Nationaal Cyber Security Centrum
Ministerie van Justitie en Veiligheid

AI-gestuurde frameworks voor offensieve doeleinden

NCSC CTI-report

Publicatiedatum: 27-11-2026

Toegestane verspreiding: TLP:CLEAR
(Traffic Light Protocol)

Dit CTI-report bevat het label TLP:CLEAR en wordt door het Nationaal Cyber Security Center (NCSC) verspreid. Het NCSC gebruikt het Traffic Light Protocol (TLP) om eenduidig te definiëren wat er met de informatie mag gebeuren. Wanneer informatie is voorzien van een TLP-aanduiding weet je met wie je deze informatie mag delen. Dit staat beschreven in de standaard van First (www.first.org/tlp).

Reacties zijn welkom op info@ncsc.nl.

Inhoudsopgave

1 Essentie 3
2 Aanleiding 4
3 Feiten 5
Noemenswaardige incidenten waarbij AI-gestuurde frameworks zijn gebruikt 8
4 Technisch onderzoek werking ARTEX 10
5 Duiding 12
6 Handelingsperspectief 15
Preventie 15
Detectie 16
Response 18
7 Appendix A: Taxonomie 19
8 Appendix B: ARTEX-artifacten 20
9 Bronnen en referenties 21

1 Essentie

- **AI-gestuurde frameworks voor offensieve doeleinden** zoals ARTEX zijn vaak ontwikkeld voor (ethisch) securityonderzoek of red teaming, maar worden gretig overgenomen door kwaadwillenden. Veel zijn open source, publiek beschikbaar en in constante ontwikkeling.

- Recent zijn er **meerdere noemenswaardige incidenten**: o.a. de DIVD, Zuid-Koreaanse financiële sector, EU-organisaties, webwinkels, telecomproviders en overheden. ARTEX komt vaak voor, maar andere frameworks worden ook ingezet.

- **Het NCSC acht het zeer waarschijnlijk dat de inzet en impact van AI-gestuurde frameworks bij digitale aanvallen zal toenemen** en dat dit de urgentie voor implementatie van adequate beveiligingsmaatregelen verder onderstreept. Recent hebben verschillende (Europese) overheidsorganisaties en CERTs vergelijkbare waarschuwingen gepubliceerd.

- **De grootste impact op het dreigingslandschap is snelheid en schaalbaarheid**: het framework fungeert als orchestrator die AI-modellen, tools en sub-agents coördineert. Taken worden parallel of non-stop uitgevoerd met beperkte menselijke interactie. Tussentijds kan een menselijke operator plannen bijstellen.
    - Aanvallen vinden zo snel plaats dat verdedigers minder of geen tijd hebben om in te grijpen. Door de schaalbaarheid kunnen verdedigers overbelast raken.

- De AI-gestuurde frameworks **werken non-stop door en hebben een groot aanpassingsvermogen**. Waar trial-en-error en ervaring opdoen voor mensen een tijdrovende klus is, kan dit met behulp van AI zeer snel.

- Bestaande beveiligingsmaatregelen blijven effectief, maar AI legt **technische achterstand sneller bloot** en **vergroot de urgentie om maatregelen te nemen**. Prioriteiten liggen op het gebied van preventie en detectie. Het hoofdstuk handelingsperspectief geeft hiervoor verschillende adviezen.

- **Er zijn verschillende typen frameworks**: van uitvoer van een volledige cyber kill chain tot specialisaties zoals kwetsbaarhedenonderzoek. ARTEX is prominent, maar alternatieven zijn CyberStrikeAI, Cairn, Strix, Hermes, HexStrike, PentestAGI, Agent Zero en recent AntiFox

- Technisch onderzoek van het NCSC naar ARTEX toont de **snelheid en lage kosten**: bij een aanval op de testomgeving *Damn Vulnerable Web Application* (DVWA) was authenticatie-omzeiling en defacement binnen ongeveer 1 minuut mogelijk voor minder dan €0,20; bij een kwetsbare GitLab CE versie duurde RCE en defacement ongeveer 35 minuten voor ongeveer €2.

- **Detecteerbaarheid is beïnvloedbaar**: zonder instructies zijn ARTEX en andere frameworks vaak herkenbaar en luidruchtig. Met eenvoudige OPSEC-instructies worden herkenbare artefacten aangepast en gaan AI-gestuurde frameworks voorzichtiger aan het werk, waardoor detectie lastiger wordt.

TLP:CLEAR

2 Aanleiding

Ontwikkelingen rondom *Artificial Intelligence* (AI) verschijnen dagelijks in het nieuws, maar de laatste weken wordt het nieuws vooral gedomineerd door berichtgeving over incidenten waarbij AI-gestuurde frameworks zijn gebruikt. Incidenten bij het Dutch Institute for Vulnerability Disclosure (DIVD), de Zuid-Koreaanse financiële sector, en bedrijven in Japan hadden een gemene deler: ARTEX, een AI-gestuurd framework met *agentic*[1] en autonome[2] capaciteiten dat kan worden ingezet voor offensieve en pentesting doeleinden.

ARTEX is echter niet het enige framework van dit soort. Inmiddels zijn er meerdere varianten van AI-gestuurde frameworks beschikbaar. Rapportages van andere incidenten wereldwijd beschrijven het gebruik van AI-gestuurde frameworks zoals Cairn, CyberStrikeAI, Hermes Agent, STRIX, HexStrike AI, PentestGPT. Dit is slechts een greep van de tientallen (publiek) beschikbare frameworks, die bovendien constant in ontwikkeling zijn.

In het CTI-report Artificial Intelligence in de Cyber Kill ChAIn[1] (d.d. 24 september 2026) beschreef het NCSC de inzet van AI door kwaadwillenden bij digitale aanvallen. Daarin werden frameworks met autonome capaciteiten al kort aangehaald. Sindsdien is de aandacht, adoptie en inzet van dit soort frameworks sterk toegenomen.

Dit CTI-report gaat dieper in op het concept van dergelijke AI-gestuurde frameworks voor offensieve doeleinden. Hoe werken ze en hoe verandert de inzet digitale aanvallen? Het NCSC heeft onderzoek gedaan naar de inzetbaarheid van onder andere ARTEX om zo de kortere doorlooptijden van aanvallen inzichtelijk te maken. Het handelingsperspectief biedt richting om jezelf beter te wapenen tegen deze verandering binnen het dreigingslandschap. Want hoewel de frameworks veel gebruik maken van traditionele aanvalsmethodes, worden kwetsbare plekken binnen jouw organisatie wel versneld blootgelegd wanneer een aanvaller een AI-gestuurd framework inzet bij een aanval.

