te doorgronden werking van die modellen, kunnen risico's met zich meenemen voor organisaties en burgers. Het is de vraag wat een grote afhankelijkheid van een technologie als AI door de hele samenleving heen zal betekenen, bijvoorbeeld wanneer Nederland hiervoor afhankelijk is van andere landen, maar ook wanneer een AI-systeem niet naar behoren functioneert. Dit raakt de sociale en politieke stabiliteit en potentieel de fysieke veiligheid van Nederland, uiteenlopend van discriminerende beslissingen tot het uitvallen van vitale processen.

3.2 Vooruitblik

AI en de bijbehorende toepassingen zullen zich razendsnel blijven ontwikkelen. Generatieve AI kan gebruikt worden bij het ontwikkelen van nieuwe AI-systemen en toepassingen. Dit maakt een exponentiële toename van AI mogelijk. Twee ontwikkelingen die samenhangen met de techrace gevoerd door bedrijven en staten zijn van invloed op de mate waarin AI bijdraagt aan de dreigingen voor de nationale veiligheid. Ten eerste zullen AI-toepassingen die nu veel hoogwaardige expertise en veel (financiële) middelen vereisen, op termijn waarschijnlijk breder en laagdrempelig beschikbaar worden. Hierdoor zal het aantal statelijke en niet-statelijke actoren groeien dat misbruik kan maken van dergelijke AI-toepassingen. Ten tweede, zullen AI-toepassingen waarschijnlijk steeds meer onderdeel worden van onze economie en maatschappij. Dit geldt vooral voor het digitale domein, maar door ontwikkelingen op het gebied van robotica en AI in toenemende mate ook in het fysieke domein. Dit zal enerzijds economische en maatschappelijke voordelen bieden maar anderzijds ook nieuwe kwetsbaarheden vormen die kwaadwillenden zullen proberen te misbruiken.

Vooralsnog vormen de ontwikkelingen van autonoom opererende AI die zelfstandig controle over mensen kan krijgen geen dreiging voor de nationale veiligheid. Er zijn experts die vrezen dat AI zelf de controle van mensen kan gaan overnemen als mensen steeds meer belangrijke beslissingen aan AI-systemen overlaten of als AI-systemen zelf hun eigen invloed willen uitbreiden. Het beheersen van AI-systemen zou moeilijker worden als ze autonoom kunnen overleven en zichzelf kunnen kopiëren en aanpassen. Geen van de huidige AI-systemen is hiertoe in staat, maar geheel uitgesloten is het niet. Zo laat recent onderzoek zien dat nieuwe geavanceerde modellen wel hiervoor relevante taken kunnen uitvoeren. Het idee dat AI controle over mensen krijgt blijft echter omstreden en vooralsnog is de kans dat dit op enig moment van invloed kan zijn op de nationale veiligheid zeer gering. De invloed van AI op de nationale veiligheid zal de komende jaren waarschijnlijk vooral bepaald worden door kwaadwillenden die doelbewust AI-toepassingen gebruiken om hun eigen doelen te verwezenlijken ten koste van de Nederlandse nationale veiligheidsbelangen.

Bijlage 1: Overzicht zes nationale veiligheidsbelangen en impactcriteria

In deze analyse worden de (mogelijke) veiligheidsdreigingen beschreven aan de hand van de zes nationale veiligheidsbelangen. Deze veiligheids-belangen en hun impactcriteria zijn uitgebreid beschreven in de "Leidraad Risicobeoordeling Rijksbrede Risicoanalyse Nationale Veiligheid" van het Analistennetwerk Nationale Veiligheid. Binnen de methodiek nationale veiligheid worden deze belangen en criteria gebruikt om te bepalen of een gebeurtenis de nationale veiligheidsbelangen raakt en of de nationale veiligheid in het geding is. In deze analyse wordt niet vastgesteld in welke mate AI een dreiging voor de nationale veiligheid is. Wel wordt de ordening van de zes veiligheidsbelangen gebruikt om te beschrijven hoe AI een dreiging voor de nationale veiligheid kan vormen. Onderstaande tabel geeft een kort overzicht van alle belangen en criteria.

