Algemene Inlichtingen- en Veiligheidsdienst
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

AI-systemen: ontwikkel ze veilig

Waarom moet je extra aandacht besteden aan veilige AI-systemen?

Een automatische scanner voor de doorvoer van goederen die onbedoeld wapens doorlaat. Een op AI-gebaseerd malware-detectieprogramma dat verkeerde trainingsdata heeft gekregen en nu niet meer werkt. Aanvallers die gevoelige gegevens uit je AI-systeem weten te achterhalen. En dat alles zonder dat je het als eigenaar op tijd door hebt. Het zijn slechts een paar voorbeelden van de mogelijk desastreuze gevolgen van gemanipuleerde AI-systemen.

Artificiële (of kunstmatige) intelligentie (AI) geeft computergestuurde machines de mogelijkheid zelfstandig problemen op te lossen[1]. Steeds meer computersystemen maken gebruik van AI of ML-modellen. Of het nu gaat om modellen voor beeldherkenning, spraaktechnologie of cybersecurity – AI kan je helpen om processen sneller, slimmer en beter uit te voeren. Ontwikkelingen op het gebied van AI gaan snel – zo snel dat het nu al belangrijk is om je AI-systemen veilig te ontwikkelen. Beveiliging is niet iets dat je achteraf nog even kunt doen, daar moet je direct vanaf het begin over nadenken (‘*security by design*’). Anders loop je het gevaar dat jouw AI-systeem niet meer werkt zoals het zou moeten, met alle gevolgen van dien.

In deze publicatie deelt het Nationaal Bureau voor Verbindingsbeveiliging (NBV) van de Algemene Inlichtingen- en Veiligheidsdienst (AIVD) manieren waarop AI-systemen aangevallen kunnen worden en hoe je je hiertegen kunt verdedigen. We lichten vijf verschillende aanvallen toe die zich specifiek op AI-systemen richten, ook wel *adversarial AI* genoemd. En we geven vijf principes die je kunt gebruiken bij het veilig ontwikkelen van een AI-systeem.

Wat is het NBV?

Het Nationaal Bureau voor Verbindingsbeveiliging (NBV) heeft als onderdeel van de Unit Weerbaarheid het doel om Nederland digitaal veilig te houden tegen statelijke dreigingen en andere Advanced Persistent Threats (APT’s). Wij zijn uniek doordat wij onze specialistische beveiligingskennis combineren met de bijzondere inlichtingenpositie die we hebben als onderdeel van de AIVD. We werken nauw samen met onze veiligheidspartners MIVD, NCTV, NCSC en CIO Rijk. Gezamenlijk helpen we de Rijksoverheid, vitale sectoren en het bedrijfsleven om bijzondere en gevoelige informatie zoals staatsgeheimen te beschermen.

1 Machine learning (ML) is onderdeel van AI en bestaat uit algoritmes die computers in staat stellen om te leren van data. Trainen is het geautomatiseerd aanpassen van een model door patronen te leren in data. Deze algoritmes kunnen complexe patronen leren die voor de menselijke geest (bijna) niet te herkennen zijn op hoge snelheid.

Zo houd je de controle over jouw AI-systemen

AI-systemen hebben een ander, aanvullend aanvalsoppervlak ten opzichte van 'traditionele' digitale systemen. Zo kunnen aanvallers proberen om jouw AI-modellen om de tuin te leiden, de werking van het systeem te saboteren, of erachter te komen hoe jouw algoritmes werken. Zonder dat je dit zelf door hoeft te hebben. Het is dan ook niet genoeg om je AI-systemen enkel van algemene cybersecuritymaatregelen te voorzien.

Aan de hand van vijf soorten aanvallen laten we zien op welke manieren AI-systemen kwetsbaar kunnen zijn. En we geven daarbij vijf principes voor het veilig ontwikkelen van een AI-model of -systeem. De principes bieden context en structuur om te helpen weloverwogen beslissingen te nemen over systeemontwerp, ontwikkelingsprocessen en helpen bij het beoordelen van specifieke dreigingen voor een AI-systeem. Deze publicatie is bedoeld om AI-experts en cybersecurityexperts binnen een organisatie met elkaar hierover in gesprek te laten gaan.

Houd continu aandacht voor kwaliteit en veiligheid

Voor het veilig ontwikkelen van AI-systemen is geen simpel stappenplan of checklist te volgen. Dat geeft mogelijk de illusie dat je daarna klaar bent en niet meer verder aan de beveiliging hoeft te werken. De principes zijn daarom meer opgesteld als een houvast om over na te denken als je de AI-experts en cybersecurityexperts van jouw organisatie bij elkaar zet. Je kunt ze vergelijken met *secure coding principles*: principes die je kunt volgen bij het veilig schrijven van een 'gewone' code.