Ontwikkeld voor onderzoeksdoeleinden, maar inzet veelal malafide

AI-gestuurde frameworks voor offensieve doeleinden zijn niet noodzakelijk ontworpen voor het uitvoeren van digitale aanvallen, maar vaak voor (ethisch) securityonderzoek en red teaming activiteiten. Desondanks adopteren kwaadwillenden dit soort frameworks en tools gretig. Dit is niet nieuw voor het AI-tijdperk en gebeurde in het verleden ook al bij pentesting frameworks zonder AI-capaciteiten zoals bijvoorbeeld Cobalt Strike en Metasploit. Bovendien zijn dit soort AI-gestuurde frameworks vaak vrij toegankelijk doordat ze publiek worden aangeboden via onder andere GitHub.

[1] Agentic AI is een geavanceerdere vorm van generatieve AI die zelfstandig acties plant en uitvoert. De AI voert dit doel uit met behulp van een set aan AI-agents die gecoördineerd acties uitvoeren.
[2] Autonome capaciteiten van AI betekent dat zelfstandig beslissingen kan nemen en acties kan uitvoeren, zonder menselijke input. In tegenstelling tot traditionele AI, die mensen nodig heeft om het te begeleiden, leert autonome AI van data, past zich aan nieuwe situaties aan en werkt onafhankelijk. Agentic AI en autonome capaciteiten zijn dus nauw verweven.

TLP:CLEAR

3 Feiten

Ook in het hedendaagse AI-tijdperk voeren kwaadwillenden digitale aanvallen uit en bepalen zelf doelwitten en doelstellingen. De AI-gestuurde frameworks die worden ingezet dienen daarbij als een instrument uit de gereedschapskist van de aanvaller.

Veel van de componenten van AI-gestuurde frameworks zijn bovendien niet nieuw en los van elkaar al langere tijd in gebruik. Hieronder vallen beschikbare LLM-modellen, maar ook traditionele (security) tooling zoals bijvoorbeeld nmap en curl. De significante ontwikkeling zit in de mate van autonomie waarmee de AI-gestuurde frameworks kunnen opereren, acties coördineren en problemen oplossen waarbij interactie met een menselijke operator maar zeer beperkt nodig is.

Zoals hierboven genoemd, zijn er inmiddels meerdere AI-gestuurde frameworks die worden ingezet bij digitale aanvallen. Omdat deze lijst niet uitputtend is en er continu nieuwe frameworks ontworpen en ingezet worden, beschrijven we in dit hoofdstuk op conceptueel niveau de werking van dergelijke frameworks en ondersteunen we dit met voorbeelden uit de praktijk.

Hoe werken frameworks die worden ingezet voor offensieve doeleinden?

Een aanval begint wanneer een kwaadwillende toegang tot beschikbare AI-modellen en tools instelt en daarna een opdracht geeft aan een gekozen framework. Deze AI-modellen kunnen lokaal draaien op hardware van de kwaadwillende of via een externe aanbieder *(inferenceprovider)* worden aangeroepen. AI-modellen van een externe aanbieder worden afgenomen via een abonnement of op basis van gebruik middels een API-koppeling.[3]

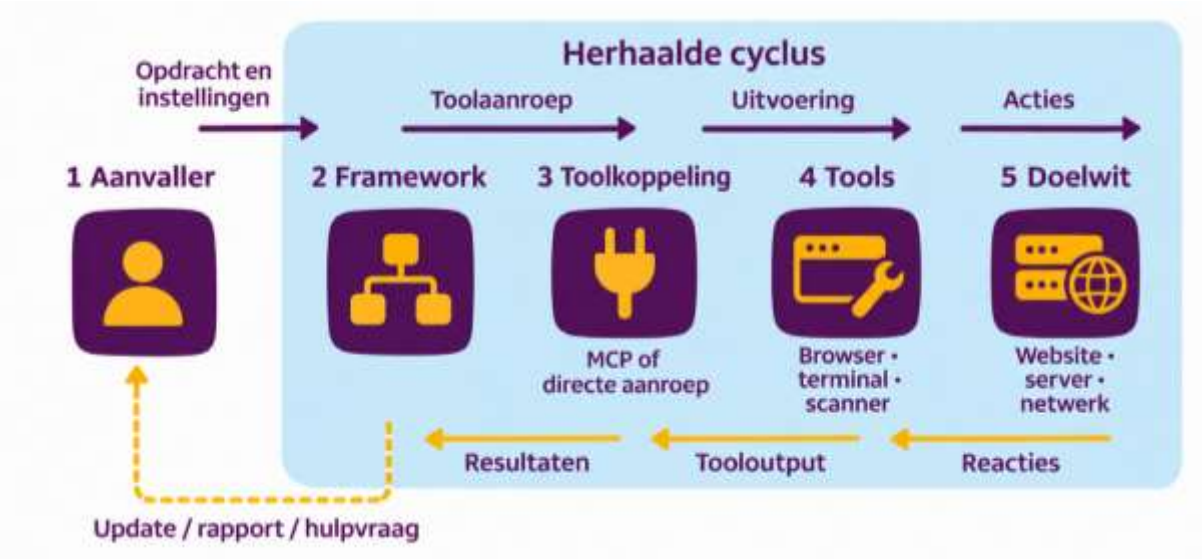
- Het gebruikte framework geeft een AI-model informatie over de opdracht, beschikbare tools en eerdere bevindingen. Het AI-model stelt vervolgens een actie voor. Het framework stuurt daarop een specifieke tool aan, zoals een browser of scanner. De koppeling met tools verloopt rechtstreeks of via een standaard zoals het Model Context Protocol (MCP)[4].
    - Een concreet voorbeeld is dat een menselijke operator de opdracht geeft om ergens binnen te dringen. Een AI-model formuleert dat een startdoel is om een webapplicatie te onderzoeken en dat daarvoor een scanner beschikbaar is. Het stelt voor om de applicatie te scannen.
    - Het framework start daarop de scanner met het juiste webadres. Vervolgens geeft het framework de scanresultaten terug aan het AI-model. Op basis daarvan kan het model bijvoorbeeld voorstellen om een gevonden inlogpagina verder te onderzoeken.

[3] Via een API-koppeling betaalt de gebruiker per token. Meer rondes, langere invoer en meerdere agents vergroten het gebruik en daarmee de kosten. Onder andere GLM 5.3 en DeepSeek 4.1 Flash en Pro zijn populair onder aanvallers. Deze modellen hebben beperkt guardrails en zijn in verhouding tot andere modellen goedkoop in het gebruik. Ook abliterated models zijn in trek. Bij deze modellen zijn de guardrails er volledig vanaf gehaald.

[4] MCP biedt een vaste manier waarop tools aan een AI-systeem laten weten wat ze kunnen, welke invoer ze nodig hebben en hoe het AI-systeem ze kan aansturen. Welke handelingen daadwerkelijk mogelijk zijn, hangt af van de tools en toegangsrechten die binnen het framework zijn ingesteld.