Belang	Impactcriteria
1. Territoriale veiligheid	1.1 Aantasting van de integriteit van het grondgebied van het Koninkrijk der Nederlanden
	1.2 Aantasting van de integriteit van de internationale positie van het Koninkrijk der Nederlanden
	1.3 Aantasting van de integriteit van de digitale ruimte
	1.4 Aantasting van de integriteit van het bondgenootschappelijke grondgebied
2. Fysieke veiligheid	2.1 Doden
	2.2 Ernstig gewonden en chronisch zieken
	2.3 Gebrek aan primaire levensbehoeften
3. Economische veiligheid	3.1 Kosten
	3.2 Aantasting van de vitaliteit van de economie van het Koninkrijk der Nederlanden
4. Ecologische veiligheid	4.1 Langdurige aantasting van het milieu en de natuur
5. Sociale en politieke stabiliteit	5.1 Verstoring van het dagelijkse leven
	5.2 Aantasting van de democratische rechtstaat
	5.3 Sociaal-maatschappelijke impact
6. Internationale rechtsorde en stabiliteit	6.1 Aantasting van de normen van staatssoevereiniteit, vreedzame co-existentie en vreedzame geschillenbeslechting
	6.2 Aantasting van de werking, legitimiteit dan wel nalevening van de internationale verdragen en normen inzake de rechten van de mens
	6.3 Aantasting van een op regels gebaseerd internationaal financieel-economisch bestel
	6.4 Aantasting van de effectiviteit, legitimiteit van de multilaterale instituties
	6.5 Instabiliteit van staten grenzend aan het Koninkrijk der Nederlanden en in de directe omgeving van de Europese Unie

Bijlage 2: Begrippenlijst

Algoritme:	Een gespecificeerde instructie om een probleem op te lossen of een berekening uit te voeren.
Artificial general intelligence (AGI):	Artificiële Intelligentie die menselijke intelligentie op alle gebieden kan evenaren.
Artificial super intelligence (ASI):	Artificiële Intelligentie die op alle gebieden superieur is aan menselijke intelligentie.
Datagestuurde AI:	AI die gebruik maakt van grote hoeveelheden data en waarbij het leerproces van algoritmes is geautomatiseerd.
Deepfake:	Een door AI gegenereerd beeld of geluidsfragment met de suggestie dat het gaat om ongemanipuleerde content.
Deep learning:	Een vorm van machine learning die gebaseerd is op de werking van neuronen in het menselijk brein. Deep learning maakt gebruik van meerlaagse netwerken, vandaar de term 'deep'.
Full AI:	Zie artificial general intelligence.
Generatieve AI:	Een vorm van AI die op basis van door de gebruiker gegeven opdrachten of vragen nieuwe content kan creëren uit bestaande data.
Internet of Things:	Netwerken van slimme apparaten die via internet met elkaar communiceren.
Kunstmatige intelligentie (AI):	Systemen die intelligent gedrag vertonen door hun omgeving te analyseren en – met een graad van autonomie – actie te ondernemen om specifieke doelen te bereiken. In het Engels aangeduid als artificial intelligence.