Maar niet alleen tijdens het ontwerp, ook bij het gebruik van je AI-systeem moet je de veiligheid in de gaten houden. Ontwikkelingen op het gebied van AI (en dus ook *adversarial AI*) gaan snel, aanvalstypen blijven zich in de toekomst ontwikkelen. Gebruik de principes om op een bewuste manier veilige AI-modellen te maken of gebruiken. We adviseren organisaties om op de hoogte te blijven van verdere ontwikkelingen in dit veld.

Op dit moment classificeert het NBV vijf verschillende categorieën van aanvallen die specifiek op AI-systemen gericht zijn:

1. **Poisoning aanvallen**
2. **Input (evasion) aanvallen**
3. **Backdoor aanvallen**
4. **Model reverse engineering & inversion aanvallen**
5. **Inference aanvallen**

Onze vijf principes zijn:

- √ **Houd je datakwaliteit op orde**
- √ **Zorg voor validatie van je data**
- √ **Houd rekening met supply chain security**
- √ **Maak je model robuust tegen aanvallen**
- √ **Zorg dat je model controleerbaar (auditable) is**

Vijf aanvallen op AI-systemen

1. Poisoning aanvallen

Met een poisoning aanval probeert een aanvaller aanpassingen te maken in je data, algoritme of model, zodat het AI-systeem wordt 'vergiftigd' en daardoor niet meer werkt zoals gewenst. Denk aan spamfilters die kwaadaardige websitelinks onjuist als veilig classificeren. Door dit soort aanvallen neemt de betrouwbaarheid van de output van je AI-systeem af.

Poisoning aanvallen vinden plaats tijdens de trainingsfase van een model. De aanval kan worden uitgevoerd door bijvoorbeeld het toevoegen van vervalste data (data-injectie), het manipuleren van bestaande data (datamanipulatie) of door het labelingsproces te verstoren. Zo leert het model foute conclusies te trekken. Ook kan het algoritme of het model zelf worden beïnvloed - voor, tijdens of na de trainingsfase. Hiervoor heeft de aanvaller schrijfrechten tot (delen van) jouw data nodig.

2. Input (evasion) aanvallen

Een input aanval, ook wel een *evasion* aanval genoemd, is bedoeld om de input voor een AI-systeem dusdanig te bewerken dat het systeem niet of onjuist werkt. Doordat de veranderingen vaak minimaal zijn, en de aanval door het menselijk oog in sommige gevallen niet waar te nemen is, is de detectie van dit soort aanvallen zeer moeilijk. Denk aan verkeersborden die door het opplakken van een post-it een geheel andere betekenis krijgen, waardoor een zelfrijdende auto ongewenste handelingen uitvoert.

Input aanvallen veranderen niets aan het AI-systeem zelf, maar geven bepaalde soorten input waardoor het systeem fouten maakt. De aanval vindt dus plaats als het AI-product al geïmplementeerd is en zorgt dat het systeem bij bepaalde invoer niet functioneert zoals het zou moeten.

3. Backdoor aanvallen

Door een *backdoor* (achterdeurtje) in een AI-model te bouwen, kan een externe partij als het ware een extra vertakking in de beslisboom toevoegen. Daarmee kan de aanvaller de uiteindelijke beslissing van het model bepalen voor specifieke invoer. Bijvoorbeeld: een aanvaller wil niet dat een model voor automatische kentekenherkenning de auto's van een criminele organisatie herkent. Hij weet toegang te krijgen tot het systeem waar dit model wordt ontwikkeld en bouwt een backdoor in, waarmee kentekens met een specifiek kenmerk niet herkend worden. Door dit kenmerk op de kentekens aan te brengen komen ze vervolgens elke keer door de scan heen.

Het kan zomaar zijn dat modellen die je gebruikt, maar niet zelf hebt ontwikkeld, een backdoor bevatten. Denk aan het downloaden van een al getraind model. Deze zou je kunnen trainen op eigen data, maar hoe het oorspronkelijke model is getraind is niet duidelijk. Daarmee kan een backdoor al in het model zitten. In sommige gevallen blijft een backdoor ook aanwezig als je het model opnieuw traint. Het is heel moeilijk, al dan niet onmogelijk, om een goed geïmplementeerde backdoor in een model te ontdekken.

4. Model reverse engineering & inversion aanvallen

Bij het *reverse engineeren* van een AI-model probeert een aanvaller erachter te komen hoe jouw model werkt. Bij *inversion* aanvallen is het doel om de dataset te reconstrueren die gebruikt is om jouw model te trainen. Deze data kan namelijk gevoelige gegevens bevatten die mogelijk interessant zijn voor een aanvaller. De aanvallen kunnen verschillende doelen dienen: bijvoorbeeld het stelen van jouw intellectueel eigendom, of het onderzoeken van zwakheden in je model.