TLP:CLEAR

- Het framework verwerkt de resultaten van gebruikte tools en bewaart bevindingen voor volgende stappen. Het kan het plan bijstellen en meerdere iteraties doorlopen voordat het aan de gebruiker rapporteert. Sommige frameworks verdelen taken over meerdere sub-agents: losse AI-agents die elk een deeltaak uitvoeren (eventueel weer met elk hun eigen tools). Afbeelding 1 illustreert schematisch deze workflow.
    - GreyNoise beschrijft naar aanleiding van een waargenomen aanvalscampagne gericht op een kwetsbaarheid in PaperCut NG/MF hoe verschillende tools zoals MimiKatz, BloodHound, NetExec en Impacket door een AI-model worden aangeroepen en ingezet.[2]
- Deze coördinerende activiteiten en het aansturen van meerdere agents maakt dat AI-gestuurde frameworks voor offensieve doeleinden ook wel *agentic* (agentisch) genoemd kunnen worden.

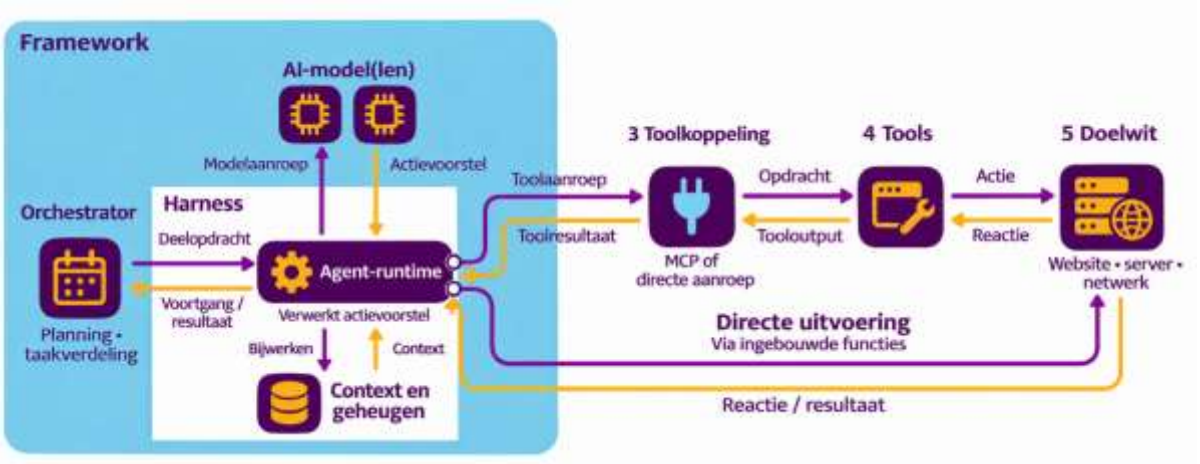


Afbeelding 1: Schematische weergave workflow AI-gestuurde frameworks

Hoe zelfstandig de uitvoer van taken verloopt, verschilt per framework en instelling. De gebruiker kan bijvoorbeeld vooraf toestemming geven voor bepaalde handelingen of tussentijds bijsturen. Ook kan de gebruiker bepalen dat alleen het eindresultaat wordt getoond en dat er moet worden doorgewerkt totdat dit resultaat succesvol is bereikt.

Samenvattend is het gekozen AI-gestuurde framework dus een agentische gereedschapskist waar binnen AI-modellen, tools en processen worden verbonden. Het framework fungeert hierbij ook als orchestrator: het bepaalt in welke volgorde deelopdrachten worden uitgevoerd, welke agents hiervoor nodig zijn en stuurt deze vervolgens aan. Dat dit allemaal soepel loopt en dat tools onderling functioneren en informatie kunnen uitwisselen is te danken aan het harness dat binnen het framework om de tools en AI-model(len) heen hangt. Dit harness bevat waar nodig specifieke skills voor specifieke taken. Afbeelding 2 geeft deze interne werking schematisch weer.

TLP:CLEAR



Afbeelding 2: Interne werking AI-gestuurde frameworks

Verschillende soorten AI-gestuurde frameworks

Niet ieder AI-gestuurd framework heeft dezelfde capaciteiten of functionaliteiten. Sommige frameworks zijn gespecialiseerd in specifieke onderzoekstaken en andere frameworks bieden een oplossing voor de gehele cyber kill chain. GitGuardian heeft 45 open source AI-gestuurde frameworks in kaart gebracht en onderverdeeld in vijf verschillende categorieën.[3] De categorieën maken onderscheid waarin ze gespecialiseerd zijn, waarvoor ze kunnen worden toegepast en welke hulpmiddelen worden gebruikt:

- Cyber kill chain: AI-gestuurd framework voert een aanval uit van begin tot eind. Van initiële reconnaissance tot het bereiken van impact in de laatste fase.
- Webapp en API: Een nauwe focus op webapps en API's. Geen uitvoer van complete aanvalspaden.
- C2-geïntegreerd : Gekoppeld aan een C2-framework, AI is de operator die via C2 tools/implants bestuurt.
- *Harnesses*: Nemen een commerciële coding-agent of AI-model (zoals Claude Code) en geven die extra offensieve skills of plug-ins. Hiermee wordt een bestaand AI-model dus verder gefinetuned of aangepast om gespecialiseerd te zijn in offensieve doeleinden
- Onderzoeksprototypes: Ontwikkeld voor onderzoeksprojecten en ontwikkelen van bijvoorbeeld Proof-of-Concepts (PoC) of het uitvoeren van specifieke tests in labomgevingen.

Veelgebruikte frameworks bij incidenten en aanvalscampagnes zijn onder andere: ARTEX, CyberStrikeAI, PentestAGI, Agent Zero, Cairn, Strix, RedAmon, HexStrike, en AntiFox. Deze lijst is niet uitputtend en aan veranderingen onderhevig. Doordat de meeste van deze frameworks open source zijn (of zijn geweest), kan iedereen tooling downloaden, reverse-engineeren en onder een andere naam weer publiceren. Zo zijn er van ARTEX bijvoorbeeld veel verschillende versies, soms met andere namen zoals ScopeWeaver.

TLP:CLEAR

Noemenswaardige incidenten waarbij AI-gestuurde frameworks zijn gebruikt

Onderstaande tabel bevat een aantal noemenswaardige incidenten van de afgelopen periode waarbij AI-gestuurde frameworks voor offensieve doeleinden zijn ingezet. Wat opvalt is dat ARTEX een veel gebruikt en effectief framework is, maar dat ook andere AI-gestuurde frameworks worden ingezet. Bovendien is deze lijst slechts een greep uit het totale aantal incidenten en publiek beschikbare bronnen.