Large language model (LLM):	Een type AI dat is ontworpen om menselijke taal te interpreteren en te genereren. LLM's zijn getraind op grote hoeveelheden tekstgegevens, waardoor ze patronen kunnen herkennen in de manier waarop mensen taal gebruiken.
Machine learning (ML):	Zie Datagestuurde AI.
Malware:	Kwaadaardige software die schadelijk is voor computers. Het wordt gebruikt om computersystemen te verstoren, gevoelige informatie te verzamelen of toegang te krijgen tot computersystemen.
Narrow AI:	AI die is gericht op een specifieke vaardigheid zoals beeld- of spraakherkenning, vaak binnen een specifieke context..
Phishing-mails:	Nepberichten die lijken te komen van een betrouwbare bron en proberen persoonlijke informatie of geld van mensen te stelen.
Prompts:	Aanwijzingen of vragen die worden gebruikt om een AI-model te activeren en tekst, afbeeldingen, audio of video te genereren.
Reinforcement learning:	Een vorm van machine learning waarbij het algoritme wordt getraind om bepaalde strategieën te volgen via een systeem van positieve en negatieve feedback.
Supervised learning:	Een vorm van machine learning waarbij een systeem wordt gevoed met door mensen gelabelde data.
Strong AI:	Zie Artificial General Intelligence.
Symbolische AI:	AI gebaseerd op precieze door mensen gecreëerde algoritmes die een computer stap voor stap volgt om te bepalen hoe deze op een bepaalde situatie reageert.
Unsupervised learning:	Een vorm van machine learning waarbij het programma wordt gevoed met ongelabelde data, waaruit het algoritme zelf patronen moet destilleren.
Weak AI:	Zie Narrow AI.

Bijlage 3: Bronnen en referenties

1 '*Zeven acties voor verantwoord innoveren met AI*', Rathenau Instituut, 2020.
2 '*Artificial intelligence: How does it work, why does it matter, and what can we do about it?*', European Parliamentary Research Service, 2020.
3 '*Artificial intelligence: How does it work, why does it matter, and what can we do about it?*', European Parliamentary Research Service, 2020.
4 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
5 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
6 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
7 Volgens het Artificial Intelligence Index Report 2023 waren private investeringen in AI in 2022 18 keer hoger dan in 2013. Toch was er in 2022 voor het eerst een daling in de private investeringen in AI te zien. In 2022 werd er wereldwijd $91,9 miljard aan private investeringen in AI gedaan, dat is 26,7% minder dan in 2021.
8 '*Kabinetsreactie WRR-rapport Opgave AI. De nieuwe systeemtechnologie*', 2022.
9 '*The AI Index 2023 Annual Report*', AI Index Steering Committee, Institute for Human-Centered AI. Stanford University, april 2023.
10 '*The AI Index 2023 Annual Report*', AI Index Steering Committee, Institute for Human-Centered AI. Stanford University, april 2023.
11 https://www.wrr.nl/publicaties/geluidsfragmenten/2021/11/11/15-opgave-ai.-de-nieuwe-systeemtechnologie '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
12 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
13 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
14 '*Artificial intelligence: How does it work, why does it matter, and what can we do about it?*', European Parliamentary Research Service, 2020.
15 '*Artificial intelligence: How does it work, why does it matter, and what can we do about it?*', European Parliamentary Research Service, 2020.
16 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
17 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021
18 '*Artificial intelligence: How does it work, why does it matter, and what can we do about it?*', European Parliamentary Research Service, 2020.