Reverse engineering van een AI-model kan worden uitgevoerd door veel uitvragen naar het model te sturen, om zo stukje bij beetje de werking in kaart te brengen. Een aanvaller is vervolgens in staat om het model in zijn eigen omgeving te draaien, om daar vervolgens te zoeken naar verdere kwetsbaarheden met bijvoorbeeld input aanvallen. Ook kan een aanvaller kijken hoe jouw model reageert op een bepaalde aanval, om daar vervolgens een tegenreactie op voor te bereiden. Model reverse engineering en inversion aanvallen zijn haast niet te detecteren.

5. Inference aanvallen

Inference aanvallen zijn gericht op het achterhalen van (potentieel geheime) trainingsdata. Modellen worden vaak getraind met grote hoeveelheden data die in veel gevallen ook persoonsgegevens of intellectueel eigendom bevatten. Inference aanvallen onderzoeken of een stuk informatie voortkwam in de trainingsdata op basis van de output van het model.

Bijvoorbeeld, je wilt weten of de foto van een persoon gebruikt is om een model voor gezichtsherkenning te trainen. Een inference aanval kan hierachter komen. Een aanvaller kan ook, met beperkte gegevens van een persoon (naam, adres, telefoonnummer), proberen missende informatie te achterhalen (zoals het rekeningnummer of BSN). Als de gegevens van deze persoon gebruikt waren om een bepaald AI-model te trainen, kan een inference aanval helpen deze gegevens te reconstureren.

Vijf principes voor het verdedigen van je AI-systemen

Met deze vijf AI-specifieke aanvallen in het achterhoofd, en de hoge cyberdreiging in het algemeen, is het belangrijk om te weten hoe je jouw AI-systemen hiertegen kunt beschermen. Het NBV heeft hiervoor vijf principes gedefinieerd, op basis van onze eigen kennis en ervaringen, en informatie van onze samenwerkingspartners (zoals TNO en NCSC-UK[2]). Deze principes zijn niet één-op-één gekoppeld aan de vijf aanvallen, en zijn aanvullend op algemene *best practices* voor het ontwikkelen van software en het beheersen van je netwerken en systemen.

Lees hiervoor bijvoorbeeld onze brochures:

- Verdedigbaar Netwerk: Hoe doe je dat?
- Cyberaanvallen door statelijke actoren – zeven momenten om een aanval te stoppen.

Het idee van onze principes is dat ze je helpen nadenken over het veilig ontwikkelen en gebruiken van AI-modellen in je organisatie. Er is geen algemeen toepasbare lijst met maatregelen waarmee je altijd veilig bent. Ook voor het beveiligen van AI-systemen geldt dat risicomanagement leidend moet zijn: wat zijn de consequenties als jouw model niet goed werkt of gestolen wordt? Bepaal op basis daarvan wat het niveau van risicoacceptatie is voor jouw organisatie, en waar je wél strenge maatregelen moet nemen. Beschouw onze principes dan ook altijd in de context van jouw organisatie en de kenmerken van jouw systemen.

Onze vijf principes zijn:

√ Houd je datakwaliteit op orde
√ Zorg voor validatie van je data
√ Houd rekening met supply chain security
√ Maak je model robuust tegen aanvallen
√ Zorg dat je model controleerbaar (auditable) is

Houd je datakwaliteit op orde

De kwaliteit van je data is hoe dan ook van groot belang als je een AI-model of -systeem ontwikkelt. Met datakwaliteit bedoelen we onder andere: hoe gestructureerd is je data? Is bekend waar de data vandaan komt en kan je de kwaliteit controleren, weet je dat er niet mee gesjoemeld is? En ook: kan je elementen in je datasets ontdekken die een negatieve invloed hebben op de prestatie van je model?

2 Principles for the security of machine learning, NCSC UK. www.ncsc.gov.uk/collection/machine-learning

Daarnaast zijn er manieren om de data te bewerken, waardoor het lastiger is voor een aanvaller om iets aan je trainingsdata of input te manipuleren. Voorbeelden hiervan zijn *datatransformaties, gradient shaping* en *NULL-labeling*. Door veel aandacht te besteden aan je datakwaliteit, verbeter je de prestatie van je model en kun je onder andere poisoning en input aanvallen tegengaan.

Zorg voor validatie van je data

Als je data gebruikt uit externe bronnen, weet je niet altijd hoe die data tot stand is gekomen. Daarom is het belangrijk om de data goed te kunnen valideren. Hoe is de dataset tot stand gekomen? Hoe zorg ik ervoor dat ik niet te afhankelijk ben van deze enkele bron?