Slachtoffer	Gebruikt framework	Belangrijkste punten
Dutch Institute for Vulnerability Disclosure (DIVD; eigen rapportage[4])	ARTEX	Aanvaller wist twee zero-day kwetsbaarheden te identificeren binnen Zammad. Verder voerde het de aanvaller ARP-spoofing en een succesvolle Man-in-the-Middle (MitM) aanval binnen het netwerk van de DIVD uit. Bij de aanval zijn grote hoeveelheden persoonsgegevens buitgemaakt.
Geanonimiseerd slachtoffer in Nederland (gerapporteerd door Eye Security[5])	ARTEX	De aanval was snel en bewandelde meerdere aanvalspaden in parallel. De aanvaller had een financieel motief en de “luidruchtigheid” van de aanval vormde geen afbreukrisico.
Doelgroepen van het National Cybersecurity Agency Italië[6]	Meerdere incidenten waarbij ARTEX, CyberStrikeAI, DeepSeek Harness – Red Team, en PentAGI zijn ingezet	NCA Italië beschrijft de grootschalige aanvalscampagnes met impact op nationaal niveau waarbij AI-gestuurde frameworks voor offensieve doeleinden zijn ingezet. NCA Italië onderstreept de urgentie voor op orde brengen van de digitale weerbaarheid.
Doelgroepen van het NCSC UK[7]	Meerdere incidenten waarbij frameworks zijn ingezet	NCSC UK beschrijft de toename van aanvallen en incidenten waarbij AI-gestuurde frameworks voor offensieve doeleinden zijn ingezet. Ook de NCSC UK onderstreept de urgentie voor op orde brengen van de digitale weerbaarheid.
Financiële sector in Zuidoost Azië (gerapporteerd door Microsoft[8])	ARTEX	Actor Storm-3358 installeerde verschillende open-source remote access trojans (RATs) na initiële compromittatie. De actor gebruikte ook andere traditionele tools i.c.m. ARTEX om scans uit te voeren.
Europese Commissie (gerapporteerd door CERT-EU[9] en TheRecord[10])	ARTEX	CERT-EU meldt verschillende ARTEX-gerelateerde activiteiten richting hun Europese doelgroepen en verwacht dat de inzet van AI-gestuurde frameworks in de toekomst zal toenemen.
O.a. semiconductor-sector en defensie-sector; (gerapporteerd door Palo Alto Unit 42[11])	Cairn	De actor kreeg toegang door de afwezigheid van basismaatregelen. Eenvoudige password-spray en brute-force aanvallen waren voldoende om initiële toegang te bewerkstelligen.
Zuid-Koreaanse financiële sector (verschillende rapportages[12,13,14])	ARTEX	Na succesvolle aanvallen op de Zuid-Koreaanse financiële sector heeft de ontwikkelaar van ARTEX de tool closed-source gemaakt. Op dat moment bestonden er al meer dan genoeg publieke forks.
Wereldwijde campagne gericht op webwinkels (gerapporteerd door Gambit Security[15])	Strix voor kwetsbaarhedenonderzoek; Cairn exploitatie en uitvoeren aanval; Hermes	Meerdere webplatformen van (grote) webwinkels succesvol aangevallen. De aanvaller wist meer dan 600.000 actieve creditcard-details buit te maken en op vijf websites card-skimming malware te

TLP:CLEAR

	voor aansturen operatie	installeren. Kosten per succesvolle compromittatie waren ongeveer $25.
Geanonimiseerd slachtoffer (gerapporteerd door Wiz.io[16])	Combinatie van CyberStrikeAI en ARTEX	De aanvaller gebruikte bij de operatie CyberStrikeAI. ARTEX werd gelijktijdig bij andere niet-succesvolle campagnes ingezet. Onderdeel van de explosieve groei van aanvallen met AI-gestuurde frameworks.
Honderden organisaties wereldwijd doelwit en slachtoffer (gerapporteerd door Microsoft[17])	ARTEX	Geautomatiseerde verkenning, password spraying en OAuth-misbruik door de Actor Storm-0940 en benutte daarna vertrouwde ConnectWise-toegang. Bij een sub-campagne werden 327 accounts bij acht organisaties gecompromitteerd.
Indiase universiteit en e-commerce platform (gerapporteerd door Palo Alto Unit 42[18])	Hermes Agent	Een actor met beperkte technische kennis is succesvol in het exfiltreren van een grote hoeveelheid gevoelige persoonsgegevens en in kaart brengen van meerdere organisaties.
Telecomproviders in o.a. België en Verenigde Staten; Meerdere overheden en ministeries wereldwijd (Gerapporteerd door Crowdstrike[19])	ARTEX	Actor Vapor Panda gebruikte ARTEX voor gelijktijdige, gerichte aanvallen bij meer dan 200 organisaties in minstens 14 landen.

TLP:CLEAR

4 Technisch onderzoek werking ARTEX

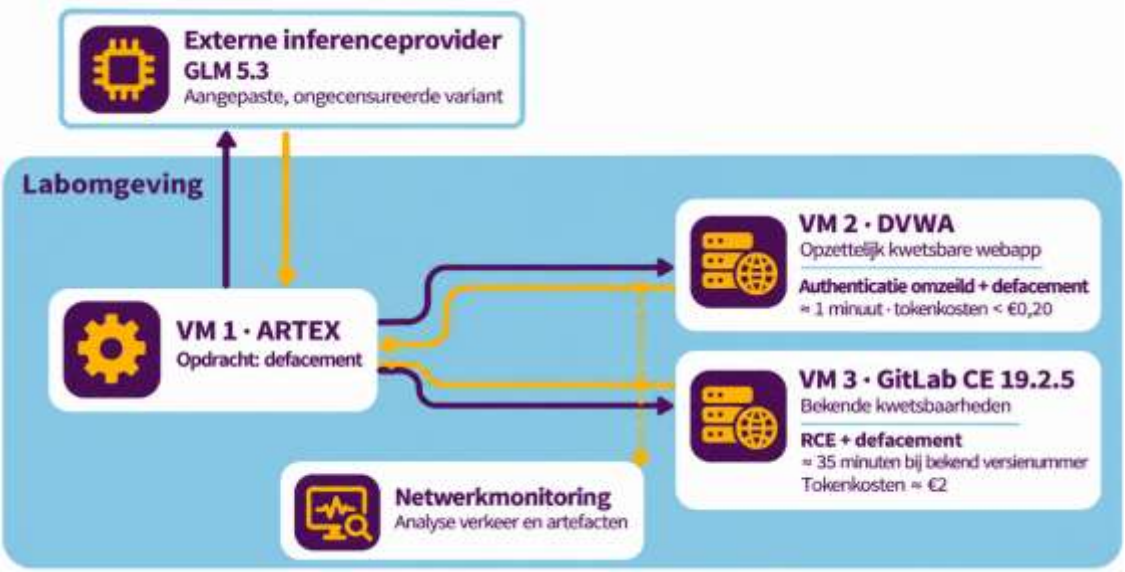
Het NCSC heeft technisch onderzoek verricht naar het AI-gestuurde framework ARTEX. Hierbij is de laatste publieke versie van ARTEX (v0.3.15[20]) ingezet om webapplicaties aan te vallen binnen een geïsoleerde labomgeving.

Het ARTEX framework was geconfigureerd om een AI-model te gebruiken dat wordt aangeboden via een externe aanbieder (inferenceprovider). Specifiek is er gebruik gemaakt van een ongecensureerde versie van het model GLM 5.3 van Z.AI.