19 '*Artificial intelligence: How does it work, why does it matter, and what can we do about it?*', European Parliamentary Research Service, 2020.
20 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021. '*What is artificial general intelligence really about? Conquering the last leg of the AI arms race*', ZDNET, 27-09-2023, https://www.zdnet.com/article/what-is-artificial-general-intelligence-really-about-conquering-the-last-leg-of-the-ai-arms-race/.
'*What's AGI, and Why Are AI Experts Skeptical?*', WIRED, 20-04-2023, https://www.wired.com/story/what-is-artificial-general-intelligence-agi-explained/.
21 '*ChatGPT and large language models: what's the risk?*', National Cyber Security Centre, 10-03-2023, https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk.
22 '*In A.I. Race, Microsoft and Google choose speed over caution.*', International New York Times, 10-04-2023, https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html.
23 '*Nederland ontwikkelt antwoord op ChatGPT: AI-taalmodel GPT-NL*', de Volkskrant, 2 november 2023.
24 '*China's version of ChatGPT has finally been made public. But will censorship limit its power?*', ABC News, 05-09-2023, https://www.abc.net.au/news/2023-09-06/china-artificial-intelligence-chatbot-chatgpt-ernie-baidu/102803758.
25 '*Russia now has its own ChatGPT, introducing GigaChat*', Cybernews, 05-05-2023, https://cybernews.com/tech/russia-chatgpt-competitor-gigachat/.
'*Saudi-China collaboration raises concerns about access to AI chips*', The Financial Times, 11-10-2023.
26 *'Google and Microsoft's AI arms race could have 'unintended consequences,'* an AI ethicist warns', CNN Business, 07-02-2023, https://edition.cnn.com/2023/02/06/media/google-microsoft-ai-reliable-sources.
27 '*The AI Index 2023 Annual Report*', AI Index Steering Committee, Institute for Human-Centered AI. Stanford University, april 2023.
28 '*How to write better ChatGPT prompts for the best generative AI results*', ZDNET, 25-09-2023, https://www.zdnet.com/article/how-to-write-better-chatgpt-prompts/.
29 '*AI-rommel verspreidt zich over alle uithoeken van het internet: drie gevaren*', Volkskrant, 08-09-2023.
30 '*ChatGPT and large language models: what's the risk?*', National Cyber Security Centre, 10-03-2023, https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk.
31 '*ChatGPT-bug lekte gesprek-geschiedenis aan andere gebruikers*', RTLnieuws, 23-03-2023, https://www.rtlnieuws.nl/tech/artikel/5373472/chatgpt-openai-sam-altman-bug-problemen.
32 '*Apple verbiedt intern gebruik van AI-tools als ChatGPT uit angst voor lekken*', Tweakers, 19-05-2023, https://tweakers.net/nieuws/209916/apple-verbiedt-intern-gebruik-van-ai-tools-als-chatgpt-uit-angst-voor-lekken.html.
33 '*AP vraagt om opheldering over ChatGPT*', Autoriteit Persoonsgegevens, 07-06-2023, https://autoriteitpersoonsgegevens.nl/actueel/ap-vraagt-om-opheldering-over-chatgpt.
34 '*ChatGPT: zegen en vloek*', IP Vakblad voor informatieprofessionals, april 2023.
35 In de Veiligheidsstrategie voor het Koninkrijk der Nederlanden 2023-2029 wordt de volgende definitie van nationale veiligheid gehanteerd: de bescherming van onze nationale veiligheidsbelangen tegen dreigingen die deze belangen kunnen schaden en daarmee maatschappelijke ontwrichting kunnen veroorzaken.
36 '*Leidraad Risicobeoordeling Rijksbrede Risicoanalyse Nationale Veiligheid*', Analistennetwerk Nationale Veiligheid, 2019.

