Auteur: drs. Marcel Feenstra, FRSA MALD MBA is werkzaam als trainer en consultant op het gebied van cybersecurity. Hij is bereikbaar via marcel@afterimage.nl

Kansen en bedreigingen in AI en cybersecurity: terugblik en actuele ontwikkelingen

In iB-Magazine, editie 2 van 2024 stonden diverse artikelen over de relatie tussen kunstmatige intelligentie (AI) en informatiebeveiliging. Inmiddels zijn we een jaar verder en de ontwikkelingen op AI-gebied hebben bepaald niet stilgestaan. De wereld lijkt volledig in de ban van AI: de koers van chipfabrikant NVIDIA is naar recordhoogte gestegen, en bedrijven zoals Samsung en Apple benadrukken de 'ingebouwde AI' als belangrijkste verkoopargument voor hun nieuwste producten. Bovendien gaat er vrijwel geen dag voorbij zonder dat tools als ChatGPT of andere Large Language Models (LLM's) worden geprezen.

Mede-redactielid Fook Hwa Tan stelde destijds dat AI weliswaar veel nieuwe mogelijkheden biedt, maar dat wij als mens nieuwe vaardigheden nodig hebben om de adviezen en resultaten van deze technologie kritisch te beoordelen en ethisch in te zetten. Transparantie over de criteria die AI gebruikt, is daarbij essentieel om te waarborgen dat privacygevoelige informatie op verantwoorde wijze wordt verwerkt en dat AI-systemen vrij zijn van ingebouwde vooroordelen (bias).

Ruben Faber en Dion Koeze wezen in diezelfde editie op de kwetsbaarheden van AI, zoals manipulatie van trainingsdata (data poisoning), modeldiefstal en injection attacks waarbij AI-systemen zelf onderdeel van de aanval worden. Daarnaast benadrukte Vincent van Dijk dat AI geen oplossing is als je niet eerst helder hebt welk probleem je probeert op te lossen: 'AI is niet het antwoord als je niet weet welke vraag je stelt.'

Dasha Simons voegde toe dat sommige AI-modellen specifiek voor één taak worden ontwikkeld, terwijl andere modellen, zoals Large Language Models als ChatGPT, breed inzetbaar zijn. Ze waarschuwde voor het risico van bias en benadrukte dat modellen altijd getoetst moeten worden op betrouwbaarheid en geschiktheid voor de beoogde taak.

Verbeterde detectie en respons

AI biedt vooral kansen door bedreigingen sneller en nauwkeuriger te detecteren dan traditionele beveiligingssystemen. Waar conventionele systemen vaak reactief zijn, kan AI proactief afwijkingen identificeren en analyseren. Een voorbeeld hiervan is Darktrace, dat gebruikmaakt van Machine Learning om afwijkingen in netwerkverkeer te

Generatieve AI maakt het bijvoorbeeld mogelijk om overtuigende deepfakes of phishing-e-mails te creëren die steeds moeilijker te onderscheiden zijn van legitieme communicatie

detecteren. Darktrace technologie, bekend als het Enterprise Immune System, leert continu wat 'normaal' netwerkgedrag is en kan onmiddellijk reageren op afwijkingen die kunnen duiden op een aanval. Zo wist Darktrace bijvoorbeeld een geavanceerde insider-aanval te detecteren waarbij een medewerker gevoelige gegevens probeerde te exfiltreren via een versleutelde verbinding. Dankzij de snelle detectie kon het beveiligingsteam tijdig ingrijpen en schade voorkomen.

Voorspellende analyses

Een ander groot voordeel van AI is de mogelijkheid om voorspellende analyses uit te voeren. Door historische gegevens te analyseren, kan AI toekomstige dreigingen voorspellen en organisaties in staat stellen om proactief beveiligingsmaatregelen te nemen.

Een voorbeeld hiervan is IBM's Watson for Cyber Security, dat patronen in cyberaanvallen herkent en toekomstige dreigingen voorspelt. Door enorme hoeveelheden ongestructureerde data (zoals artikelen en blogs) te combineren met gestructureerde beveiligingslogs, kan Watson niet alleen actuele dreigingen identificeren, maar ook voorspellingen doen over nieuwe aanvalsmethoden. Cargills Bank in Sri Lanka gebruikte Watson in combinatie met IBM QRadar SIEM om bedreigingen vroegtijdig te detecteren en te classificeren, wat resulteerde in een aanzienlijke verbetering van hun beveiliging.

Automatisering van beveiligingstaken

AI kan routinetaken zoals netwerkmonitoring, kwetsbaarheidsscans en toegangsbeheer automatiseren. Dit vermindert de werkdruk van beveiligingsteams en stelt hen in staat om zich te richten op complexere dreigingen.