- Dit is een aangepaste versie van een “open-weights” model, waarbij ingebouwde veiligheidsfilters en gedragsregels (guardrails) zijn omzeild of verwijderd. Hierdoor weigert het AI-model niet wanneer bijvoorbeeld gevraagd wordt om offensieve activiteiten zoals een cyberaanval uit te voeren. Het opzetten van een nieuwe ARTEX omgeving vereist weinig technische skills en is in korte tijd mogelijk.

Als testdoelwitten werden twee webapplicaties opgezet in aparte virtuele machines. In het eerste deel van het onderzoek werd de Damn Vulnerable Web Application (DVWA) gebruikt, een opzettelijk kwetsbare referentieapplicatie. In het tweede deel werd een twee maanden oude versie van GitLab Community Edition (v19.2.5) gebruikt, welke publiek bekende kwetsbaarheden bevat. In beide gevallen is ARTEX gevraagd om een defacement uit te voeren. Tijdens het onderzoek is netwerkverkeer gemonitord om in kaart te brengen hoe de aanval eruitzag en welke artefacten eruit gehaald konden worden.

- Bij het eenvoudigere doelwit, DVWA, lukte het ARTEX om binnen ongeveer een minuut de authenticatie te omzeilen en binnen dezelfde periode een defacement uit te voeren. De kosten qua tokengebruik voor deze aanval bedroegen minder dan €0,20.
  Bij de GitLab-omgeving duurde de aanval langer. ARTEX had moeite om de exacte versie van de GitLab-omgeving te achterhalen. Met kennis van het versienummer duurde het ongeveer 35 minuten om een Remote Code Execution (RCE) kwetsbaarheid in te zetten en de loginpagina te defacen. De kosten qua tokengebruik hiervoor lagen rond de €2.



TLP:CLEAR

OPSEC-instructies verminderen luidruchtigheid aanval

Een opvallende bevinding betreft de detecteerbaarheid van de aanval. Wanneer alleen een doelwit werd opgegeven, was het gedrag van ARTEX zeer herkenbaar, en luidruchtig. ARTEX liet duidelijke artefacten achter, zoals dat “ARTEX” expliciet werd benoemd in de User-Agent van verzoeken.

- Wanneer er instructies werden meegegeven om de operationele beveiliging (*OPSEC*) te verbeteren, werden deze herkenbare artefacten niet meer gegenereerd.
- De User-Agent werd aangepast naar "Mozilla/5.0 (X11; Linux x86_64) Chrome/126.0". Ook werden benamingen zoals "Artx" niet meer gebruikt.
- In de kwetsbare GitLab-omgeving is een account gemaakt genaamd 'DEV' met het generieke e-mailadres: dev.alice2@example.com.

De resultaten van deze eenvoudige testopstelling tonen aan dat een AI-gestuurd framework voor het uitvoeren van aanvallen in een geïsoleerde omgeving snel en relatief goedkoop kan zijn, en dat de detecteerbaarheid beïnvloed wordt door meegegeven OPSEC-instructies.

TLP:CLEAR

5 Duiding

Bij een meerderheid van de hedendaagse incidenten maken aanvallers gebruik van AI-gestuurde frameworks voor offensieve doeleinden. Dit is met een goede reden: het behaalt resultaten, werkt snel en raakt niet uitgeput zoals een menselijke operator.

De verandering binnen het huidige dreigingslandschap is echter niet betere, meer geavanceerde of zelfs nieuwe aanvallen. Het is de capaciteit om aanvallen op grote schaal en sneller uit te voeren. Een menselijke operator met een AI-gestuurd framework wordt een team die in parallel aanvallen op meerdere doelwitten kan uitvoeren. Taken die voorheen dagen duurden, kunnen in uren of zelfs minuten worden uitgevoerd. Bovendien kan een framework non-stop doorwerken en gaat het alle parameters, endpoints, accounts en andere ingangen systematisch af.

- De **toename in snelheid** betekent dat verdedigers minder tijd hebben om in te grijpen. Aanvallen kunnen volledig zijn uitgevoerd voordat ze worden opgemerkt. Dit is een aanzienlijk verschil vergeleken met digitale aanvallen voor de komst van AI.
- Door de **schaalbaarheid** en het kunnen uitvoeren van meerdere aanvalspaden in parallel (en in het slechtste geval meerdere aanvallen tegelijkertijd) kan dit bovendien de verdediging van organisaties overbelasten en lam leggen. De hoeveelheid alerts en benodigde handelingen van verdedigers zijn op dat moment niet meer bij te benen.
    - Het eindrapport van de hack bij de DIVD toont ter illustratie de enorme hoeveelheid queries, alerts en Indicators of Compromise (IoC's) die er gegenereerd worden.[21]

Het NCSC acht het zeer waarschijnlijk dat de inzet van AI-gestuurde frameworks bij digitale aanvallen zal toenemen met toenemende impact. Dit onderstreept de urgentie voor implementatie van adequate beveiligingsmaatregelen.

- Ook NCSC UK[22], NCA Italië[23], de Duitse BSI[24], het Zwitserse NCSC[25], CERT-EU[26], de Amerikaanse FBI[27] en het Canadian Centre for Cyber Security[28] hebben recent vergelijkbare waarschuwingen gepubliceerd.

Een effectief instrument binnen het aanvalsarsenaal

Het NCSC schreef al eerder dat de inzet van AI de instapdrempel voor digitale aanvallen verlaagt en dat aanvallers minder technisch onderlegd hoeven te zijn.[29] De inzet van AI-gestuurde frameworks voor offensieve doeleinden verlaagt deze drempel nog verder.

- Gedurende een aanval doet een aanvaller operationele ervaring op en verbeteren zijn capaciteiten. Niet door zelf aan de knoppen te zitten en acties uit te voeren, maar door in gesprek te zijn met het AI-model binnen het gebruikte framework. Het framework fungeert daarbij als malafide adviseur en voert tegelijkertijd ook acties en aanvallen uit.
- Dit betekent tijdens een aanval en bij opeenvolgende aanvallen de kwaadwillende leert en zich verbetert op onderdelen waar de aanval eerder vastliep of misschien gedetecteerd werd. Dit aanpassingsvermogen wordt door AI aanzienlijk versneld en maakt trial-en-error

TLP:CLEAR

geen tijdrovende klus waarbij een menselijke operator tussentijds nieuwe vaardigheden moet ontwikkelen.

- AI-gestuurde frameworks en AI-modellen zijn constant in ontwikkeling en verbeteren zich voortdurend**. Het NCSC acht het waarschijnlijk dat in de komende maanden het uitvoeren van meer geavanceerde aanvallen voor steeds meer actoren binnen bereik ligt met de inzet van AI-gestuurde frameworks.**
- Bovendien zijn de kosten van een aanval inmiddels laag. Er zijn genoeg AI-gestuurde frameworks open-source en gratis te downloaden. Daarnaast zorgen de lage kosten en brede beschikbaarheid van AI-modellen via externe aanbieders ervoor dat een aanval slechts enkele euro's hoeft te kosten of zelfs volledig vanaf eigen hardware uitgevoerd kan worden. Ook kunnen aanvallers illegaal gebruik maken van doorverkochte gestolen API-sleutels van AI-modellen.
- De combinatie van de lage kosten en de snelheid waarmee AI-gestuurde frameworks opereren maakt dat individuen kunnen uitgroeien tot een zeer capabele actor die op grote schaal en binnen korte tijd aanvallen kan uitvoeren.

