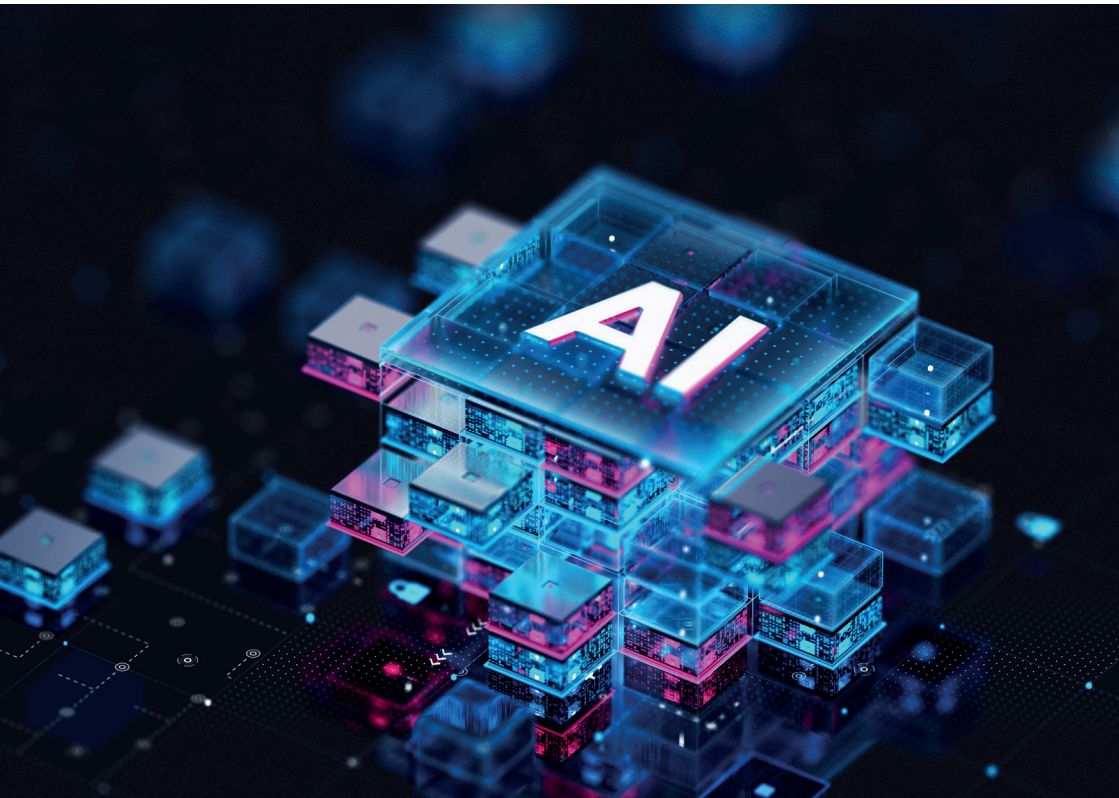
Generatieve AI:
een transformatieve impact op cybersecurity



Generatieve AI vraagt om een nieuwe cybersecurityaanpak

De Algemene Inlichtingen- en Veiligheidsdienst (AIVD) en de Rijksinspectie Digitale Infrastructuur (RDI) zien dat AI een transformatieve impact zal hebben op het gebied van cybersecurity. Deze impact is dusdanig anders van aard dat een nieuwe cybersecuritybenadering nodig is om deze impact het hoofd te bieden.

In deze publicatie lichten de AIVD en de RDI toe welke kenmerken van generatieve AI zorgen voor deze transformatieve impact. Daarbij wordt het AI Cybersecurity Kwadrant geïntroduceerd. Het kwadrant helpt om de relatie tussen generatieve AI en cybersecurity te duiden binnen de razendsnelle ontwikkelingen op beide gebieden. Deze duiding vormt vervolgens de basis voor het handelingsperspectief.

Als een vervolg zullen factsheets uitgebracht worden over verschillende onderwerpen die in deze publicatie aan bod komen. Hierin wordt gedetailleerder ingegaan op de nieuwe uitdagingen en risico's van generatieve AI.

Wat is generatieve AI?