Een voorbeeld hiervan is Splunk SOAR (Security Orchestration, Automation, and Response). Met deze tool kunnen beveiligingsteams workflows automatiseren, waardoor ze sneller kunnen reageren op veelvoorkomende bedreigingen. Zo gebruikte het Britse financiële platform Tide Splunk SOAR om hun incidentrespons met een factor vijf te versnellen, doordat handmatige taken vrijwel volledig werden geautomatiseerd.

Threat intelligence-integratie

AI kan worden ingezet om real-time bedreigingsinformatie te verzamelen en te analyseren. Door AI-gestuurde threat intelligence te gebruiken, kunnen organisaties sneller reageren op nieuwe kwetsbaarheden en aanvalstechnieken. Een voorbeeld hiervan is het platform Trellix, dat AI gebruikt om bedreigingsinformatie te correleren met interne beveiligingsdata. Dit leidt tot snellere besluitvorming en effectievere risicobeheersing. Tijdens verkiezingen in een Amerikaanse staat zorgde deze aanpak ervoor dat potentiële dreigingen vroegtijdig konden worden geadresseerd.

De keerzijde van de medaille

Hoewel AI aanzienlijke voordelen biedt, brengt ze ook nieuwe risico's met zich mee. Cybercriminelen gebruiken dezelfde technologie om complexere aanvallen uit te voeren. Generatieve AI maakt het bijvoorbeeld mogelijk om overtuigende deepfakes of phishing-e-mails te creëren die steeds moeilijker te onderscheiden zijn van legitieme communicatie.

In 2019 werd een Britse CEO opgelicht voor 243.000 dollar nadat hij dacht te telefoneren met zijn baas, terwijl criminelen een met AI gegenereerde stem gebruikten. Dit soort aanvallen wordt steeds toegankelijker en vormt een serieuze bedreiging.

Daarnaast kunnen AI-modellen zelf het doelwit zijn van aanvallen. Bij een aanval zoals model inversion proberen kwaadwillenden gevoelige gegevens te achterhalen die zijn gebruikt om een machine learning-model te trainen. Onderzoekers hebben aangetoond dat dit type aanval persoonlijke informatie uit gezichtsherkenningssystemen kan blootleggen. Dit benadrukt het belang van robuuste beveiligingsmaatregelen rondom AI-modellen.
Een andere dreiging wordt gevormd door adversarial attacks, waarbij kleine aanpassingen aan inputgegevens AI-systemen misleiden. Onderzoekers toonden bijvoorbeeld aan dat kleine objecten op de weg een Tesla konden verwarren, met mogelijk ernstige gevolgen.

Balans tussen kansen en bedreigingen

Om AI effectief in te zetten in cybersecurity, moeten organisaties een balans vinden tussen kansen en risico's. Hier volgen enkele strategieën:

- Continue aanpassing: zorg ervoor dat AI-systemen regelmatig worden bijgewerkt met nieuwe dreigingsinformatie om hun effectiviteit te behouden.
- Menselijk toezicht: hoewel automatisering nuttig is, blijft menselijk toezicht essentieel om de juiste beslissingen te waarborgen.
- Beveiliging van modellen: bescherm AI-modellen tegen diefstal en manipulatie door sterke toegangscontroles en encryptie toe te passen.
- Bewustwordingstraining: train medewerkers om nieuwe dreigingen zoals deepfakes te herkennen en adequaat te reageren.

Conclusie

De ontwikkelingen op het gebied van AI en cybersecurity gaan razendsnel. Vooral generatieve AI, met steeds krachtigere modellen, blijft zich in hoog tempo ontwikkelen. Dit biedt zowel kansen als nieuwe bedreigingen. Voor professionals in informatiebeveiliging is het essentieel om van deze ontwikkelingen op de hoogte te blijven en hun strategieën continu aan te passen. Door een goede balans te vinden tussen kansen benutten en risico's beperken, kunnen organisaties profiteren van AI zonder hun veiligheid in gevaar te brengen.

Referenties

(1) https://www.linkedin.com/pulse/power-trio-how-ai-quantum-computing-cybersecurity-our-lokhandwala-hlxmf
(2) https://www.ibm.com/case-studies/cargills-bank-ltd
(3) https://www.splunk.com/en_us/customers/success-stories/tide.html
(4) https://www.trellix.com/assets/case-studies/trellix-election-casestudy.pdf
(5) https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
(6) https://spectrum.ieee.org/real-time-deepfakes
(7) https://drlee.io/ai-security-model-hacking-with-model-inversion-attacks-techniques-examples-and-real-world-a23b5fff272a
(8) https://neptune.ai/blog/adversarial-machine-learning-defense-strategies