Nationaal Cyber Security Centrum
Ministerie van Justitie en Veiligheid

Artificial Intelligence in de Cyber Kill ChAIn

NCSC CTI-report

Publicatiedatum: 24 september 2026

Reacties zijn welkom op info@ncsc.nl.

Inhoudsopgave

1 Inleiding

Artificial Intelligence (hierna AI) wordt in steeds meer systemen en processen geïntegreerd. Dit biedt kansen en mogelijkheden om de digitale weerbaarheid naar een nieuw niveau te tillen. Tegelijkertijd verandert deze ontwikkeling het dreigingslandschap, want ook kwaadwillenden ontdekken AI en weten het steeds beter te gebruiken. Dit CTI-report focust zich op dat laatste: hoe zetten aanvallers AI in bij digitale aanvallen?

Dit CTI-report belicht hoe onderdelen van digitale aanvallen in de In, Through, en Out-fase van de Unified Kill Chain[i] met behulp van AI verbeterd, versneld of geautomatiseerd worden. Deze ontwikkelingen hebben impact op het dreigingslandschap en vragen om versterking en aanpassingen van de bestaande beveiligingsmaatregelen.

- In dit rapport beschrijft het NCSC op basis van feitelijke waarnemingen hoe aanvallers op dit moment (generatieve) AI inzetten voor het uitvoeren van taken.
- Het NCSC speculeert niet over op welke manier AI nu of in toekomst ingezet zou kúnnen worden. Dit rapport is daarmee geen verkenning van toekomstige scenario's.
- In hoeverre AI gebruikt is in een aanval, is niet altijd evident. Voor de benoemde voorbeelden hebben beveiligingsonderzoekers op basis van specifieke kenmerken of informatie kunnen vaststellen dat AI is gebruikt bij de aanval. In andere gevallen is het niet te achterhalen of AI is gebruikt of bestaat er slechts een vermoeden dat AI is gebruikt. Die gevallen worden niet beschreven in dit rapport.
- Het gebruik van kant-en-klare tooling, die voorafgaand aan een cyberaanval met behulp van AI is ontwikkeld, is buiten scope voor deze publicatie. Ook aanvallen op AI-modellen vallen buiten scope.

Wat verstaan we onder AI?

AI verwijst naar systemen die op basis van informatie uit context van de vraag of input(data) met een zekere mate van zelfstandigheid een actie kunnen ondernemen of antwoord op vragen kunnen aandragen. AI werkt met algoritmen die in software geschreven zijn. Een algoritme kun je zien als een recept; het is een set aan instructies voor het bereiken van een bepaald doel. Deze algoritmen kan je weer verder opdelen in drie verschillende groepen:

- Expertsysteem, niet-zelflerende AI
- Machine Learning, wel zelflerende AI

[i] De Unified Kill Chain is een conceptueel model dat wordt gebruikt voor het beschrijven van cyberaanvallen. Voor meer informatie: https://www.unifiedkillchain.com/.

- Generatieve AI produceert (onder andere) *nieuwe* content in de vorm van data - zoals audio, tekst, video en afbeeldingen. De meest bekende generatieve AI-modellen zijn op dit moment taalmodellen (Large Language Models, LLM's).
    - Agentic AI is een geavanceerdere vorm van generatieve AI die zelfstandig acties plant en uitvoert. De gebruiker geeft een doel, waarna de AI tussenstappen bepaalt, informatie verzamelt, digitale hulpmiddelen gebruikt en de aanpak aanpast op basis van de resultaten. De AI voert dit doel uit met behulp van een set aan AI-agents die gecoördineerd acties uitvoeren. Dit maakt Agentic AI geschikt voor complexe taken, maar vergroot ook het risico op ongewenste acties wanneer de opdracht onduidelijk is, de AI verkeerde conclusies trekt of te veel rechten heeft.

Een uitleg van een Expertsysteem, Machine Learning, en een uitgebreidere uitleg van generatieve AI is te vinden in *Appendix A: Wat is AI?*. De voorbeelden in dit rapport gaan vooral over het gebruik van generatieve AI door kwaadwillenden, bijvoorbeeld voor het genereren van malware of uitvoeren van andere aanvalsstappen. Ook de voorbeelden waarbij AI (grotendeels) autonoom acties uitvoert, zijn vormen van generatieve AI.

Disclaimer:

De verzameling voorbeelden en beschrijvingen van Techniques, Tactics, and Procedures (TTPs) in dit rapport vormen geen uitputtende lijst van alle manieren hoe kwaadwillenden AI inzetten. Het is een selectie van een aantal kenmerkende voorbeelden die de veranderingen van het dreigingslandschap illustreren. In sommige gevallen bevatten rapporten van incidenten geen technische indicatoren bevatten om de claims eigenstandig te verifiëren en zijn de rapporten soms meer speculatief van aard. Dit is ook onderdeel van een breder vraagstuk over hoe je AI-gebruik met zekerheid kan detecteren wanneer voor de hand liggende indicatoren zoals bijvoorbeeld prompt-instructies of comments in malafide code niet aanwezig zijn.

Bovendien gaan de ontwikkelingen van AI snel en kan maanden het dreigingslandschap er over 6 tot 12 maanden weer anders uitzien vanwege de nieuwe capaciteiten die AI-modellen bieden. Het NCSC blijft je informeren over nieuwe ontwikkelingen, maar adviseert ook om zelf alert te blijven op veranderingen.

2 Analyse waargenomen AI-aanvalstechnieken

Dit CTI-report belicht, aan de hand van de Unified Kill Chain (UKC)[1], hoe onderdelen van digitale aanvallen in de In, Through, en Out-fase met behulp van AI geautomatiseerd, verbeterd of versneld kunnen worden.

Tijdens het uitvoeren van een cyberaanval dient een kwaadwillende ten minste drie operationele fasen te doorlopen: binnendringen (In), beweging door het netwerk (Through) en het behalen van (strategische) doelstellingen (Out). Deze operationele fasen bevatten tactische stappen (of tactieken), waarmee de aanvaller diens operationele doel probeert te behalen. De Unified Kill Chain (UKC) beschrijft de operationele doelstelling en bijbehorende tactieken volgordelijk. Alhoewel niet iedere stap altijd netjes achter elkaar wordt uitgevoerd, zal een kwaadwillende min of meer al deze fases doorlopen bij het uitvoeren van een digitale aanval.

2.1 Getting In

Gedurende de In-fase van de UKC probeert de aanvaller toegang te krijgen tot de beveiligde omgeving van het doelwit. Het doel van deze stap in de aanvalsketen is om een vorm van aanwezigheid te creëren in de omgeving van zijn doelwit. Een voorbeeld van dit proces is gevisualiseerd in *Appendix B: Visualisatie In-fase*.

Deze sectie belicht vier manieren waarop kwaadwillenden AI inzetten voor het binnendringen van digitale systemen en netwerken: (1) *verkennen van het aanvalsoppervlak*, (2) *kwetsbaarhedenonderzoek*, (3) *genereren van malware* en (4) *verbinden van meerdere stappen*.

Verkennen van het aanvalsoppervlak

Aanvallers gebruiken al jaren scantools om grote aantallen domeinen, IP-adressen en internetdiensten te onderzoeken, de digitale infrastructuur van doelwitten verkennen om zo aanvalsoppervlakken in kaart te brengen. Met AI kunnen deze activiteiten verder worden geautomatiseerd.

- Geavanceerde actoren bouwden tooling waarmee AI-modellen verkenningsactiviteiten (reconnaissance) uitvoeren, zo rapporteren beveiligingsonderzoekers van Google Threat Intelligence[2] en Hunt.io[3]. Denk hierbij aan het in kaart brengen van informatie over beoogde slachtoffers, zoals informatie over de organisatiestructuur en met welke partners er wordt samengewerkt. Het gebruik van deze tooling verkort vooral de handmatige analysetijd.
- Een dreigingsactor (*UNC6418*) gebruikte AI voor het verzamelen van gevoelige accountgegevens en mailadressen, waarna de verzamelde accounts doelwit werden van een phishingcampagne.[4]

Samenvattend

AI wordt door kwaadwillenden ingezet voor het geautomatiseerd verwerken van scanresultaten

en het daaruit rangschikken van potentieel relevante doelwitten op basis van voorkeuren van de aanvaller, om zo te bepalen welke vervolgstap het meest kansrijk is. Deze ontwikkeling zorgt ervoor dat de tijd tussen het initiëren van een scan en het uitvoeren van de vervolgstap korter wordt.

Kwetsbaarhedenonderzoek

AI kan onbekende kwetsbaarheden (zero-days) vinden, maar ook op basis van een patchbeschrijving zoeken naar de kwetsbaarheid en deze misbruiken (*n*-day). Beide methoden worden ingezet door aanvallers. Verschillende rapportages van beveiligingsbedrijven beschrijven dat verschillende soorten actoren in verschillende mate gebruikmaken van AI voor het vinden en uitbuiten van zowel bekende als onbekende kwetsbaarheden.