37 '*National Security Commission on Artificial Intelligence, Final report*', National Security Commission on Artificial Intelligence, 2021.
38 '*In Oekraïne zie je traditionele oorlogsvoering, maar het land is tegelijk een proefterrein voor de nieuwste oorlogstechnologie van de VS*', De Morgen, 26 jul 2023.
39 '*National Security Commission on Artificial Intelligence, Final report*', National Security Commission on Artificial Intelligence, 2021.
40 '*Militaire AI is veel breder dan een autonoom wapen, zegt Roy Lindelauf. Ook nepnieuws valt eronder*', NRC, 18-01-2024.
41 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
42 '*In Oekraïne zie je traditionele oorlogsvoering, maar het land is tegelijk een proefterrein voor de nieuwste oorlogstechnologie van de VS*', De Morgen, 26 jul 2023.
43 '*'The Gospel': how Israel uses AI to select bombing targets in Gaza*', The Guardian, 01-12-2023, https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets?ref=upstract.com.
44 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
45 '*Mistgordijnen nemen toe in cyberdomein*', Ministerie van Defensie/MIVD, 23-06-2022, https://magazines.defensie.nl/specials/2022/01/mistgordijnen-nemen-toe-in-cyberdomein.
46 '*Dreigingsbeeld Statelijke Actoren 2*', AIVD, MIVD en NCTV, 2022.
47 '*National Security Commission on Artificial Intelligence, Final report*', National Security Commission on Artificial Intelligence, 2021.
48 '*Mistgordijnen nemen toe in cyberdomein*', Ministerie van Defensie/MIVD, 23-06-2022, https://magazines.defensie.nl/specials/2022/01/mistgordijnen-nemen-toe-in-cyberdomein.
49 '*Dreigingsbeeld Statelijke Actoren 2*', AIVD, MIVD en NCTV, 2022.
50 '*Mistgordijnen nemen toe in cyberdomein*', Ministerie van Defensie/MIVD, 23-06-2022, https://magazines.defensie.nl/specials/2022/01/mistgordijnen-nemen-toe-in-cyberdomein.
51 '*Dreigingsbeeld Statelijke Actoren 2*', AIVD, MIVD en NCTV, 2022.
52 *Putin Wants Russia to Win the Artificial Intelligence Race. Here's Why it Won't*, The Moscow Times, Ben Dubow, Nov. 14, 2023
Will Russia Rule the World Through AI? Assessing Putin's Rhetoric Against Russia's Reality, Pages 36-60, The RUSI Jounal vol 164, Published online: 29 Nov 2019
The Impact of AI on Strategic Stability is What States Make of It: Comparing US and Russian Discourses, Nadibaidze & Miotto, 2023
53 '*Defensie Strategie Data Science en AI 2023-2027*', ministerie van Defensie, 2023.
54 '*OPWNAI: Cybercriminals starting to use ChatGPT*', Check Point Research, 06-01-2023, https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/.
55 '*Beyond the Safeguards: Exploring the Security Risks of ChatGPT*', Erik Derner and Kristina Batistic, 13-05-2023, https://arxiv.org/abs/2305.08005.
'*Themarapportage cyberdreigingen*', Analistennetwerk Nationale Veiligheid, 2022.
56 '*Beyond the Safeguards: Exploring the Security Risks of ChatGPT*', Erik Derner and Kristina Batistic, 13-05-2023, https://arxiv.org/abs/2305.08005.
'*ChatGPT and large language models: what's the risk?*', National Cyber Security Centre, 10-03-2023, https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk.