Resultaten op dit moment nog wisselend

De resultaten van digitale aanvallen met behulp van AI-gestuurde frameworks voor offensieve doeleinden zijn op dit moment nog wisselend. De frameworks kunnen zeer geavanceerde creatieve oplossingen op bepaalde beveiligingsmaatregelen te omzeilen maar vervolgens terugvallen op simpele en minder effectieve handelingen in de volgende stap.

- Onderzoekers van Recorded Future omschrijven dit als de *jagged edge*: aanvallers behalen significante resultaten, op grotere schaal, en met gevorderde technische capaciteiten, maar falen tegelijkertijd in simpele taken en operationele beveiliging (OPSEC) die niet passen bij een ervaren aanvaller.[30]
- Ook creëren de inzet van autonome AI frameworks nog vaak veel ruis. Daarom zal de inzet van deze frameworks op dit moment vooral ingezet worden bij aanvallen waarbij succesvolle detectie van een aanval geen groot probleem van de aanvaller vormt.
- Naarmate de frameworks en gebruikte AI-modellen zich ontwikkelen, zal ook de hoeveelheid ruis afnemen. Zeker wanneer menselijke operators beter technisch onderlegd zijn, meer ervaren worden en hun prompts verbeteren om langer en beter buiten het zichtveld van verdedigers te blijven. Zoals het onderzoek in het vorige hoofdstuk al aantoonde: simpele OPSEC-instructies bemoeilijken de detectie van een dergelijk framework zoals ARTEX.

Niveau van autonomie neemt toe

Het CTI-report AI in de Cyber Kill ChAIn introduceerde een taxonomie die de mate van autonomie bij de inzet van AI bij digitale aanvallen beschrijft. Deze taxonomie is ook in appendix A van dit rapport terug te vinden.

- Voorbeelden uit de eerste helft van 2026 vielen voornamelijk in de categorie AI-1 t/m AI-3 waarbij de menselijke operator nog een belangrijke uitvoerende rol speelt in het maken van tussentijdse beslissingen en uitvoeren van taken (Human-in-the-Loop).

TLP:CLEAR

- Met de toenemende inzet van AI-gestuurde frameworks de afgelopen maanden verschuift het zwaartepunt naar AI-4 (AI-geregisseerd). Het is zeer waarschijnlijk dat nieuwe ontwikkelingen op het gebied van AI het niveau van autonomie van AI-gestuurde frameworks zal verhogen en de menselijke operator een steeds kleinere rol in de uitvoer heeft en alleen nog doelen stelt en bijstuurt waar nodig.

Technische achterstand pijnpunt bij incidenten

Net als menselijke aanvallers zoeken AI-gestuurde frameworks en AI-modellen bij digitale aanvallen in het begin naar laaghangend fruit: misconfiguraties, kwetsbare systemen, inloggegevens, secrets, API-sleutels en andere “open deuren”. Misbruik hiervan is geen baanbrekende nieuwe aanvalsmethode.

- Wel geldt hier ook weer dat AI de force multiplier is. Waar een menselijke operator voorheen poging voor poging te werk moest gaan, kan met behulp van AI dit op grote schaal en parallel worden uitgevoerd. Daarnaast geeft het AI-model ook analyse en advies over de buitgemaakte data, vervolgstappen en kan deze ook autonoom uitvoeren op basis van de wensen van de menselijke operator.
- Bestaande beveiligingsmaatregelen blijven effectief, maar de inzet van AI-gestuurde frameworks bij digitale aanvallen legt technische achterstand bij organisaties versneld bloot.
- Daarmee is de urgentie voor de juiste en effectieve maatregelen toegenomen.

TLP:CLEAR

6 Handelingsperspectief

De ontwikkelingen rondom AI staan niet los van bestaande beveiligingsmaatregelen en de basisprincipes. AI-gestuurde frameworks voor offensieve doeleinden zetten bijna altijd traditionele en bekende aanvalsmethodes in. Daardoor is het essentieel dat de basisprincipes op orde zijn binnen jouw organisatie.

De gebruikte voorbeelden in dit rapport tonen aan hoe efficiënt AI-gestuurde frameworks misconfiguraties te vinden en uit te buiten. Ook slecht beveiligde accounts en identiteiten (of wanneer deze veel rechten hebben) geven AI-gestuurde frameworks en hun operators vrij spel. Het NCSC raadt organisaties daarom verder aan een aantal maatregelen een verhoogde prioriteit te geven.

Preventie

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?

- **Scan continu het eigen netwerk**. Inventariseer welke webapplicaties, VPN-systemen, en beheerinterfaces met het internet verbonden zijn. Dit heet ook wel *External Attack Surface Management*.
    - Identificeer publieke API's, secrets, credentials, trek deze in en vervang ze wanneer dit niet de bedoeling is. Sla gevoelige tokens en sleutels niet op in broncode of configuratiebestanden.
    - Verwijder systemen en applicaties die niet publiek bereikbaar hoeven te zijn.
    - Schakel ongebruikte functionaliteiten uit.
- **Bescherm informatie over jouw netwerk**: Publiceer OpenAPI-specificaties, source maps, runtimeconfiguraties en interne API-routes alleen wanneer dat noodzakelijk is.[31] Beveilig deze informatie op een need-to-know basis en voorkom dat tokens, client-ID's of andere gevoelige gegevens erin voorkomen.
- **Patch kwetsbaarheden zo snel mogelijk:** Ontwikkelingen op het gebied van AI zorgen ervoor dat kwetsbaarheden sneller gevonden en misbruikt kunnen worden. Meer informatie over kwetsbaarhedenbeheer is te vinden in de NCSC-factsheet Verbeter je kwetsbaarhedenbeheer
- **Voorkom onveilige configuraties**. Breng configuratie-instellingen en toegangsrechten in kaart. Pas het principe van 'least privilege' toe. Scheid beheerders- en gebruikersaccounts,

TLP:CLEAR

schakel anonieme toegang uit wanneer het niet noodzakelijk is en besteed specifiek aandacht aan gastaccounts en rechten van applicaties. Pas onveilige standaardconfiguraties aan voordat een systeem in gebruik wordt genomen. Voor meer informatie over misconfiguraties, zie de NCSC-website.

Bescherm identiteiten

Beperk gebruikers- en applicatierechten 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, kan die geen processen manipuleren zonder deze administratieve rechten. Maak gebruik van sterke wachtwoorden en implementeer multifactorauthenticatie voor alle toegang. Kies je authenticatiemiddel op basis van het bijbehorende risico.