- Een cybercriminele actor heeft AI ingezet om een zero-day kwetsbaarheid te ontdekken voor het omzeilen van twee-factorauthenticatie. De actor gebruikte een exploitatiescript waarin Google indicaties vond van het gebruik van AI.[2]
- AI wordt gebruikt voor het geautomatiseerd zoeken naar Proof-of-Concept (PoC) code voor bekende kwetsbaarheden, zo beschrijven beveiligingsonderzoekers van Palo Alto Unit42. De AI-agent verkende geautomatiseerd kwetsbare doelwitten waarna het ook pogingen tot exploitatie ondernam. De exploitatiepogingen mislukten, waarna de AI-agent op eigen initiatief naar PoC-code voor een andere kwetsbaarheid zocht en daarmee opnieuw exploitatiepogingen uitvoerde. Ook die pogingen mislukten; de systemen van het doelwit waren correct geconfigureerd, waardoor de exploit niet werkte. Handmatige exploitatiepogingen door dezelfde aanvaller met andere kwetsbaarheden, waren wel succesvol.[5]
- Google en Microsoft documenteerden hoe geavanceerde actoren experimenteren met het gebruik van AI voor kwetsbaarhedenanalyse. Google beschrijft dat een geavanceerde actor (*APT45*) duizenden prompts heeft verstuurd om verschillende bekende kwetsbaarheden te analyseren en exploits te valideren.[2] Microsoft beschrijft dat de actor *Emerald Sleet* (*Kimsuky*) al meerdere jaren AI gebruikt(e) voor het onderzoeken van bekende kwetsbaarheden.[6]
- Beveiligingsbedrijf Gambit Security beschrijft hoe een kwaadwillende bij een aanvalscampagne gericht op Mexicaanse overheden AI heeft gebruikt voor de ontwikkeling van exploits. In één sessie maakte de AI-agent een zelfstandig Python-exploit met ondersteuning voor proxies. In zeven minuten voerde het systeem acht opeenvolgende aanpassingen uit. Het probeerde verschillende payloadvormen, Unicode-tekens en base64-encoding totdat het een werkende variant had. Gambit Security stelt dat het na de eerste weigering van het AI-model, ongeveer 40 minuten duurde tot willekeurige code kon worden uitgevoerd (*remote code execution; RCE*). Het arsenaal van de aanvaller bevatte daarnaast 20 aangepaste exploits voor 20 verschillende kwetsbaarheden.[7]

Samenvattend
Kwaadwillenden experimenteren met AI en gebruiken het om kwetsbaarheden te vinden en uit te buiten. De drempel om bekende kwetsbaarheden uit te buiten wordt lager en geavanceerdere

actoren kunnen de capaciteiten waarmee ze (onbekende) kwetsbaarheden onderzoeken uitbreiden.

Genereren van malware

Met behulp van AI-modellen kunnen kwaadwillenden malware-code genereren. Dit vereist vaak dat vangrails (*guardrails*)[ii] van modellen worden omzeild of dat er toegang is tot een AI-model zonder vangrails. Met behulp van AI hoeft de aanvaller zelf geen code meer te schrijven, maar slechts prompts die instrueren wat de malware zou moeten kunnen.

- Onderzoekers van beveiligingsbedrijf Cisco Talos beschrijven aan de hand van meerdere incidenten hoe kwaadwillenden AI inzetten om als software engineer te fungeren. Hoewel AI wordt gebruikt voor *resource development*, zoals voor het genereren van malware, stellen de onderzoekers van Cisco Talos dat de technische capaciteiten van de aanvaller de impact bepalen. Vooral geavanceerdere aanvallers ontwikkelen geavanceerde malware en tooling met AI. In dat geval geldt AI als *force multiplier*.[8]
- CERT.PL rapporteerde over een functionaliteit voor het beschadigen van bestanden (*file corruption*) in wiper-malware (*LazyWiper*), die mogelijk met AI is ontwikkeld. Dit baseert CERT.PL onder andere op het feit dat de specifieke functie voor het beschadigen van bestanden in de code afwijkt van de rest van de code op het gebied van onder andere structuur.[9]
- De PowerShell infostealer-malware (CHERRYPIE/ChocoShell) die een dreigingsactor (*UNC7005*) gebruikte in een campagne gericht op initiële toegang, bevat volgens Google Threat Intelligence Group indicaties van LLM-gegenereerde code.[10]
- Het NCSC heeft tijdens eigen onderzoek van enkele aanvalspogingen zelf ook malware-samples aangetroffen waarbij zeer waarschijnlijk AI is gebruikt om deze te genereren. Indicaties hiervoor waren onder andere te zien in de specifieke opbouw van de code. Het NCSC beoordeelt deze samples als niet-geavanceerd en van technisch laagwaardige kwaliteit. De onderzochte aanvalspogingen waren niet succesvol.

Samenvattend

Met AI kan een aanvaller sneller malware ontwikkelen. AI vergroot het potentieel van het ontwikkelproces en kan, mede afhankelijk van de technische vaardigheid van menselijke operator die de AI-agent instrueert, resulteren in meer geavanceerde code. De rol van de aanvaller verandert van ontwikkelaar van malware naar opdrachtgever. Bovendien kan AI, in tegenstelling tot een menselijke programmeur, een continu ontwikkelproces ondersteunen zolang aanwezige rekenkracht/tokenbudget dat toelaat.

Verbinden van meerdere onderdelen van de aanvalsketen

Met behulp van AI kunnen onderdelen van een aanvalsketen aan elkaar worden verbonden.

[ii] Guardrails zijn regels en filters in een AI-model die input en output controleren en filteren op ongewenste content en gedrag.

- EvilTokens, een Phishing-as-a-Service toolkit (PHaaS), gebruikt AI voor gepersonaliseerde phishingmails bij device code phishing[iii]-aanvallen.[11] Na succesvolle accountovername bracht EvilTokens via Microsoft Graph de organisatie, functies, managers, medewerkers en financiële relaties van het slachtoffer in kaart. Met behulp van Groq API en het model *llama-3.1-8b-instant* verwerkt EvilTokens vervolgens maximaal 5.000 gestolen e-mails, in batches van 250 berichten.[12] Het eerste model zocht naar betalingen, facturen, leveranciers en lopende financiële gesprekken. Het tweede model (*llama-3.3-70b-versatile*) voegde de resultaten samen, kende een score voor bruikbaarheid toe en schreef drie gerichte fraude-e-mails. Een derde model zorgde voor vertalingen wanneer dit nodig was. Zodoende leiden de resultaten van de eerste aanval, geautomatiseerd tot een nieuwe volgende aanval.
- Hoewel automatisering binnen PhaaS-toolkits niet nieuw is, toont EvilTokens wel aan hoe het zich, met behulp van de integratie van AI-functionaliteiten, aanpast aan het gedrag van het doelwit en daarna ongestructureerde informatie zoals e-mails, schrijfstijlen, financiële relaties en bestaande gesprekken op grote snelheid analyseert en verwerkt voor nieuwe aanvallen.
- In een andere casus beschrijven beveiligingsonderzoekers van Cisco Talos hoe cybercriminelen (UAT-10147) gebruikmaakten van AI voor het uitvoeren van meerdere aanvalsstappen. De groep zette AI onder meer in voor het genereren van aanvalsinstructies, exploitatiescripts en post-compromittatie-activiteit.[13]

Samenvattend

Bij de bovenstaande voorbeelden gebruiken aanvallers vooral bestaande aanvalstechnieken en automatiseren ze deze, waardoor de aanvallen efficiënt en vaak versneld kunnen worden uitgevoerd. AI voert een aantal stappen in een doorlopende aanvalsketen uit, met beperkte tussentijds interactie van de aanvaller zelf. Een uitgebreide verkenning van het aanvalsoppervlak en onderzoek naar kwetsbaarheden met behulp van AI ondersteunt betere exploit ontwikkeling. Een social engineering-aanval wordt dankzij AI effectiever, door versneld gebruik van bemachtigde informatie na een accountovername. Ook genereren, testen en verbeteren AI-modellen malware waarbij tijdens de ontwikkeling de AI al nadenkt over persistentie, ontwijken van detectie en C2-kanalen. Kortom: de In-fase wordt sneller en efficiënter doorlopen.

Tegelijkertijd betekent de inzet van AI niet automatisch dat aanvallen ook beter zijn wanneer technisch minder onderlegde actoren dit proberen of geen verificatie van de output uitvoeren. Malware kan onbruikbare code bevatten. Ook kunnen acties van AI-modellen en agents veel ruis veroorzaken die gedetecteerd kan worden. Vooral voor actoren waarbij onder de radar blijven minder van belang is en die snel hun doel willen bereiken, kan AI uitkomst bieden. Denk hierbij aan opportunistische digitale aanvallen voor financieel gewin.