57 '*How threat actors are using AI and other modern tools to enhance their phishing attempts*', Talos, 13-04-2023, https://blog.talosintelligence.com/ai-and-other-modern-tools-enhance-phishing/
'*ChatGPT - The impact of Large Language Models on Law Enforcement*', Europol, 2023.
58 '*Cybersecuritybeeld Nederland 2023*', NCTV, 2023.
59 '*Cybersecuritybeeld Nederland 2023*', NCTV, 2023.
60 '*ChatGPT and large language models: what's the risk?*', National Cyber Security Centre, 10-03-2023, https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk.
61 '*Themarapportage cyberdreigingen*', Analistennetwerk Nationale Veiligheid, 2022.
62 Wanneer je bijvoorbeeld aan ChatGPT vraagt hoe je een bom kan maken, of hij een mail kan schrijven waarmee je iemand bedreigt of code kan schrijven waarmee je kan hacken zal de chatbot aangeven dat hij daar niet bij kan helpen.
63 '*ChatGPT - The impact of Large Language Models on Law Enforcement*', Europol, 2023.
'*WormGPT: New AI Tool Allows Cybercriminals to Launch Sophisticated Cyber Attacks*', The Hackers News, 15-07-2023, https://thehackernews.com/2023/07/wormgpt-new-ai-tool-allows.html.
64 '*WormGPT – The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks*', Slashnext, 13-07-2023, https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/
'*WormGPT: New AI Tool Allows Cybercriminals to Launch Sophisticated Cyber Attacks*', The Hackers News, 15-07-2023, https://thehackernews.com/2023/07/wormgpt-new-ai-tool-allows.html.
'*WormGPT Is a ChatGPT Alternative With 'No Ethical Boundaries or Limitations''*, PC, 14-07-2023, https://www.pcmag.com/news/wormgpt-is-a-chatgpt-alternative-with-no-ethical-boundaries-or-limitations
'*New AI Tool 'FraudGPT' Emerges, Tailored for Sophisticated Attacks*', The Hackers News, 26-07-2023, https://thehackernews.com/2023/07/new-ai-tool-fraudgpt-emerges-tailored.html.
65 '*ChatGPT - The impact of Large Language Models on Law Enforcement*', Europol, 2023.
66 '*LLM meets malware: starting the era of autonomous threat, Security Affairs*', 14-06-2023, https://securityaffairs.com/147447/malware/llm-meets-malware.html.
67 '*Human-competitive AI will disrupt the cyber security industry; prepare now!*', Partnership for Cyber Security Innovation, september 2023.
68 '*Human-competitive AI will disrupt the cyber security industry; prepare now!*', Partnership for Cyber Security Innovation, september 2023.
69 '*SentinelOne sticks generative AI into its stuff because 2023 gotta 2023*', The Register, 24-04-2023, https://www.theregister.com/2023/04/24/rsa_sentinelone_ai_threat_hunting/
'*Google brings generative AI to cybersecurity*', TechCrunch, 24-04-2023, https://techcrunch.com/2023/04/24/google-brings-generative-ai-to-cybersecurity/?guccounter=1
70 '*Human-competitive AI will disrupt the cyber security industry; prepare now!*', Partnership for Cyber Security Innovation, september 2023.
71 '*Leidraad Risicobeoordeling Rijksbrede Risicoanalyse Nationale Veiligheid*', Analistennetwerk Nationale Veiligheid, 2019.
72 '*These experts are racing to protect AI from Hackers. Time is running out*', ZDNET, 2023, https://www.zdnet.com/in-depth/innovation/these-experts-are-racing-to-protect-ai-from-hackers-time-is-running-out/#ftag=RSSbaffb68

73 '*OpenAI: geringe kans dat ChatGPT recept geeft voor biologische wapens*', RTLnieuws, 01-02-2024, https://www.rtlnieuws.nl/economie/artikel/5432595/kleine-kans-dat-openai-recept-geeft-voor-biologische-wapens
'*Capabilities and risks from frontier AI: A discussion paper on the need for further research into AI risk*', AI Safety Summit, oktober 2023.
74 '*The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study.'*,Mouton, Christopher A., Caleb Lucas, and Ella Guest (RAND Corporation), 2024. https://www.rand.org/pubs/research_reports/RRA2977-2.html.
75 *Simulating 500 million years of evolution with a language model*, Hayes, T, et al. 2024
76 '*Capabilities and risks from frontier AI: A discussion paper on the need for further research into AI risk*', AI Safety Summit, oktober 2023.
77 '*Leidraad Risicobeoordeling Rijksbrede Risicoanalyse Nationale Veiligheid*', Analistennetwerk Nationale Veiligheid, 2019.
78 '*Zeven acties voor verantwoord innoveren met AI*', Rathenau Instituut, 2020.
79 '*National Security Commission on Artificial Intelligence, Final report*', National Security Commission on Artificial Intelligence, 2021.
80 '*Zeven acties voor verantwoord innoveren met AI*', Rathenau Instituut, 2020.
81 '*De-risking authoritarian AI. A balanced approach to protecting our digital ecosystems.'*, Australian Strategic Policy Institute, juli 2023.
82 '*De-risking authoritarian AI. A balanced approach to protecting our digital ecosystems.'*, Australian Strategic Policy Institute, juli 2023.
83 '*Leidraad Risicobeoordeling Rijksbrede Risicoanalyse Nationale Veiligheid*', Analistennetwerk Nationale Veiligheid, 2019.
84 TU Delft, ABN/NLAIC Position Paper AI als versneller van de energietransitie
85 '*Ghost in the machine – addressing the consumer harms of generative AI*', Norwegian Consumer Council, juni 2023.
'*TechScape: Turns out there's another problem with AI – its environmental toll*', The Guardian, 01-08-2023, https://www.theguardian.com/technology/2023/aug/01/techscape-environment-cost-ai-artificial-intelligence.
'*Kunstmatige intelligentie vreet stroom, één opdracht hetzelfde als een uur een lamp aan*', NOS, 31-05-2023, https://nos.nl/nieuwsuur/artikel/2477186-kunstmatige-intelligentie-vreet-stroom-een-opdracht-hetzelfde-als-een-uur-een-lamp-aan.
'*AI slurpt energie: 'Kan over vier jaar net zoveel stroom als Nederland gebruiken''*, NOS, 10-10-2023, https://nos.nl/artikel/2493605-ai-slurpt-energie-kan-over-vier-jaar-net-zoveel-stroom-als-nederland-gebruiken.
86 '*Leidraad Risicobeoordeling Rijksbrede Risicoanalyse Nationale Veiligheid*', Analistennetwerk Nationale Veiligheid, 2019.
87 '*ChatGPT - The impact of Large Language Models on Law Enforcement*', Europol, 2023.
'*Disinformation Researchers Raise Alarms About A.I. Chatbots*', The New York Times, 08-02-2023.
88 https://www.rijksoverheid.nl/onderwerpen/desinformatie-nepnieuws
89 '*Elections spotlight generative ai and deep fakes*', Check Point Research, 05-09-2023, https://research.checkpoint.com/2023/elections-spotlight-generative-ai-and-deep-fakes/.