Generatieve AI is een specifieke vorm van kunstmatige intelligentie die nieuwe inhoud kan genereren, zoals tekst, beeld, code en meer. Deze technologie is complex en ontwikkelt zich in een ongekend tempo. Het unieke aan generatieve AI is dat het niet alleen informatie kan analyseren, maar ook kan creëren. Generatieve AI is daarmee in zeer diverse situaties toepasbaar.

Vanwege deze brede toepasbaarheid, wordt voor dit soort generatieve AI-systemen ook wel de term 'General Purpose AI' gebruikt. Dit is ook de term die in de Europese AI-verordening wordt gehanteerd: "*AI-modellen voor algemeen gebruik.*"

Deze vorm van AI zal een transformatieve impact hebben op cybersecurity. De impact zal dusdanig van aard zijn dat aanvullende expertise en investeringen nodig zijn om voldoende weerbaar te blijven. In deze publicatie lichten de AIVD en de RDI de betekenis van de ontwikkelingen voor cybersecurity toe.

AIVD en RDI

Als bovengenoemde partijen hebben wij ieder een eigen rol ten aanzien van de digitale weerbaarheid van Nederland. We hebben gezamenlijk verschillende expertises op het gebied van AI en cybersecurity en werken met elkaar samen om de nieuwste ontwikkelingen te duiden en onze doelgroepen hierover te informeren. Wij versterken de weerbaarheid en cybersecurity van de Nederlandse samenleving tegen bedreigingen die voortkomen uit onder meer de ontwikkeling van generatieve AI. Met deze publicatie willen we onze doelgroepen helpen om grip te krijgen op de kansen en risico's van generatieve AI.

Deze brochure is een vervolg op eerdere publicaties, zoals de expertblog AI: Cruciaal moment in de geschiedenis of een hype? (AIVD, RDI en NCSC, juni 2023, te vinden op www.ncsc.nl) en de brochure AI-systemen: ontwikkel ze veilig (AIVD, februari 2023, te vinden op www.aivd.nl).

Wat maakt generatieve AI transformatief?

Generatieve AI heeft specifieke eigenschappen die het fenomeen onderscheiden van andere typen cybersecurityuitdagingen. Vier van deze eigenschappen zetten we hieronder verder uiteen. In combinatie met de snelheid van de ontwikkelingen die generatieve AI nu doormaakt, zorgen deze eigenschappen voor een transformatieve impact. Dit betekent dat generatieve AI nieuwe typen dreigingen introduceert en de verdediging fundamentele veranderingen zal moeten ondergaan.

Huidige cybersecuritybenaderingen zijn niet toereikend om deze transformatieve impact het hoofd te bieden. De ontwikkelingen volgen elkaar snel op en de technische expertise bij vrijwel alle organisaties loopt achter.

Onderstaande eigenschappen vormen extra uitdagingen voor cybersecurity. Tegelijkertijd bieden deze eigenschappen nieuwe kansen voor de eigen verdediging.

Complexiteit en snelheid

Dreigingen aangedreven door generatieve AI kunnen zich snel ontwikkelen en veel 'slimmer' zijn in hun uitvoering. Traditionele beveiligingssystemen zijn mogelijk niet opgewassen tegen deze complexere bedreigingen. Eerder waren dit soort complexere aanvallen voorbehouden aan statelijke actoren, die meer mogelijkheden en capaciteiten hebben. Door de bredere beschikbaarheid van AI krijgen niet alleen zij meer mogelijkheden: ook criminelen kunnen hierdoor gemakkelijker complexere aanvallen uitvoeren.

Tegelijkertijd kan generatieve AI organisaties in staat stellen om hun cybersecurity te verbeteren, bijvoorbeeld bij het detecteren van aanvallen of het onderzoeken van netwerkverkeer. De toepassing van generatieve AI in de cybersecurityorganisatie kan dus voordelen bieden, als het verstandig en veilig wordt gebruikt.