[iii] Bij Device Code Phishing misbruikt een aanvaller het authenticatieproces voor apparaten met beperkte interfaces, zoals printers en smart TV's. Een gebruiker wordt misleid om een authenticatiecode te delen met een aanvaller, waardoor de aanvaller een geauthentiseerde sessie krijgt met het betreffende account.

2.2 Hacking through

In de Through-fase van de Unified Kill Chain probeert een aanvaller de verkregen toegang uit te breiden door rechten te verhogen (*privilege escalation*) en zich door het netwerk te bewegen (*lateral movement*) richting het doel of een nieuwe positie. Tijdens deze aanvalsfase opereert de aanvaller in de systemen van het slachtoffer. Een voorbeeld van dit proces is gevisualiseerd in *Appendix C: Visualisatie Through-fase*.

Deze sectie beschrijft drie manieren waarop kwaadwillenden AI inzetten voor het uitbreiden van een verworven toegangspositie: (1) *post-compromittatie verkenning*, (2) *context-gebaseerd en dynamisch genereren en uitvoeren van malafide code* en (3) *beweging in het netwerk: verhogen van rechten (privilege escalation)* en *beweging door het netwerk (lateral movement)*.

Post-compromittatie verkenning

AI automatiseert netwerkverkenningen en identificeert mogelijkheden tot beweging door het netwerk. AI genereert daarvoor scripts, voert autonoom acties uit of analyseert grote hoeveelheden informatie. Drie voorbeelden hiervan:

- Een AI-model haalde systeeminformatie van een gecompromitteerd systeem op en analyseerde deze om mogelijkheden voor laterale beweging te identificeren.[7]
- Een AI-agent zocht tijdens aanvallen op organisaties in de overheids- en financiële sectoren in Latijns-Amerika naar netwerkinformatie, inloggegevens in configuratiebestanden en SSH-sleutels. Een aantal verkenningstaken zijn volgens de onderzoekers van TrendAI vrijwel per direct succesvol uitgevoerd, wat wijst op gebruik van AI omdat handmatige uitvoering van dergelijke taken doorgaans tijdrovend en foutgevoelig is.[14]
- Ook wisten aanvallers met behulp van AI het netwerk te verkennen en hierbij bereikbare bestanden te identificeren en accounts met bruikbare rechten middels SUID/SGID te vinden.[15]

Samenvattend

Snelheid en flexibiliteit blijken de voornaamste meerwaarden van het gebruik van AI in de post-compromittatiefase. AI versterkt bestaande methodes en kan dit op grotere schaal geautomatiseerd uitvoeren.

Malware past zich tijdens de aanval aan de omgeving en context aan

Kwaadwillenden gebruiken malware met een AI-component waarmee malafide activiteiten tijdens uitvoering dynamisch gegenereerd worden. De malware past zich dan aan de omgeving aan door een ingebouwde functionaliteit die zich toespitst op het geïnfecteerde systeem.

Deze malware wordt voornamelijk in de context van AI-gecontroleerde cyberaanvallen uitgevoerd: de malware dient een specifiek doel en wordt daartoe ingezet.

- CERT-UA (Oekraïne) en Google beschrijven de inzet van LAMEHUG/PROMPTSTEAL die via een API-koppeling (HuggingFace) communiceert met een LLM (Qwen2.5-Coder-32B-

Instruct). Op deze manier kan er via prompts informatie over het systeem en netwerk worden verkregen. Ook kan de malware specifiek zoeken naar Microsoft Office documenten, en deze exfiltreren met SFTP of HTTP POST verzoeken. Zowel CERT-UA als Google geven aan dat APT28 deze malware zou hebben gebruikt.[16,17]

- QUIETVAULT/S1ngularity-malware gebruikt AI-modellen die beschikbaar zijn op het systeem van het slachtoffer. Met een command line-interface (CLI) van bekende AI-modellen die aanwezig zijn op het systeem van het slachtoffer, zoekt de malware naar gevoelige bestanden zoals .env-bestanden, SSH-sleutels, npm- en GitHubtokens en cryptowallets. De malware werd verspreid via de software-toeleveringsketen, door veelgebruikte publieke softwarecomponenten te besmetten[iv].[18,19]
- Experimentele malwarevarianten, zoals PROMPTFLUX en PROMPTLOCK tonen dat AI-modellen ook worden ingezet voor het testen met dynamische obfuscatie en versleuteling.[17,20] Ook kan dergelijke custom malware ontwikkeld worden als dropper voor andere tooling zoals Sliver (een pentest-framework).[21]
- HONESTCUE-malware bevraagt de Gemini API en ontvangt malafide code dat direct in memory wordt uitgevoerd. Hierbij laat de gegenereerde code geen bestanden of sporen achter op het netwerk omdat de code direct wordt uitgevoerd. Hoewel Google HONESTCUE nog niet linkt aan actieve campagnes en toont de malware volgens Google wel aan hoe actoren met beperkte technische kennis dergelijke capaciteiten ontwikkelen.[4]

Samenvattend

AI maakt het mogelijk om code te genereren op basis van meerdere parameters, zoals systeeminformatie en een doelstelling vanuit de actor. Malware kan zich aan een omgeving aanpassen dankzij de adaptieve capaciteiten die ontstaan door het gebruik van AI. Door AI heeft een aanvaller minder voorkennis van het doelsysteem nodig bij het maken van de malware en werkt de malware mogelijk op een bredere set aan doelsystemen.

Het dynamisch genereren van malafide code levert operationeel voordeel voor aanvallers op. Malware is breder inzetbaar en kan functionaliteiten creëren op basis van omgevingsinformatie en doelstellingen. Bovendien introduceert deze functionaliteit uitdagingen voor verdedigers, omdat de malware zich telkens anders gedraagt.

Preventie- en detectiemaatregelen op basis van technische kenmerken (zoals hashes en malware-specifiek gedrag) zijn minder effectief tegen dit type malware, omdat de aanpassing aan de context leidt tot telkens nieuwe code en gedragingen. De aanvaller vergroot daarmee diens kans dat de malafide code succesvol wordt uitgevoerd.

[iv] Het verspreiden van malware via de toeleveringsketen is een trend en een actueel relevante dreiging. Voor meer informatie over deze dreiging, zie https://www.ncsc.nl/expertblogs/het-nieuwe-trojaanse-paard-zit-in-je-dependencies.

Beweging in het netwerk: verhogen van rechten en lateraal bewegen

Voor het bewegen door het netwerk kan ook AI worden gebruikt. Uit onderstaande voorbeelden blijkt dat aanvallers hiermee experimenteren op het gebied van verhogen van rechten en laterale bewegingen.

- Beveiligingsonderzoekers van Sysdig rapporteerden in november 2025 over een aanval waarbij de aanvaller binnen 10 minuten na initiële toegang diens rechten verhoogde. De aanvaller gebruikte een mogelijk met AI gegenereerd script, om middels een Lambda-functie rechten te verhogen. Vervolgens probeerde de aanvaller meerdere accounts over te nemen door account-IDs te enumereren en toegang tot deze accounts te krijgen. Tijdens de pogingen voor laterale beweging namen de onderzoekers ook AI-hallucinaties waar zoals het genereren een account-ID wat niet bestond en een account-ID van een extern account.[22]
- Bij aanvallen op Mexicaanse overheidsinstanties verhoogde een AI-agent rechten door een SSH-sleutel toe te voegen aan de lijst met geautoriseerde root-sleutels, na het scannen van bewerkbare crontabs (geplande uitvoering van een script of commando). Een menselijke operator gaf tussentijds instructies.[7] Deze casus vertoont overlap met de hierboven aangehaalde incidenten in Latijns-Amerika die door TrendAI zijn onderzocht.[14] Hoewel een menselijke operator deze aanval ook kan uitvoeren, nam een AI-model hier het werk grotendeels uit handen.

Samenvattend

AI helpt aanvallers met het vinden van routes binnen het netwerk om lateraal te bewegen of rechten te verhogen. Deze acties vereisen kennis van de netwerkinfrastructuur en de werking daarvan. AI kan aanvallers helpen bij het interpreteren van verkenningsinformatie en daarmee helpen met het uitvoeren van de juiste acties. Verschillende voorbeelden tonen de interesse van kwaadwillenden om AI te gebruiken voor het verhogen van rechten of laterale beweging in een netwerk.

2.3 Taking it out

De Out-fase is de laatste set aan stappen binnen de Unified Kill Chain. Hierin probeert een aanvaller doelen te behalen, zoals data-exfiltratie en het versleutelen van systemen. Doelen verschillen per aanval en data exfiltratie en impact hoeven elkaar niet altijd op te volgen. De aanvaller kiest op dit moment wat het beste bij de intentie achter de aanval past. De aanvalsactiviteiten in deze fase zijn meestal bedoeld om:

- Waardevolle informatie of gegevens in kaart te brengen en te verzamelen voorafgaande aan exfiltratie (*collection*);
- Waardevolle informatie of gegevens te stelen (*exfiltration*);
- Digitale processen of systemen te verstoren of saboteren (*impact*).