90 '*Elections spotlight generative ai and deep fakes*', Check Point Research, 05-09-2023, https://research.checkpoint.com/2023/elections-spotlight-generative-ai-and-deep-fakes/.
91 '*People are 'Jailbreaking' ChatGPT to Make It Endorse Racism, Conspiracies*', Vice, 08-02-2023, https://www.vice.com/en/article/n7zanw/people-are-jailbreaking-chatgpt-to-make-it-endorse-racism-conspiracies.
92 '*Disinformation Researchers Raise Alarms About A.I. Chatbots*', The New York Times, 08-02-2023.
93 '*Instant Videos Could Represent the Next Leap in A.I. Technology*', International New York Times, 05 apr 2023.
94 '*PoisonGPT: How we hid a lobotomized LLM on Hugging Face to spread fake news*', Mithril Security, 09-07-2023, https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/
95 '*Themarapportage polarisatie, extremisme en terrorisme*', Analistennetwerk Nationale Veiligheid, 2022.
96 '*Ghost in the machine – addressing the consumer harms of generative AI*', Norwegian Consumer Council, juni 2023.
97 '*Themarapportage polarisatie, extremisme en terrorisme*', Analistennetwerk Nationale Veiligheid, 2022.
98 '*Could AI swamp social media with fake accounts?*', BBC, 14-02-2023, https://www.bbc.com/news/business-64464140?at_medium=RSS&at_campaign=KARANGA
'*US cyber chiefs warn AI will help crooks, China develop nastier cyberattacks faster*', The Register, 12-04-2023, https://www.theregister.com/2023/04/12/us_chatgpt_threat/.
99 '*Nederland en VS verstoren Russische digitale beïnvloedingsoperatie*', AIVD, 09-07-2024, https://www.aivd.nl/actueel/nieuws/2024/07/09/nederland-en-vs-verstoren-russische-digitale-beinvloedingsoperatie
100 '*Risk in focus: Generative AI and the 2024 election cycle*', CISA, 18-01-2024.
101 '*AI-versie van Joe Biden roept Democraten op niet te gaan stemmen*', De Standaard, 23-01-2024.
102 '*Een digitale kopie van je eigen stem*', Volkskrant, 08-01-2023.
103 '*Themarapportage polarisatie, extremisme en terrorisme*', Analistennetwerk Nationale Veiligheid, 2022.
104 Dreigingsbeeld Terrorisme Nederland, NCTV, Juni 2024.
105 Dreigingsbeeld Terrorisme Nederland, NCTV, Juni 2024.
106 '*Early terrorist experimentation with generative artificial intelligence services*', Tech Against Terrorism, November 2023.
107 '*AI Jihad: Deciphering Hamas, Al-Qaeda and Islamic State's Generative AI Digital Arsenal*', Daniel Siegel, 19-02-2024, https://gnet-research.org/2024/02/19/ai-jihad-deciphering-hamas-al-qaeda-and-islamic-states-generative-ai-digital-arsenal/
108 '*These ISIS news anchors are AI fakes. Their propaganda is real.*', The Washington Post, 17 mei 2024.
109 '*The Next Paradigm-Shattering Threat? Right-Sizing the Potential Impacts of Generative AI on Terrorism*', David Wells (Middle East Institute), maart 2024.
110 '*Toezicht op AI & Algoritmes*', Autoriteit Persoonsgegevens, https://autoriteitpersoonsgegevens.nl/uploads/imported/toezicht_op_ai_en_algoritmes.pdf
111 '*Toezicht op AI & Algoritmes*', Autoriteit Persoonsgegevens, https://autoriteitpersoonsgegevens.nl/uploads/imported/toezicht_op_ai_en_algoritmes.pdf