Buitengewone personalisatie
Met generatieve AI kunnen aanvallen snel op maat worden gemaakt voor specifieke doelen op geautomatiseerde wijze. Dat maakt de dreigingen minder voorspelbaar. Hierdoor kunnen bestaande beveiligingsmechanismen die gebaseerd zijn op bekende patronen tekortschieten. Andersom kan generatieve AI juist goed worden gebruikt om nieuwe dreigingen te onderkennen. En maakt het 'personalisatie' van de eigen verdediging mogelijk.

Schaalbaarheid
Generatieve AI-technologieën maken het mogelijk om aanvallen op grotere schaal uit te voeren dan voorheen mogelijk was. Dit betekent dat een enkele aanvaller met beperkte middelen een impact kan hebben die veel groter is dan tot nu toe is gezien. De bestaande cybersecuritybenaderingen gaan vaak nog niet uit van aanvallen met een grotere impact. Wederom is deze trend ook mogelijk voor de verdedigende kant: met generatieve AI kunnen sneller cybersecurityanalyses worden gemaakt en maatregelen worden getroffen.

Autonomie
Generatieve AI is in staat om systemen te laten werken met *verminderde menselijke* tussenkomst:

- *Taakdeductie:* de mate waarin generatieve AI in staat is om een opdracht of taak uit te splitsen in subtaken, waardoor de AI in staat is de opdracht beter uit te voeren. Verschillende bestaande generatieve AI-systemen, zoals SWE-agent, zijn de eerste voorbeelden die op deze manier een opdracht opdelen in subtaken. De wijze waarop zij dit doen, is achteraf vaak niet helemaal te herleiden noch volledig te reproduceren.

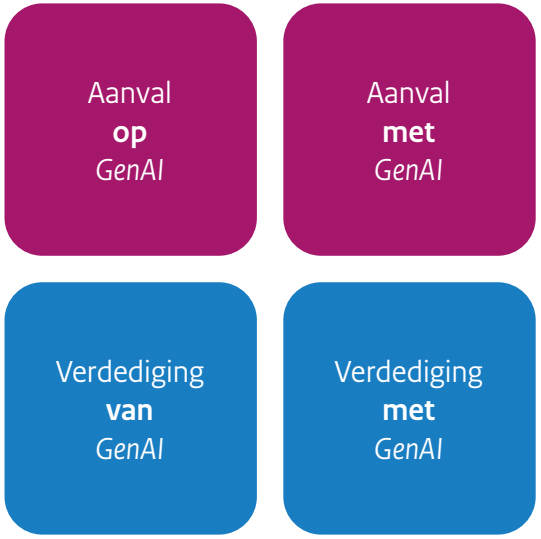
- *Taakdiversiteit:* de mate waarin generatieve AI in staat is om binnen één systeem verschillende taken uit te voeren, zoals tekstgeneratie, fotogeneratie en videogeneratie tegelijkertijd.

De AIVD en de RDI verwachten dat de ontwikkelingen op het gebied van taakdeductie en taakdiversiteit meer samenkomen op termijn. Deze samenkomst leidt tot nieuwe uitdagingen: de combinatie van enerzijds een hoge mate van taakdeductie en anderzijds een hoge mate van taakdiversiteit vergroot de kans dat generatieve AI onverwachte en ongewenste gedragingen vertoont. Ook daarmee dienen organisaties rekening te houden in hun cybersecurity-benadering wanneer zij generatieve AI toepassen. Daarbij is het van belang om een goede afweging te maken over het mandaat dat gegeven wordt aan generatieve AI-oplossingen: waar worden de systemen voor gebruikt en wat 'mag' de AI-oplossing doen?

Nieuwe uitdagingen: het AI Cybersecurity Kwadrant

Om de uitdagingen en kansen nader te duiden introduceren we het volgende *AI Cybersecurity Kwadrant*.

Hieronder worden per vlak van het kwadrant belangrijke aandachtspunten beschreven.



Het kwadrant bestaat uit twee assen. De bovenste vlakken hebben betrekking op de offensieve aspecten van generatieve AI. Dit zijn dus de dreigingen die gepaard kunnen gaan met generatieve AI.
De onderste as heeft betrekking op het verdedigen van AI-systemen en het gebruik van AI voor het verdedigen van systemen tegen cyberaanvallen, in het kader van weerbaarheid en veilig gebruik. Op de linkerzijde van het kwadrant is het generatieve AI-systeem het doel van de aanval of verdediging; rechts wordt generatieve AI gebruikt als het middel voor de aanval of verdediging.