Gevoelige data lokaliseren en verzamelen

Een van de doelen van een aanvaller kan zijn om gevoelige gegevens te exfiltreren. Uit onderstaande voorbeelden blijkt dat aanvallers in tenminste twee gevallen vooraf AI hebben gebruikt om gevoelige data te lokaliseren en te verzamelen.

- In het eerdergenoemde rapport van securitybedrijf Gambit Security wordt onder andere beschreven hoe de aanvaller (na de In- en Through-fase) verschillende verkenningsactiviteiten bij de Mexicaanse Belastingdienst heeft uitgevoerd, waaronder het verzamelen van tabellen met persoonsgegevens en het in kaart brengen van de hele database. Op deze manier onderzocht de aanvaller waar waardevolle persoonsgegevens opgeslagen staan, zoals namen, adressen en biometrische gegevens.[7]
- AI-leverancier Anthropic stelt dat een geavanceerde aanvaller Claude Code in de dataverzamelingsfase vrijwel autonoom heeft laten handelen.[23] De actor gaf Claude de opdracht om zelfstandig databases en systemen te doorzoeken, gegevens te extraheren, resultaten te analyseren om vertrouwelijke informatie te identificeren en de bevindingen te categoriseren op basis van hun inlichtingenwaarde. Bij dit incident verwerkte AI zelfstandig grote hoeveelheden data en identificeerde automatisch waardevolle informatie, zonder dat er menselijke analyse nodig was. De actor hoefde alleen nog de bevindingen te controleren en vervolgacties te bevestigen.

Samenvattend

AI is ingezet als operationele assistent door op basis van concrete opdrachten binnen het systeem te zoeken naar gevoelige data. Het AI-model voerde de collectie van data niet zelfstandig uit, maar verminderde wel het handmatige werk. Daarnaast is data met behulp van AI vrijwel geautomatiseerd verzameld en geanalyseerd.

Data-exfiltratie en doelen behalen

Na het lokaliseren en verzamelen van gevoelige data wil de aanvaller deze data vaak exfiltreren om hiermee het initiële doel van de operatie te kunnen bereiken, bijvoorbeeld voor afpersing, fraude of spionagedoeleinden. Een aanvaller kan AI bijvoorbeeld als codeassistent inzetten om een tool te genereren voor data-exfiltratie.

- Een aanvaller ontwikkelde met behulp van AI exfiltratie-scripts in python om op grote schaal gevoelige configuratie- en inloggegevens buit te maken. De data zijn via HTTPS verstuurd naar *webhook[.]site* waardoor het leek op legitiem verkeer naar een bestaande SaaS-service.[13]
- Aanvallers hebben ook data geëxfiltreerd door slachtoffers te misleiden met legitiem ogende AI-functionaliteiten in PDF-readers, waarin een backdoor (FlutterShell) verpakt zat.[24] Slachtoffers konden de malafide PDF-reader vragen om een samenvatting van een document te maken. In plaats van het verzoek direct naar een AI-model te sturen, ving de actor het document als Man-in-the-Middle af. Het slachtoffer ontving wel een samenvatting van het document, maar had dus onbewust zijn eigen data geëxfiltreerd naar de aanvaller. Via malvertising werd de backdoor verspreid.
- Bij de S1ngularity-aanval op Nx in 2025 paste aanvaller npm-versies van Nx aan met malafide code. Na installatie gaven deze kwaadaardige versies opdrachten aan lokale, aanwezige AI-codeassistenten, zoals Claude Code, Gemini CLI en Amazon Q, en omzeilden daarbij de gebruikelijke goedkeuringsmechanismes richting beheerders. De codeassistenten kregen de opdracht om het bestandssysteem te doorzoeken en gevoelige bestanden te vinden (zie ook het hoofdstuk *Context-gebaseerd dynamisch genereren en uitvoeren van malafide code* hierboven). AI ondersteunde daarmee vooral het snel en breed vinden van waardevolle gegevens. De daadwerkelijke exfiltratie verliep via reguliere malwarecode en de GitHub CLI. De buitgemaakte informatie werd geüpload naar openbare repositories.[25,26]
- Een vierde voorbeeld toont hoe een actor een eigen tool met behulp van AI genereerde, in plaats van gebruik te maken van publiek beschikbare hacktools.[14] Hierdoor was de kans kleiner dat de actor werd gedetecteerd door traditionele beveiligingsoplossingen die afhankelijk zijn van bekende malware signatures. Deze tool bevatte onder andere functionaliteiten voor data exfiltratie.

Samenvattend

AI voert acties in de Out-fase zoals dataexfiltratie nog zelden volledig autonoom uit. De beschreven voorbeelden laten zien dat de meerwaarde van AI vooral zit in het vinden, interpreteren en selecteren van gegevens voor exfiltratie en in het snel ontwikkelen of aanpassen van exfiltratie-tooling of scripts. Een overkoepelende analyse van Microsoft onderschrijft dit beeld.[27]

Impact: verstoring en sabotage

Het NCSC heeft tijdens het onderzoek geen bronnen of publicaties gevonden die door AI gefaciliteerde verstoring of sabotage van systemen of netwerken beschrijven.

3 Duiding

Uit de analyse blijkt dat in diverse soorten campagnes kwaadwillenden AI hebben gebruikt. Deze ontwikkeling zal de komende tijd verder zeer waarschijnlijk doorzetten. Dit heeft impact op het toekomstige dreigingslandschap en vereist een reactie van verdedigers.

De komst van AI-modellen heeft een significante invloed op de snelheid waarmee aanvallers de stappen uit de Unified Kill Chain tot uitvoering kunnen brengen. Kwaadwillenden zijn nu in staat om sneller en soms makkelijker hun (strategische) doelen te behalen. AI-modellen kunnen worden bevraagd op tactische en technische stappen waar kwaadwillenden zelf eerder vastliepen. Dit verlaagt de drempel voor minder technisch onderlegde aanvallers. AI-modellen kunnen daarnaast deze stappen sneller en geautomatiseerd uitvoeren. Deze ontwikkelingen hebben impact op hoe verdedigers moeten omgaan met digitale aanvallen.

In de aangehaalde voorbeelden is te zien dat aanvallers AI vooral inzetten om bekende en bestaande aanvalstechnieken te automatiseren of autonoom te laten uitvoeren. De voorbeelden tonen hoe sommige onderdelen van een aanval binnen enkele minuten plaatsvinden, terwijl dit voorheen handmatig een stuk langer zou duren. Wanneer actoren beter overweg kunnen met AI-modellen of de capaciteiten van AI-modellen zich verder ontwikkelen, kunnen deze doorlooptijden mogelijk nog korter worden. AI is dus een tool die tijdens aanvallen wordt ingezet en wordt daarmee een onderdeel van de gehele dreiging.

Veel AI-activiteit in de In-fase; Through en Out (nog) beperkt

AI wordt vooral waargenomen in de In-fase. Reconnaissance en verkennende activiteiten, genereren van malware, opzetten van infrastructuur en onderzoek naar kwetsbaarheden zijn taken waarin AI-modellen zeer snel en effectief resultaten kunnen bereiken.

De omvang qua inzet van AI tijdens de Through- en Out-fase is in de analyse van het NCSC aanzienlijk lager dan tijdens de In-fase. Dat de In-fase een sterke vertegenwoordiging heeft kan deels verklaard worden dat iedere aanval begint in de In-fase. Een aanval kan immers niet starten in de Through- of Out-fase. Onderzoek van Anthrophic dat malafide gebruik van Claude in kaart brengt met een heatmap op de Mitre ATT&CK Navigator, onderschrijft dit beeld.[28]

De concentratie van voorbeelden van AI-gebruik in de In-fase betekent niet automatisch dat aanvallers vooral hier AI inzetten. Dat voor de In-fase desondanks toch het meest wordt waargenomen, kan een aantal redenen hebben:

- Aanvallen beginnen met verkennende (*reconnaissance*) activiteiten, er is infrastructuur nodig en een aanvaller moet middelen zoals malware of phishing-berichten ontwikkelen. Bovendien is lang niet iedere aanval succesvol en worden de Through- en Out-fase in dat geval helemaal niet doorlopen.
- Succesvolle en volledige detectie van activiteit in de Through- en Out-fase valt samen met voldoende logging en monitoring. Wanneer die niet afdoende is, worden incidenten misschien niet opgemerkt of wordt forensische analyse bemoeilijkt om een

aanvalsverloop volledig in kaart te brengen. Het op orde brengen van logging en monitoring wordt dus urgenter.