112 '*Leidraad Risicobeoordeling Rijksbrede Risicoanalyse Nationale Veiligheid*', Analistennetwerk Nationale Veiligheid, 2019.
113 '*Dreigingsbeeld Statelijke Actoren 2*', AIVD, MIVD en NCTV, 2022.
114 '*Opgave AI. De nieuwe systeemtechnologie*', WRR, 2021.
115 BNC-fiche Mededeling Europese Normalisatiestrategie, Kamerbrief inzake open strategische autonomie
116 '*Who is winning the AI race: China, the EU or the United States?*', Center for Data Innovation, januari 2021, '*Chinese techmiljardairs én overheid slijpen de messen voor AI-strijd met VS*', Financieel Dagblad, 14 jul 2023.
117 '*Chinese techmiljardairs én overheid slijpen de messen voor AI-strijd met VS*', Financieel Dagblad, 14-07-2023.
118 '*ChatGPT: hoe Europa in een halfjaar de slag om de chatbot verloor*', Financieel dagblad, 14-04-2023.
119 '*Who is winning the AI race: China, the EU or the United States?*', Center for Data Innovation, januari 2021.
120 '*Chinese techmiljardairs én overheid slijpen de messen voor AI-strijd met VS*', Financieel Dagblad, 14 jul 2023.
'*Will China overtake the U.S. on AI? Probably not. Here's why.*', The Washington Post, 6 July 2023.
121 '*The AI Index 2024 Annual Report*', AI Index Steering Committee, Institute for Human-Centered AI. Stanford University, 2024.
122 '*Chinese techmiljardairs én overheid slijpen de messen voor AI-strijd met VS*', Financieel Dagblad, 14 jul 2023.
123 '*VS willen China ook afsnijden van AI-clouddiensten*', AG Connect, 06-07-2023, https://www.agconnect.nl/business/artificial-intelligence/vs-willen-china-ook-afsnijden-van-ai-clouddiensten.
124 '*Will China overtake the U.S. on AI? Probably not. Here's why.*', The Washington Post, 6 July 2023, '*The AI Index 2023 Annual Report*', AI Index Steering Committee, Institute for Human-Centered AI. Stanford University, april 2023, DBSA2
125 AIVD jaarverslag 2023

December 2024

Deze publicatie is een gezamenlijke uitgave van:

Algemene Inlichtingen- en
Veiligheidsdienst (AIVD)

Militaire Inlichtingen- en
Veiligheidsdienst (MIVD)

Nationaal Coördinator
Terrorismebestrijding en Veiligheid
(NCTV)