Aanval

Aanvallen op generatieve AI

Hierbij gaat het om aanvallen die gericht zijn op het verstoren of misleiden van AI-systemen, inclusief pogingen om de trainingsdata te beïnvloeden of de AI op ongepaste wijze te manipuleren.

AI-systemen kunnen zelf worden aangevallen met zogenaamde adversarial attacks

Bovendien kunnen generatieve AI-systemen zelf worden aangevallen, met nieuwe vormen van digitale aanvallen: adversarial attacks. Bij een 'succesvolle' aanval gaat een systeem ongewenst of schadelijk gedrag vertonen. De ontwikkeling en het gebruik van generatieve AI-systemen kunnen op meerdere manieren worden aangevallen. Voorbeelden hiervan zijn inversion-aanvallen, waarbij een aanvaller probeert om potentieel gevoelige trainingsdata van een model te achterhalen. Het detecteren van zulke adversarial attacks is bovendien erg moeilijk. Ook is de integriteit van een model vrijwel niet 100% vast te stellen, aangezien een groot deel van de toeleveringsketen buiten de organisatie ligt en aanbieders van AI vaak intransparant zijn op dit vlak. Concreet betekent dit dat een model bijvoorbeeld een backdoor kan bevatten.

Naast kwaadwilligheid kan ook onwetendheid over hoe generatieve AI werkt leiden tot onbedoelde gevolgen. Dit soort risico's neemt toe wanneer organisaties op meer plekken in hun werkproces generatieve AI-systemen gaan inzetten.

Aanval **met** *GenAI*

Aanvallen met behulp van generatieve AI

In dit deel van het kwadrant vinden we aanvallen waarbij GenAI wordt ingezet als middel voor het uitvoeren van cyberaanvallen.

Deepfakes

Aanvallers kunnen generatieve AI gebruiken om niet van echt te onderscheiden teksten, foto's of video's te creëren. Hierdoor neemt de kans op chantage, reputatieschade en datalekken toe. Denk aan het fenomeen van 'CEO-fraude', een vorm van phishing met deepfake-audio en -beeldmateriaal. Of aan het om de tuin leiden van biometrische beveiligingssystemen die gebruikmaken van stemherkenning, door middel van een deepfake-stem.

Veranderingen in gangbare cyberaanvallen

Generatieve AI kan diverse onderdelen van de Cyber Kill Chain ondersteunen[1] en bestaande cybersecurity risico's op verschillende manieren vergroten. Zo maakt AI het eenvoudiger om een doelgerichte aanval op te zetten. Het meest bekende voorbeeld hiervan zijn phishing mails die met generatieve AI op grote schaal gepersonaliseerd worden met informatie van bijvoorbeeld sociale media. Daarnaast stelt generatieve AI hackers in staat om complexe systemen efficiënter te penetreren, bijvoorbeeld door automatisch te zoeken naar kwetsbaarheden en efficiënter malwarecode te schrijven. Ook maakt generatieve AI het voor hackers mogelijk om vragen te stellen over gespecialiseerde systemen, zoals SCADA-systemen in de kritieke infrastructuur, die normaliter diepgaande expertise vereisen. Dergelijke systemen worden zo potentieel toegankelijker voor minder gespecialiseerde actoren die ongeautoriseerd toegang willen verkrijgen.

[1] De Cyber Kill Chain is een hulpmiddel dat wereldwijd gebruikt wordt om inzicht te geven in de werkwijze van cyberaanvallers. Meer lezen over de Cyber Kill Chain, met voorbeelden? Deze keten wordt uiteengezet in de gezamenlijke AIVD-MIVD-publicatie Cyberaanvallen door statelijke actoren – zeven momenten om een aanval te stoppen (juni 2021).

Verdediging

Verdediging **van** *GenAI*

Verdediging van generatieve AI

Verdedigingsmaatregelen zijn gericht op het beschermen van AI-systemen tegen externe aanvallen en interne misbruiken. Dit omvat het veiligstellen van trainingsdata en het model, evenals het ontwikkelen van robuuste systemen die zichzelf kunnen verdedigen tegen aanvallen.