- Tot slot is het belangrijk om nogmaals te benadrukken dat het niet altijd mogelijk is om te identificeren of AI is gebruikt wanneer er geen toegang is tot infrastructuur van de actor of telemetrie vanuit een AI-leverancier van gebruikte AI-modellen. Hierdoor kan het aantal incidenten waarbij AI is gebruikt in de Through- en Out-fase in werkelijkheid hoger liggen. Het NCSC heeft meerdere incidentrapporten geanalyseerd waarin claims van AI-gebruik zeer beperkt onderbouwd werden en daardoor niet te verifiëren waren. Deze casussen zijn buiten beschouwing gelaten.

Plotten van incidenten op mate van autonomie

De inzet van AI kent verschillende gradaties van autonomie, waarbij aan de ene kant van het spectrum AI niet of beperkt wordt ingezet waarbij de mens volledige controle heeft. Aan de andere kant van het spectrum opereert AI autonoom en stelt de mens vooraf alleen een doel op en heeft verder geen controle. Het NCSC hanteert hierbij 6 categorieën die gebaseerd zijn op een aantal bestaande frameworks (voor een uitgebreide toelichting en beschrijving van deze taxonomie, zie *Appendix D: Taxonomie*):

- AI-0 – Geen inzet van AI
- AI-1 – AI-gegenereerd
- AI-2 – AI-versterkt
- AI-3 – AI-gecontroleerd
- AI-4 – AI-geregisseerd
- AI-5 – AI-autonoom

Uit de analyse van incidenten waarbij AI (vermoedelijk) is ingezet en rapportages die AI-gebruik beschrijven, ziet het NCSC voornamelijk AI-gebruik in de eerste drie categorieën (AI-1, AI-2 en AI-3). Een aantal AI-leveranciers rapporteerden over incidenten tijdens capaciteitentests voor hun AI-modellen.[29,30] Deze incidenten vallen in categorie AI-4 en AI-5.

- Het NCSC schat dat incidenten bij Nederlandse organisaties waarbij AI wordt ingezet primair binnen de categorieën A-1 t/m A-3 vallen, waarbij AI dus vooral als ondersteuning wordt ingezet of taken geautomatiseerd worden uitgevoerd. AI dient hier dan als force multiplier van bestaande capaciteiten en activiteiten.

Incidenten in de categorie AI-0, waarbij geen AI wordt gebruikt, vinden ook nog veelvuldig plaats. Ook dit soort incidenten kunnen zeer grote consequenties hebben. De basisprincipes van cybersecurity moeten hoe dan ook op orde zijn, ongeacht of AI wel of niet door aanvallers gebruikt wordt.

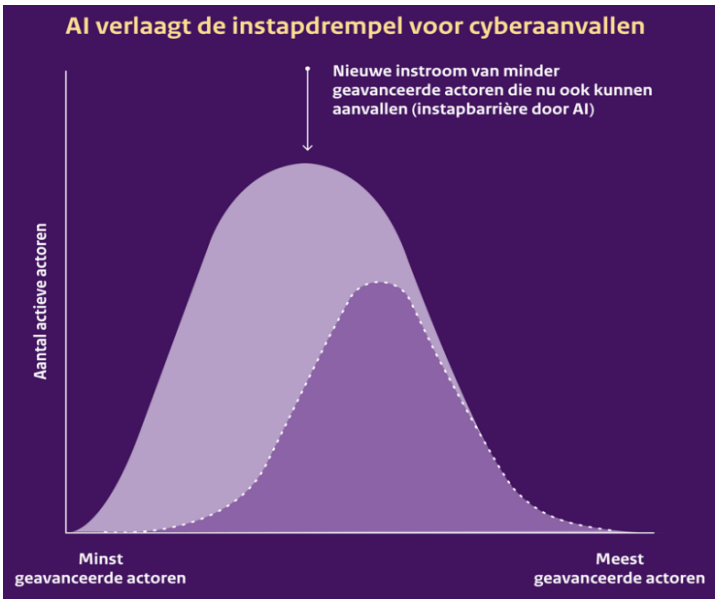
Dreigingslandschap verandert tot op zekere hoogte

Op dit moment heeft AI de capaciteiten om onderdelen van de aanvalsketen te versnellen en te automatiseren. Veel van de genoemde voorbeelden maken nog steeds gebruik van beproefde tools en technieken. Wel zijn er een aantal veranderingen op te merken:

- Met behulp van AI hoeven aanvallers niet langer iedere stap handmatig eerst te evalueren, maar kan een AI ook tussentijdse stappen evalueren en zelf beslissingen maken. Hierdoor kan de totale tijd van begin tot eind van een aanval aanzienlijk korter zijn.
- Ook kan een AI-agent doorwerken terwijl de menselijke operator niet beschikbaar is. Dit maakt een aanval voor de aanvaller sneller en efficiënter.
- Daarentegen zijn essentiële beveiligingsmaatregelen vaak niet op orde bij slachtoffers van dergelijke aanvallen. Denk hierbij aan API-keys en secrets die publiek benaderbaar zijn, gebruik van slechte wachtwoorden, misconfiguraties, niet-gepatchte kwetsbaarheden. De voorbeelden laten zien dat AI goed is in het snel vinden van dergelijke beveiligingsrisico's om ze uit te buiten waarbij een geavanceerde aanval niet nodig is.
- Geavanceerde actoren verkennen de mogelijkheden van AI en gebruiken het waar mogelijk, maar zijn er niet van afhankelijk om succesvol te zijn. Ze hebben nu al de technische expertise en capaciteiten om complexe aanvallen uit te voeren en hun doelen te bereiken. Desondanks schat het NCSC in dat ook deze geavanceerde actoren aanvallen sneller en wendbaarder maken en hun capaciteiten uitbreiden met behulp van AI wanneer dit toegevoegde waarde heeft.

Dergelijke ontwikkelingen maken AI dus vooral een force multiplier. Het versterkt en vergroot capaciteiten van kwaadwillenden. AI kan immers taken uit handen nemen, troubleshooten, en advies geven over te nemen strategieën. Kortom, kwaadwillenden hoeven AI niet per se in te zetten om betere malware of exploits te bouwen. Uitvoer- en doorlooptijden versnellen of verbeteren kan aanvallen al succesvoller maken.

- De beschikbare kracht en capaciteiten van AI betekenen niet automatisch dat mensen zonder technische kennis de meest geavanceerde aanvallen kunnen uitvoeren. Zo kan een kwaadwillende wellicht makkelijk malware genereren, maar is het daadwerkelijk inzetten hiervan een stuk complexer. Desalniettemin, zullen digitale aanvallen over de gehele linie wel meer geavanceerd worden en daarmee ook moeilijker te detecteren worden. Dit is schematisch weergegeven in *afbeelding 1*.



Afbeelding 1: toename capaciteiten van aanvallers

- Het NCSC acht het zeer waarschijnlijk dat naarmate AI-modellen meer beschikbaar en toegankelijk worden, de inzet van AI bij aanvallen ook zal toenemen.

De impact van AI op detectie van aanvallen

Naast de algemene duiding op het dreigingslandschap zijn er ook enkele specifieke aandachtpunten bij het toenemende gebruik van AI door kwaadwillenden.
AI kan (tijdige) detectie hinderen omdat aanvallers zich beter en sneller aankunnen aanpassen aan de context van het slachtoffernetwerk.

- AI maakt ook analyse, verwerking en verder misbruik van buitgemaakte data efficiënt. Detectie moet zich meer inzetten op waarnemen van afwijkend gedrag ten opzichte van vooropgestelde baselines en zal de ontwikkelingen die AI teweegbrengt moeten incorporeren.
- AI kan aanvallers ondersteunen bij (nieuwe) obfuscatietechnieken en polymorfe malware[v] om zoveel mogelijk ongezien te opereren. Doordat code gegenereerd kan worden door AI en dynamisch aangepast kan worden tijdens uitvoering kan deze ontwikkeling in een verdere stroomversnelling komen.
    - Dit blijven echter wel zeer specifieke use-cases en met name polymorfie malware wordt lang niet altijd ingezet. Desondanks stelt AI aanvallers wel in staat om dergelijke tools met minder inspanning te ontwikkelen
- Een aanval kan ook mislukken door hallucinaties of *guardrails* van AI-modellen. Ook kan het gebruik van een AI-model binnen een slachtoffernetwerk voor veel ruis (en dus bijbehorende logs en artefacten die gegenereerd worden) zorgen wat voor verdedigers kansen biedt om een aanval te detecteren. Zonder gebruik van AI is de kans op vergelijkbare ruis kleiner.

Ontwikkelingen gaan snel: aandacht vereist

De analyse in dit rapport is gebaseerd op waarnemingen in de afgelopen maanden. Ontwikkelingen op het gebied van AI gaan vooralsnog snel waarbij iedere paar maanden er nieuwe modellen beschikbaar komen. Frontier AI-modellen lopen voorop maar open-source en open-weight[vi] modellen lopen gemiddeld maar zes maanden achter en ook met non-frontiermodellen kunnen kwaadwillenden impact bereiken. Een te grote focus op frontier modellen kan ertoe leiden dat dreigingen gemist worden.