Bij extern verkregen data heb je vaak geen invloed op de data. Een data provider kan bijvoorbeeld besluiten andere labels toe te voegen, waardoor jouw model minder nauwkeurig wordt. Daarom is het belangrijk om dit continu te monitoren en waar mogelijk proactief te controleren.

Houd rekening met supply chain security

Zodra een kant-en-klaar model wordt gedownload of voor jou door anderen wordt ingericht, is de kans op bijvoorbeeld een backdoor reëel. Het tegengaan van een backdoor is zeer lastig als je het model niet zelf kunt doorgronden. Als je het model zelf kunt bouwen of beoordelen wordt de introductie van een backdoor veel lastiger. Daar is veel kennis, kunde en tijd voor nodig.

Soms is het niet mogelijk om zelf een model te bouwen, omdat je simpelweg niet de juiste soort of hoeveelheid data hebt. Of omdat je de toegang tot de rekenkracht mist die nodig is. In dat geval kijk je naar de garanties die er zijn om de leverancier te vertrouwen. Dat wordt ook wel *supply chain security* genoemd: zorgen dat je grip hebt op je toeleveranciers en de kwaliteit van de producten en diensten die zij leveren.

Je kunt dan bijvoorbeeld de extern binnengehaalde data of modellen steekproefsgewijs beoordelen op mogelijke fouten. Ook kun je technieken gebruiken om te zorgen dat onjuiste en/of data met verkeerde bedoelingen zo min mogelijk impact hebben op je model: zie daarvoor ook het principe over datakwaliteit. Maak verder zo veel mogelijk gebruik van producten en diensten van vertrouwde leveranciers.

Maak je model robuust tegen aanvallen

De robuustheid van je AI-model is de mate waarin het model goed kan functioneren bij afwijkende input, veranderingen in de data of pogingen tot misbruik. Alle voorgaande principes uit deze folder helpen maken je model al meer robuust. Dit principe voegt daar aan toe: zorg dat je je model traint tegen mogelijke aanvallen.

Bijvoorbeeld: met *adversarial training* kan je zorgen dat je model aangepaste, kwaadaardige data herkent en hier tegen bestand is. Hierdoor krijgt je model meer weerstand tegen adversarial input en raakt het hierdoor minder van slag. Of je kan detectiemechanismen ontwikkelen die gemanipuleerde input onderscheiden van gewenste input. Nog een optie is jouw model en trainingsdata beschermen tegen aanvallen door *rate limiting*, waarmee je beperkt hoe vaak een actie uitgevoerd kan worden binnen een bepaalde tijd.

Ten slotte adviseren wij om *red teaming* oefeningen met je AI-model te houden. Hierbij probeert een team van aanvallers (met meer of minder voorkennis van je model) kwetsbaarheden in het model te ontdekken en te misbruiken. Bijvoorbeeld met de vijf aanvalsmethoden die hierboven beschreven zijn. Red teaming wordt van oudsher toegepast in traditionele IT-infrastructuren en wordt steeds gangbaarder voor AI-modellen. Red teaming kan de zwakke punten in de beveiliging van jouw model tijdig ontdekken.

Zorg dat je model controleerbaar (auditable) is

Een AI-model geeft voorspellingen, maar vaak is het onduidelijk hoe deze voorspelling tot stand is gekomen. Zelfs als het model goed “werkt”, kun je moeilijk uitleggen waarom. Herkent jouw beeldherkenningsmodel echt de paarden op afbeeldingen, of wordt het model om de tuin geleid door een onopvallend watermerk? Wat maakt dat het model het ene type malware wel herkent, en het andere niet?

Als je bij het bouwen en trainen van je model vooraf rekening houdt met controleerbaarheid, dan zal het achteraf minder als een black box voor je zijn. Dit wordt ook wel *Explainable AI* genoemd. Bij simpele modellen kan je makkelijk achterhalen waarom bepaalde keuzes worden gemaakt. Bij meer geavanceerde modellen kan dit in zekere mate ook. Als je zelf begrijpt hoe je model werkt, kan je hier ook controles en testcasussen voor ontwikkelen.

Heb je vragen?

Loop je bij het veilig ontwikkelen van je AI-systeem tegen vragen aan? Bel ons op: 079-320 50 50 en vraag naar het NBV, of mail naar nbv-ai@minbzk.nl. We helpen je graag om jouw organisatie digitaal weerbaarder te maken.

De principes van het NBV helpen je om jouw AI-systeem in de context van beveiliging te zien - en passende maatregelen te nemen.

Algemene Inlichtingen- en Veiligheidsdienst
Postbus 20010 | 2500 EA Den Haag
T (079) 320 50 50

Februari 2023