Aanvullend aanvalsoppervlak en complexiteit van AI-systemen

De verdediging van generatieve AI-systemen draait om bescherming: van de gegevens en de tools die door het AI-systeem wordt gebruikt, van de AI-algoritmen zelf en van de ondersteunende infrastructuur. Deze drie zaken zorgen ervoor dat generatieve AI-systemen een groter aanvalsoppervlak hebben dan andere informatiesystemen. Bovendien is de verwachting dat organisaties generatieve AI-systemen breed zullen inzetten, waardoor het aanvalsoppervlak nog meer toeneemt. Bij toegang tot zo'n AI-systeem heeft de aanvaller bovendien ineens een heel krachtig Zwitsers zakmes in handen. Generatieve AI vereist dan ook aanvullende beveiliging bovenop de standaard cybersecurity-maatregelen die een organisatie al zou moeten nemen: er is een groter oppervlak dat beschermd moet worden, onder andere tegen de hierboven genoemde adversarial attacks. De complexiteit en ondoorzichtigheid van generatieve AI-systemen vormen hierbij een extra uitdaging.

Verdediging **met** *GenAI*

Verdediging met behulp van generatieve AI

Generatieve AI wordt ook gebruikt als een hulpmiddel om te verdedigen tegen cyberaanvallen, door middel van real-time dreigingsdetectie en respons, of het combineren van en handelen op dreigingsinformatie uit verschillende bronnen.

Afhankelijkheid van generatieve AI en soevereiniteitsvraagstukken

De AIVD en de RDI voorzien dat de inzet van generatieve AI op defensief gebied onvermijdelijk zal worden, vanwege de snelle ontwikkelingen van generatieve AI op offensief gebied. Generatieve AI biedt hier zeer veel kansen, omdat het organisaties in staat stelt om grote hoeveelheden en verschillende soorten data te analyseren, voorspellingen te maken en mogelijk automatische responsacties uit te voeren.

De risico's op onvoorspelbaar of ongewenst gedrag nemen echter toe wanneer het generatieve AI-systeem meer 'mandaat' krijgt om taken uit te voeren. Daarnaast bestaan er afhankelijkheidsrisico's bij de inzet van generatieve AI-modellen die ontwikkeld zijn door een externe partij. Het is altijd belangrijk om bewust te zijn van deze extra risico's. Het gaat onder andere om de mate van privacy en bescherming van gevoelige data, naleving van regelgeving (onder andere de Europese AI-verordening) en het voorkomen van vendor lock-in. De meeste state of the art generatieve AI-modellen zijn afkomstig uit niet-Europese landen, waar de EU-verordening niet zal gelden. De Universiteit van Stanford heeft eerder laten zien dat de compliance met deze verordening nog onvoldoende is.

Hoe nu verder?

In het licht van deze analyse is het duidelijk dat de komst van generatieve AI aanzienlijke uitdagingen met zich meebrengt. Hoewel generatieve AI positieve toepassingen heeft, moeten organisaties ook voorbereid zijn op nieuwe soorten uitdagingen en dreigingen. Het is van belang om dit nieuwe risicodomein te erkennen en hiernaar te handelen – zowel om generatieve AI op een verstandige manier te gebruiken, als om de veranderende dreiging door generatieve AI tegen te kunnen gaan.

Vanuit onze kennis- en expertiserol raden we organisaties dringend aan het AI Cybersecurity Kwadrant te gebruiken om de kansen, uitdagingen en risico's voor de eigen organisatie inzichtelijk te maken. Daarnaast is het noodzakelijk de risicoanalyse regelmatig bij te werken om in de pas te blijven met de razendsnelle ontwikkelingen rondom generatieve AI.

Lees meer over AI:

Scan de QR-code om de brochure te lezen.

Deze publicatie is een gezamenlijke uitgave van:

Algemene Inlichtingen- en Veiligheidsdienst (AIVD)
aivd.nl

Rijksinspectie Digitale Infrastructuur (RDI)
rdi.nl

September 2024