- Dit betekent dat verwachtingen in dit rapport een beperkte houdbaarheid hebben en over 6 tot 12 maanden opnieuw geëvalueerd zullen moeten worden.
- Hoewel er op dit moment nog niet eerder beschreven aanvalsmethodes zijn geobserveerd, kan dit met de komst van nieuwe AI-modellen of betere adaptatie door

[v] Polymorfe malware heeft functionaliteiten om onderdelen van de code aan te passen terwijl de kernfunctionaliteit behouden blijft. Dit bemoeilijkt detectie omdat de malware er steeds net wat anders uitziet waardoor herkenning op basis van voorop vastgestelde eigenschappen niet mogelijk is.

[vi] Frontiermodellen zijn de meest geavanceerde modellen die tegelijkertijd volledig gesloten zijn. Een gebruiker kan de broncode of parameters en gewichten waarop de AI besluiten maakt niet inzien. Bij open-source modellen is zowel de broncode als trainingsdata en parameters openbaar. Open-weight modellen zitten hier tussenin: de parameters kan je downloaden en lokaal gebruiken, maar licentie en broncode blijven eigendom van de leverancier.

geavanceerde actoren veranderen. Het NCSC blijft informeren wanneer er nieuwe ontwikkelingen zijn.

De snelheid van AI-ontwikkelingen vereisen dat organisaties alert blijven en aanpassen als dat nodig is. AI is inmiddels niet meer te negeren en zal in toenemende mate een rol gaan spelen binnen onze samenleving. Daarentegen betekent niet ieder nieuw AI-model automatisch nieuwe aanvalsmethodes, maar zoals de voorbeelden ook meermaals doen uitwijzen, zijn die voor een succesvolle compromittatie vaak ook nog helemaal niet nodig.

4 Handelingsperspectief

De ontwikkelingen rondom AI staan niet los van bestaande beveiligingsmaatregelen en de basisprincipes. Deze blijven ook in het huidige dreigingslandschap essentieel. Organisaties die AI beschouwen als een afzonderlijk of nieuw technisch vraagstuk lopen het risico de kern van het probleem te missen.

De ontwikkelingen op het gebied van AI vragen wel om een herijking van de digitale weerbaarheid en bestaande beveiligingsmaatregelen. In de praktijk betekent dit dat jij bestaande principes in het licht van kortere reactietijden en moeilijke detectie moet herevalueren en aanscherpen waar nodig.

Verklein je aanvalsoppervlak

Om jouw organisatie goed te beschermen tegen digitale dreigingen moet je weten in welke mate jouw netwerken en systemen blootgesteld zijn aan dreigingen van buitenaf. De inzet van AI door kwaadwillenden zorgt ervoor dat het verkleinen van het aanvalsoppervlak prioriteit moet krijgen. Met exposuremanagement breng je de zichtbaarheid van jouw digitale assets in kaart én breng je de weerbaarheid ervan naar een passend niveau. Meer informatie over exposuremanagement en het verkleinen van je aanvalsoppervlak is te vinden in de NCSC-factsheet Hoe richt je exposuremanagement in?

Verbeter je kwetsbaarhedenbeheer

Kwetsbaarhedenbeheer (*patchmanagement*) is een cyclisch proces waarin kwetsbaarheden en de weerbaarheid bevorderende maatregelen worden geanalyseerd en beoordeeld. We delen dit proces op in vijf fases: beoordelen, prioriteren, behandelen, evalueren, verbeteren.

Het doel van kwetsbaarhedenbeheer is om de blootstelling aan risico's van de organisatie te verkleinen. Ontwikkelingen op het gebied van AI zorgen ervoor dat kwetsbaarheden sneller gevonden en misbruikt kunnen worden. Dit vereist dat daarom dat je kwetsbaarhedenbeheer op een passend niveau is ten opzichte van de huidige realiteit. Meer informatie over kwetsbaarhedenbeheer is te vinden in de NCSC-factsheet Verbeter je kwetsbaarhedenbeheer.

Identiteitsbeheer: Implementeer authenticatiebeleid en Least Privilege

Beperk gebruikers- en applicatierechten op tot het hoogstnoodzakelijke niveau voor de uitvoering van werkzaamheden. Laat eindgebruikers applicaties niet met administratieve rechten uitvoeren. Zelfs als een aanvaller toegang krijgt tot een gebruikersaccount, dan kan die geen processen manipuleren zonder deze administratieve rechten. Maak gebruik van sterke wachtwoorden en verplicht het gebruik van een wachtwoordmanager. Implementeer vervolgens multifactorauthenticatie (MFA) voor alle toegang, met een nadruk op beheerderstoegang en toegang op afstand (zoals VPN en Remote Desktop Protocol (RDP)).

Voorkom dat een aanvaller vrij door het netwerk kan bewegen door endpoints en systemen logisch te segmenteren. Hanteer het uitgangspunt dat netwerkverkeer tussen netwerken in het algemeen niet toegestaan is en voeg vervolgens specifieke firewallregels toe voor verkeer dat wel nodig is.

- Meer informatie over multifactorauthenticatie is te vinden in de NCSC-factsheet Gebruik tweefactorauthenticatie.
- Meer informatie over het organiseren van toegang is te vinden in de NCSC-factsheet Hoe organiseer ik identiteit en toegang.

Conditional Access

Conditional Access betekent dat authenticatie of toegang afhankelijk wordt gemaakt van bepaalde condities (voorwaarden), zoals:

- Locatie, bijvoorbeeld alleen toegang vanuit Nederland.

- Apparaat, bijvoorbeeld alleen vanaf een specifiek apparaat.

- Tijdstip, bijvoorbeeld alleen toegang tijdens kantooruren.

- Gebruiker of groep, bijvoorbeeld managers krijgen andere toegang dan normale gebruikers.

Certificate-based authenticatie

Een vertrouwde derde partij, doorgaans een Certificate Authority (CA), bevestigt door middel van een onderzoek de identiteit van een gebruiker. Vervolgens verstrekt de CA een digitaal certificaat waarin de CA de identiteit van de gebruiker bevestigt. Dit certificaat wordt vervolgens door de gebruiker als bewijsstuk gebruikt om toegang te krijgen.

Passkeys op basis van FIDO2

Dit is een open standaard van de FIDO Alliance. FIDO2 gebruikt asymmetrische cryptografie en biedt een hoge mate van beveiliging. Dit komt omdat de inloggegevens uniek zijn voor elk systeem en enkel daar te gebruiken zijn. Inloggegevens kunnen niet worden onderschept op een valse website en hergebruik van de inloggegevens is onmogelijk. De privésleutel en eventuele biometrische gegevens verlaten het geregistreerde apparaat niet en worden niet opgeslagen op een server. Privésleutels kunnen alleen individueel uitlekken, dit maakt mogelijke aanvallen minder schaalbaar. Bovendien, omdat bij een datalek alleen publieke kunnen sleutels uitlekken is er geen risico op misbruik. Tot slot is de opgeslagen publieke sleutel niet traceerbaar.

Identificeer en herstel misconfiguraties

De basis van weerbaarheid ligt in het voorkomen van onveilige configuraties voordat deze kunnen worden misbruikt. Stel een actueel overzicht op van alle platformen, cloudomgevingen en applicaties en breng configuratie-instellingen en toegangsrechten in kaart. Pas het principe van 'least privilege' toe. Scheid beheerders- en gebruikersaccounts, schakel anonieme toegang uit wanneer het niet noodzakelijk is en besteed specifiek aandacht aan gastaccounts en rechten van applicaties. Verplicht multifactor authenticatie (MFA) voor alle beheerdersaccounts. Pas onveilige standaardconfiguraties aan voordat een systeem in gebruik wordt genomen. Verschillende voorbeelden laten zien hoe AI-gestuurde aanvallen niet succesvol zijn doordat systemen veilig geconfigureerd zijn. Voor meer informatie over misconfiguraties, zie de NCSC-website.

Versterk logging en monitoring

Grootschalig AI-gebruik door aanvallers kan detectie moeilijker maken. Een aanval kan bijvoorbeeld meer lijken op het gedrag van een medewerker. Een aanvaller kan geloofwaardige accounts of profielen maken of normale bedrijfsprocessen nabootsen. Met AI kunnen ook snel nieuwe middelen worden ontwikkeld. Dat kan de detectie van nieuwe en onbekende malware lastiger maken. Er zijn daarnaast aanwijzingen dat aanvallers sneller kunnen handelen en hun werkwijze sneller kunnen aanpassen aan het netwerk van het slachtoffer.