Detectie

Monitoring, detectie en logging

Het beschikbaar hebben van adequate logging vormt het fundament om aanvallen tijdig te detecteren en te kunnen mitigeren. Met de snelheid waarmee AI-gestuurde frameworks opereren is het hebben van zicht en daarop kunnen acteren een belangrijke prioriteit. Afwezigheid van logging betekent niet automatisch dat er geen malafide activiteit is geweest.

- **Zorg voor inzicht in het gebruik en potentieel misbruik** van alle systemen, endpoints en interne AI-diensten en tools door deze te laten loggen. Sla deze logging centraal op en monitor op misbruik.[32,33]
- **Correleer detectie uit verschillende logbronnen met elkaar.** Scans naar misconfiguraties lijken in veel gevallen routine, maar in combinatie met andere signalen kunnen deze wijzen op een geautomatiseerde aanval.[34,35,36]
- **Stel alerts in voor afwijkend gedrag op basis van snelheid.** Een kenmerk van AI-gestuurde frameworks is dat ze bepaalde aanvalsmethodes in zeer korte tijd (en soms parallel) uitvoeren totdat alles is geprobeerd of een doel is bereikt. Dergelijke snelheden worden bij normaal gedrag van mensen niet verwacht.
- **Voor Microsoft-omgevingen** zijn minimaal de volgende logs zijn om aanvallen van AI-gestuurde frameworks te kunnen detecteren:
    - webserver- en WAF-logs,
    - endpoint-proces- en bestandsgegevens,
    - interactieve en niet-interactieve aanmeldingen,
    - service-principal-aanmeldingen,
    - Microsoft Graph-activiteit en auditlogs van Key Vault, Storage en cloudbeheer.
- Meer informatie over hoe Microsoft Defender dit soort aanvallen kan detecteren is te lezen in een blog van Microsoft MVP Jeffrey Appel.[37]
- ThreatDefence heeft een database met detectieregels voor meerdere AI-gestuurde frameworks.[38] Het NCSC heeft de werking van deze regels niet eigenstandig geverifieerd.

TLP:CLEAR

Digitaal struikeldraad

AI-gestuurde frameworks speuren tijdens aanvallen constant naar gevoelige gegevens en andere waardevolle informatie. Ze vallen daarbij ook gretig honeytokens, canary files en andere valstrikken aan wat kansen biedt voor verdedigers.

Met misleiding-gebaseerde detectie plaats je valstrikken in jouw digitale omgeving: realistische nep-assets (decoys) die interactie met een aanvaller uitlokken. Denk aan slecht-beveiligde admin-accounts (honeyaccounts), eenvoudig toegankelijke, gevoelige bestanden of nep-API-sleutels (honeytokens) of kwetsbare systemen (honeypots).

- Het is van belang om hiervoor wel geloofwaardige bestanden en tokens te ontwerpen. Een aanvaller kan aangeven dat de AI-agent moet uitkijken voor dit soort valstrikken en deze niet aan te vallen.
- Verder beschrijven meerdere onderzoekers hoe bepaalde honeytokens en andere technieken specifiek AI-gestuurde frameworks en AI-modellen om de tuin leiden en tijd laten verspillen.
    - Zo kan een honeytoken worden aangepast om dynamisch de hoeveelheid informatie die wordt aangeboden aan te passen op basis van de hoeveelheid aanvragen om de AI-agent in een eindeloze loop vast te houden.[39]
    - Tracebit beschrijft verschillende vormen van *context bombs*. Het doel is om de nieuwsgierigheid en resultaatgerichtheid van AI-modellen te gebruiken tegen zichzelf om zo AI-agents onnodig tijd te laten verspillen.[40]

Threat hunting

Maak gebruik van de assume breach denkwijze waarbij je je organisatie voorbereidt op een succesvolle digitale aanval. Het is namelijk geen garantie dat detectie alle aanvallen zal detecteren en op tijd mitigeren. De kans is altijd aanwezig dat je als organisatie slachtoffer wordt van een succesvolle aanval.

- Bij veel van de recente aanvallen waarover gerapporteerd is zijn er herkenbare *Intent IDs* te detecteren. Dit zijn kenmerken die toebehoren aan deelacties die door AI-agents worden uitgevoerd, een soort interne taaknummers van het framework. Die worden achtergelaten op het netwerk van het slachtoffer. Doordat dit vaak specifieke cijferreeksen en strings betreft kan hier gericht naar gezocht worden. Beveiligingsbedrijf Wiz.io beschrijft in een blog hoe je hier onderzoek naar kan doen.[41]
- ARTEX maakt bovendien gebruik van bepaalde User-Agent-strings die activiteit van ARTEX aanduiden.[42] Ook andere AI-gestuurde frameworks vertonen soortgelijk gedrag. De kanttekening hierbij is dat een aanvaller dergelijke opvallende user agent strings kan vermijden met simpele aanpassingen van prompts waardoor hier specifiek naar zoeken geen garantie is of er wel of geen malafide activiteit is geweest.
- Appendix B heeft een aantal voorbeelden van ARTEX-artefacten en signalen.

TLP:CLEAR

Response

Onderzoek kunnen doen na een cyberaanval is belangrijk, het is goed denkbaar dat niet meteen duidelijk is wat de staat van de compromittatie van jouw systemen en gegevens is en hoe te herstellen. Daarom wil je forensic ready zijn: de paraatheid hebben om digitaal forensisch onderzoek te (laten) doen. Dit artikel vertelt je wat je daarvoor zelf kunt doen en waarvoor je ondersteuning kunt inschakelen.

- Een incidentresponsplan (IRP) is essentieel om voorbereid te zijn op een cyberincident. Met een IRP stel je interne en externe processen vast, identificeer je rollen/verantwoordelijkheden/taken, stel je een communicatieplan op en maak je afspraken. Wie mag bijvoorbeeld de productie stilzetten tijdens een malware-infectie ('stekkermandaat')?. Meer informatie is te vinden op de website van het NCSC.
- Wapen je jezelf tegen cyberincidenten met impact op de beschikbaarheid, vertrouwelijkheid en integriteit van jouw systemen en data. Stel hiervoor een Disaster Recovery Plan (DRP) op. Het plan bevat procedures om na de calamiteit weer terug te keren naar de normale bedrijfsvoeringssituatie.

Veilig gebruik van AI binnen jouw organisatie
Ook verdedigers kunnen AI inzetten om zich beter te wapenen tegen AI-gestuurde aanvallen. De noodzaak hiervoor zal in de toekomst alleen maar toenemen. Niet alleen om te helpen bij triage en afhandelen van security alerts, maar ook voor het geautomatiseerd nemen van beslissingen. Ook voor het scannen van jouw netwerkomgeving en aanvalsoppervlak (zoals beschreven onder Preventie), kan AI een uitkomst bieden.