Beoordeel je huidige detectiemogelijkheden, logging en monitoring opnieuw. Traditionele statische Indicators of Compromise (IoC's), vaste signatures en eenvoudige volumeregels kunnen daardoor minder betrouwbaar worden. Je kunt de detectie bijvoorbeeld op de volgende manieren aanpassen:

- Let op snelheid en tempo. Kijk hoe snel acties elkaar opvolgen, hoeveel systemen kort na elkaar worden benaderd, hoe snel een aanvalstechniek of gedrag na een mislukte poging verandert en hoeveel acties per minuut plaatsvinden.
- Stel identiteit centraal, zoals gebruiker, apparaat en sessie, in plaats van alleen IP-adressen. Zo wordt het moeilijker om detectie te ontwijken door van IP-adres of hulpmiddel te veranderen.
- Leg meer context rond acties vast en bekijk opeenvolgende acties in samenhang. Registreer wie een proces startte, vanuit welk proces en op welk systeem. Leg ook vast welke netwerkverbindingen ontstonden. Een losse actie is misschien niet verdacht, terwijl de reeks acties dat wel kan zijn.
- Ook de samenhang tussen bestandstype en -grootte kan een indicatie zijn voor malafide activiteiten. Denk bijvoorbeeld aan een klein .ISO-bestand of juist een groot JPEG-bestand die niet passen binnen het normale netwerkverkeer.
- Monitor intern beschikbare AI-endpoints op ongebruikelijke verzoeken en acties om aanvallen, die interne AI-modellen en systemen misbruiken, te detecteren.

Controleer ook de responstijd op signalen. Beoordeel of je die voor de belangrijkste gevallen kunt verkorten. Lees meer over het goed inrichten van logging en monitoring.

5 Appendix A: Wat is AI?

AI verwijst naar systemen die op basis van informatie uit de omgeving of input(data) met een zekere mate van zelfstandigheid een actie kunnen ondernemen of antwoord kunnen aandragen. AI werkt met algoritmen die in software beschreven zijn.

Een algoritme kun je zien als een recept; het is een set aan instructies voor het bereiken van een bepaald doel. Deze algoritmen delen we in drie verschillende groepen op:

1. Expertsysteem / niet-zelflerende AI

Dit zijn technieken waarin machines voorgeprogrammeerd zijn door mensen, om menselijke beslissingen na te bootsen. Deze systemen zijn gebaseerd op vast geprogrammeerde regels die door mensen ontworpen zijn. In de praktijk betekent dit dat het meestal een systeem met een heleboel if-statements is. Een voorbeeld van een expertsysteem is een tool voor het uitvoeren van steekproeven bij subsidies, of denk aan een tool om te berekenen wat je maximale hypotheek kan zijn.

2. Machine Learning

Dit zijn algoritmen die patronen uit een dataset halen. Soms zijn deze patronen voor de mens bijna niet zichtbaar of onmogelijk te achterhalen door de grote hoeveelheid data. Dit levert modellen op die (door de selectie van de trainingsdataset) op specifieke use-cases van toepassing zijn. Een voorbeeld van Machine Learning is gezichtsherkenning of een prijsberekeningsalgoritme.

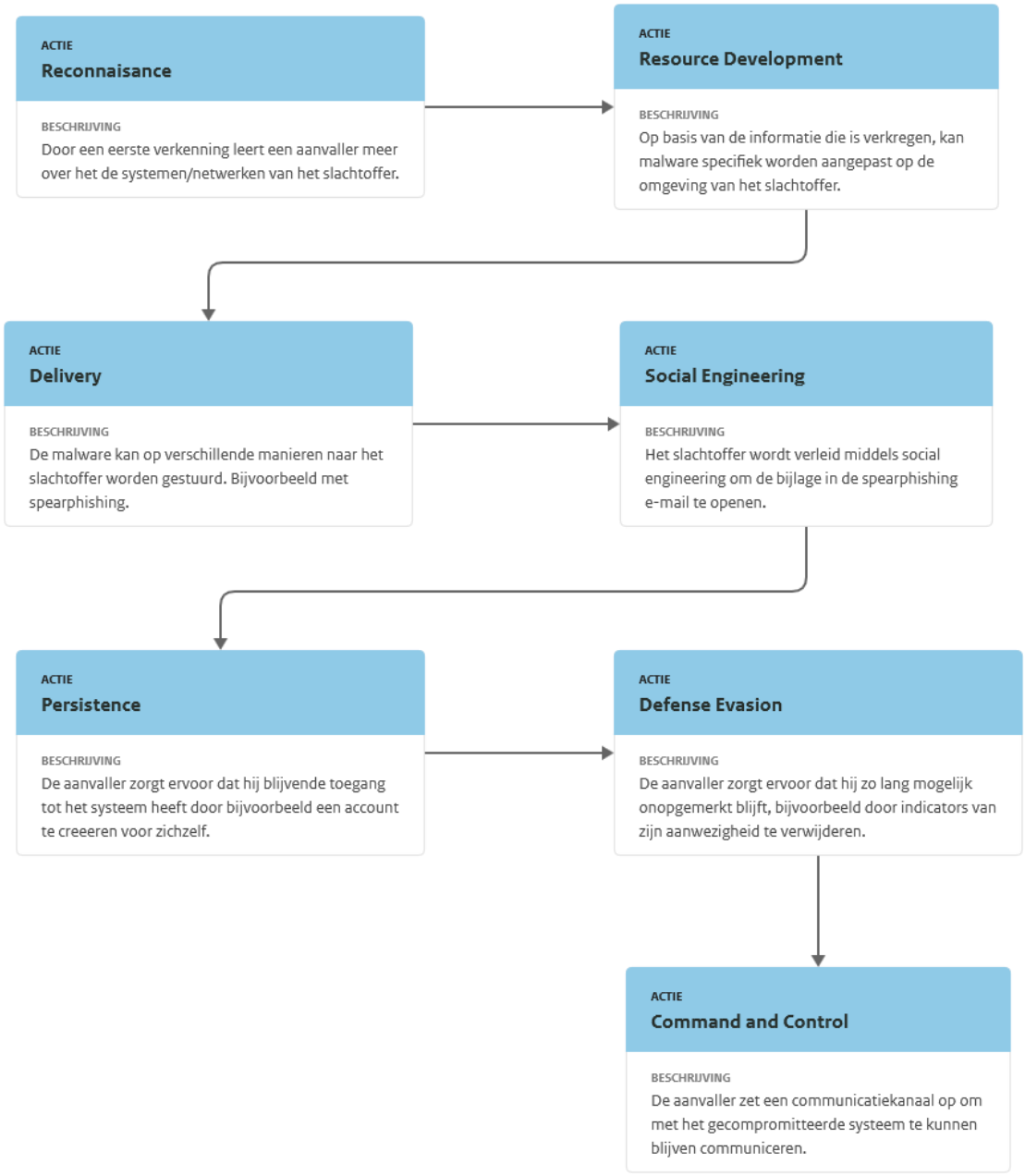
3. Generatieve AI

Met Generatieve AI kan onder andere nieuwe content in de vorm van data - zoals audio, tekst, video en afbeeldingen - worden geproduceerd. De meest bekende generatieve AI-modellen zijn op dit moment taalmodellen (Large Language Models, LLM's). Deze algoritmen zijn een evolutie van de Machine Learning-algoritmen. Ze zijn getraind op een enorme hoeveelheid zeer diverse data. Met deze data maken de algoritmen een kansberekening om te bepalen wat het beste volgende woord is. Hoe meer data de computer krijgt, hoe beter de computer leert om patronen te herkennen én om zelf soortgelijke data te genereren.[31]