Belangrijk hierbij is dat je dit wel op een gecontroleerde en veilige manier doet. Dit betekent dat je tools die aanvallers inzetten zoals bijvoorbeeld ARTEX of andere frameworks niet zomaar los moet laten, aangezien dit onvoorziene consequenties kan hebben. Ook de keuze voor te gebruiken AI-modellen vereist voorbereiding. Ongeverifieerde modellen kunnen mogelijk gevoelige gegevens en informatie lekken naar kwaadwillenden. Zo toonde onderzoek van Dragos aan hoe verschillende AI-modellen van inferenceproviders geheime informatie en secrets lekten naar kwaadwillenden.[43]

Meer informatie over veilig gebruik van AI binnen jouw organisatie is te vinden op de website van het NCSC.

TLP:CLEAR

7 Appendix A: 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[44], OWASP[45], CSA[46], NVIDIA[47], AWS[48], Anthrophic[49] en Mitre[50].

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

TLP:CLEAR

8 Appendix B: ARTEX-artifacten

Artefact	Locatie
artex-enrich/*	User-Agent in web-, proxy-, WAF- en Graph-logs
norma/*	User-Agent in web- en cloudlogs
artex-…-i<nummer> of pentest-…-i<nummer>	Accountnamen, cloudresources, tags, workflows, bestanden en User-Agents
tasks/<taak>/i<nummer>	Opdrachtregels, proceslogs, scripts en foutmeldingen
exp<nummer>-worker-i<nummer>	Proces-, sessie- of opdrachtregelgegevens
Behavior:Win64/ArtexAI	Microsoft Defender Antivirus

Meer artefacten zijn ook te vinden in een recente blog van ThreatDefence.[51] Belangrijke kanttekening is dat deze door de aanvaller makkelijk gewijzigd kunnen worden. Afwezigheid is dus geen garantie dat er geen malafide activiteit heeft plaatsgevonden.

TLP:CLEAR

9 Bronnen en referenties

[1] https://www.ncsc.nl/artificial-intelligence-ai/artificial-intelligence-in-de-cyber-kill-chain
[2] https://www.greynoise.io/blog/ai-orchestrated-campaign-against-papercut-ng-mf
[3] https://blog.gitguardian.com/ai-pentesting-tools-beyond-artex/
[4] https://csirt.divd.nl/cases/DIVD-2026-00014/full_disclosure/
[5] https://labs.eye.security/how-we-defended-against-an-ai-driven-attack/
[6] https://www.acn.gov.it/portale/w/impiego-di-agenti-di-ia-in-campagne-di-compromissione-automatizzate
[7] https://www.ncsc.gov.uk/blogs/familiar-weaknesses-exploited-faster-defending-against-agentic-cyber-attacks
[8] https://www.microsoft.com/en-us/security/blog/2026/11/06/artex-abuse-expands-as-cybercriminals-target-financial-sector/
[9] https://www.cert.europa.eu/blog/from-penetration-testing-to-cyber-threat-beware-of-artex
[10] https://therecord.media/european-commission-confirms-cyberattack-using-ai
[11] https://unit42.paloaltonetworks.com/cairn-framework-ai-autonomous-attacks/
[12] https://www.crowdstrike.com/en-us/blog/unknown-threat-actor-uses-artex-to-target-south-korean-finance/
[13] https://threatbook.io/blog/artex-attacks-south-korean-financial-institutions
[14] https://www.zscaler.com/blogs/security-research/threat-actor-abuses-agentic-ai-pentesting-framework-target-south-korean
[15] https://gambit.security/blog-posts/autonomous-ai-agents-online-retailers-25-a-company
[16] https://www.wiz.io/blog/agentic-attacks-artex-offensive-ai-frameworks
[17] https://www.microsoft.com/en-us/security/blog/2026/10/21/storm-0940-uses-artex-agentic-ai-connectwise-supply-chain-compromise/
[18] https://unit42.paloaltonetworks.com/ai-assisted-cybercrime-analysis/
[19] https://www.crowdstrike.com/en-us/blog/vapor-panda-uses-artex-to-scale-global-campaign/
[20] https://www.virustotal.com/gui/file/bfcc783609d5f57f3d5e25725ec4b693c9dbc45d3ff9029c4f10cc552e328dcb
[21] https://csirt.divd.nl/downloads/DIVD-2026-00014/DIVD-2026-00014-forensics-report.pdf
[22] https://www.ncsc.gov.uk/blogs/familiar-weaknesses-exploited-faster-defending-against-agentic-cyber-attacks
[23] https://www.acn.gov.it/portale/w/impiego-di-agenti-di-ia-in-campagne-di-compromissione-automatizzate
[24] https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Offensive-agentische-KI_261115.pdf
[25] https://www.bacs.admin.ch/en/26-ki-agenten-1-en
[26] https://www.cert.europa.eu/blog/from-penetration-testing-to-cyber-threat-beware-of-artex
[27] https://www.ic3.gov/CSA/2026/261113.pdf
[28] https://www.cyber.gc.ca/en/news-events/cyber-threat-actors-increasingly-use-automated-ai-enabled-tools-cyber-attacks
[29] https://www.ncsc.nl/expertblogs/ai-legt-de-vinger-op-de-kwetsbare-plek
[30] https://www.recordedfuture.com/blog/tag-217-artex-south-korean-banks
[31] https://unit42.paloaltonetworks.com/artex-and-ai-powered-red-teaming-tools/
[32] https://www.ncsc.nl/risicomanagement/routekaart-risicomanagement/doorlopende-monitoring
[33] https://github.com/JSCU-NL/logging-essentials
[34] https://www.elastic.co/security-labs/threat-command/cyberstrikeai-aws-cloudtrail-intrusion
[35] https://www.puppygraph.com/learn/security-correlation-rules
[36] https://www.splunk.com/en_us/blog/learn/it-event-correlation.html
[37] https://jeffreyappel.nl/artex-old-attacks-at-machine-speed-and-how-microsoft-defender-keeps-up/
[38] https://github.com/ThreatDefence/agentic-attack-detections
[39] https://plaintext.do/en/blog/hacking-back-ai-pentesting-agents/
[40] https://tracebit.com/blog/context-bombs-that-misdirect-ai-attackers
[41] https://www.wiz.io/blog/agentic-offensive-frameworks-intent-ids-artex-cairn

[42] https://www.microsoft.com/en-us/security/blog/2026/11/25/what-defenders-need-to-know-artex-agentic-ai-framework-threat-actors-abuse/
[43] https://www.dragos.com/blog/ai-supply-chain-security-canary-tokens
[44] https://www.sophos.com/en-us/content/sophos-ai-security-2026-report
[45] https://owasp.org/APTS/standard/4_Graduated_Autonomy/
[46] https://cloudsecurityalliance.org/blog/2026/01/28/levels-of-autonomy
[47] https://developer.nvidia.com/blog/agentic-autonomy-levels-and-security/
[48] https://aws.amazon.com/blogs/security/the-agentic-ai-security-scoping-matrix-a-framework-for-securing-autonomous-ai-systems/
[49] https://www.anthropic.com/research/attack-navigator
[50] https://atlas.mitre.org/
[51] https://threatdefence.com/article/artex-agentic-attack-framework-detection-hunting

TLP:CLEAR

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
@ncsc_nl

December 2026