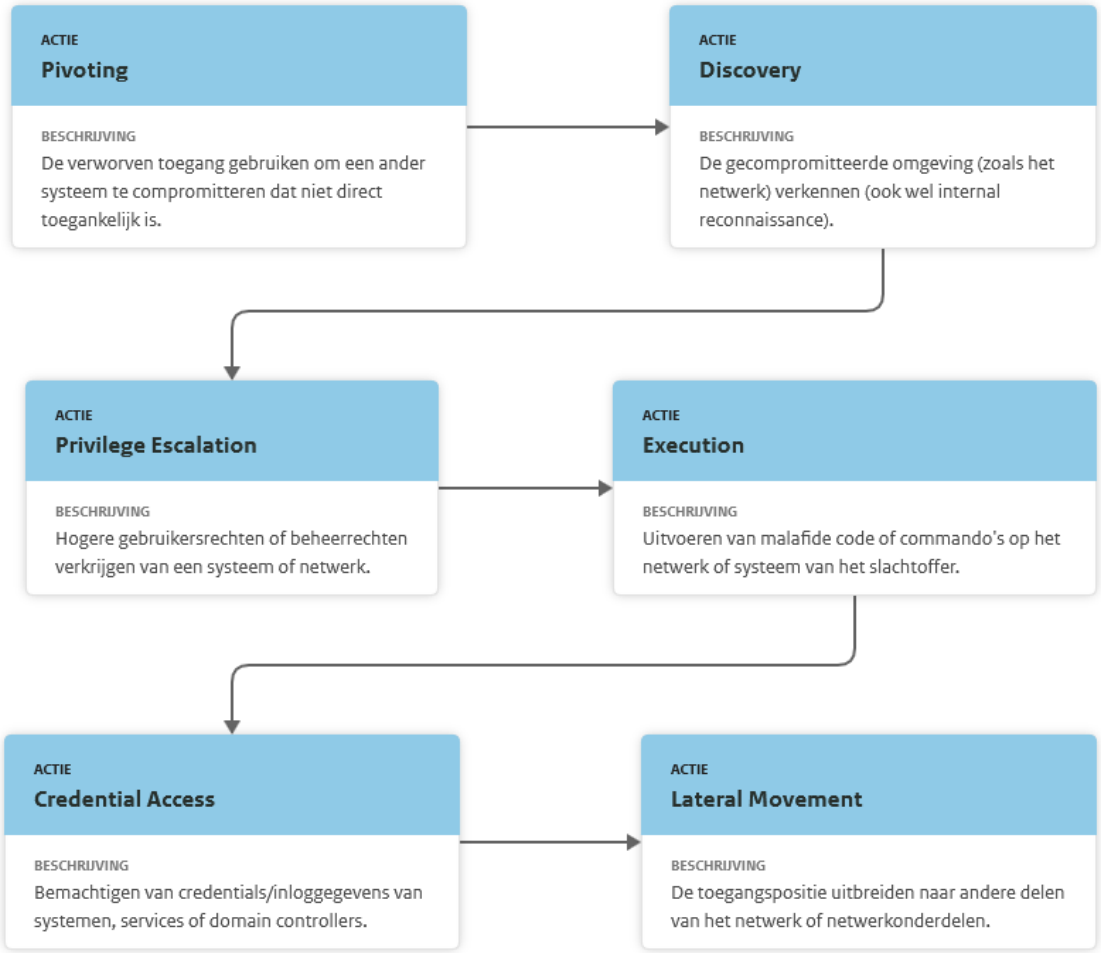
3a. Agentic AI

Agentic AI is een subcategorie van Generatieve AI. Het is een meer geavanceerd type waarin de AI zelfstandig acties plant en uitvoert. De gebruiker geeft een doel, waarna de AI tussenstappen bepaalt, informatie verzamelt, digitale hulpmiddelen gebruikt en de aanpak aanpast op basis van de resultaten. De AI voert dit doel uit met behulp van een set aan AI agents die gecoördineerd acties uitvoeren. Dit maakt Agentic AI geschikt voor complexe taken, maar vergroot ook het risico op ongewenste acties wanneer de opdracht onduidelijk is, de AI verkeerde conclusies trekt of te veel rechten heeft.

6 Appendix B: Visualisatie In-fase



7 Appendix C: Visualisatie Through-fase



8 Appendix D: Taxonomie

AI kan in verschillende gradaties worden toegepast, en ook aanvallers gebruiken en misbruiken AI-modellen in verschillende gradaties. In eenvoudige gevallen betekent dit dat een aanvaller een AI-model inzet als een meer capabele zoekmachine. Echter zijn er inmiddels ook voorbeelden waarbij een aanvaller met behulp van AI autonoom (onderdelen van) een aanval uitvoert. Daartussenin bevinden zich nog een aantal gradaties. Om incidenten waarbij gebruik wordt gemaakt van AI te labelen, maar ook om de aard van incidenten inzichtelijk te maken, volgt hieronder een taxonomie met daarin een aantal gradaties hoe AI is ingezet bij cyberaanvallen.

Deze taxonomie geeft op tactisch niveau extra uitleg over incidenten. Dit biedt de mogelijkheid om onderscheid aan te brengen in de mate van geavanceerdheid van de toepassing van AI. De taxonomie is gebaseerd op elementen uit bestaande taxonomieën van Sophos[32], OWASP[33], CSA[34], NVIDIA[35], AWS[36], Anthrophic[28] en Mitre[37].

Gradatie	Classificatie	Uitleg	Rol van de mens
AI-0	Geen AI gebruikt	Geen AI inzet.	Mens volledig in control
AI-1	AI-gegenereerd	Mens geeft opdrachten. AI levert output/ artefact.	Human-in-the-loop
AI-2	AI-versterkt	AI versterkt operaties en helpt mens bij maken van beslissingen.	Human-in-the-loop
AI-3	AI-gecontroleerd	AI voert meerdere taken binnen door mens vooraf opgestelde opdracht. Mens beslist over zaken buiten scope.	Human-on-the-loop
AI-4	AI-geregisseerd	Mens bepaalt het doel en behoudt strategisch overzicht. AI coördineert meerdere fases en vraagt alleen op kritieke punten om goedkeuring.	Human-on-the-loop
AI-5	AI-autonoom	AI initieert, past zich aan en verandert doelen om zo missie te behalen. Mens speelt geen rol bij maken van beslissingen tot eind output.	Human-out-of-the-loop

9 Referenties

[1] https://www.unifiedkillchain.com/

[2] https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access

[3] https://hunt.io/blog/chinese-operators-claude-deepseek-government-intrusion

[4] https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

[5] https://unit42.paloaltonetworks.com/autonomous-ai-cyber-attack-campaign/

[6] https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

[7] https://gambit.security/blog-posts/a-single-operator-two-ai-platforms-nine-government-agencies-the-full-technical-report

[8] https://blog.talosintelligence.com/keep-going-bro-youve-got-this-a-data-driven-look-at-how-adversaries-are-weaponizing-ai/

[9] https://cert.pl/uploads/docs/CERT_Polska_Energy_Sector_Incident_Report_2025.pdf

[10] https://cloud.google.com/blog/topics/threat-intelligence/distinct-clusters-target-individuals-of-interest-to-russia

[11] https://www.microsoft.com/en-us/security/blog/2026/04/06/ai-enabled-device-code-phishing-campaign-april-2026/

[12] https://www.sekoia.com/blog/eviltokens-an-ai-augmented-phishing-as-a-service-for-automating-bec-fraud-part-2

[13] https://blog.talosintelligence.com/uat-10147-chinese-speaking-adversary-integrates-agentic-ai-into-post-compromise-operations/

[14] https://www.trendmicro.com/en_us/research/26/e/vibe-hacking-two-ai-augmented-campaigns-target-government-and-financial-sectors-in-latin-america.html

[15] https://hunt.io/blog/thailand-ministry-finance-targeted-with-hermes-ai-agent#PostExploitation_Tools

[16] https://cert.gov.ua/article/6284730

[17] https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools

[18] https://www.stepsecurity.io/blog/supply-chain-security-alert-popular-nx-build-system-package-compromised-with-data-stealing-malware

[19] https://www.wiz.io/blog/s1ngularity-supply-chain-attack

[20] https://www.eset.com/nl/over/newsroom/persberichten-overzicht/persberichten/eset-ontdekt-promptlock-de-eerste-door-ai-aangedreven-ransomware/

[21] https://unit42.paloaltonetworks.com/ai-use-in-malware/#post-175752-_vrgnq2kwajin

[22] https://www.sysdig.com/blog/ai-assisted-cloud-intrusion-achieves-admin-access-in-8-minutes

[23] https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf

[24] https://unit42.paloaltonetworks.com/flutterbridge-new-fluttershell-backdoor/

[25] https://snyk.io/blog/weaponizing-ai-coding-agents-for-malware-in-the-nx-malicious-package/

[26] https://nx.dev/blog/s1ngularity-postmortem

[27] https://www.microsoft.com/en-us/security/blog/2026/03/06/ai-as-tradecraft-how-threat-actors-operationalize-ai/

[28] https://www.anthropic.com/research/attack-navigator

[29] https://openai.com/index/hugging-face-model-evaluation-security-incident/

[30] https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals

[31] https://www.autoriteitpersoonsgegevens.nl/themas/algoritmes-ai/algoritmes-uitgelegd/generatieve-ai

[32] https://www.sophos.com/en-us/content/sophos-ai-security-2026-report

[33] https://owasp.org/APTS/standard/4_Graduated_Autonomy/

[34] https://cloudsecurityalliance.org/blog/2026/01/28/levels-of-autonomy

[35] https://developer.nvidia.com/blog/agentic-autonomy-levels-and-security/

[36] https://aws.amazon.com/blogs/security/the-agentic-ai-security-scoping-matrix-a-framework-for-securing-autonomous-ai-systems/

[37] https://atlas.mitre.org/

Uitgave
Nationaal Cyber Security Centrum (NCSC)
Postbus 117, 2501 CC Den Haag
Turfmarkt 147, 2511 DP Den Haag
070 751 5555

Meer informatie
www.ncsc.nl
info@ncsc.nl