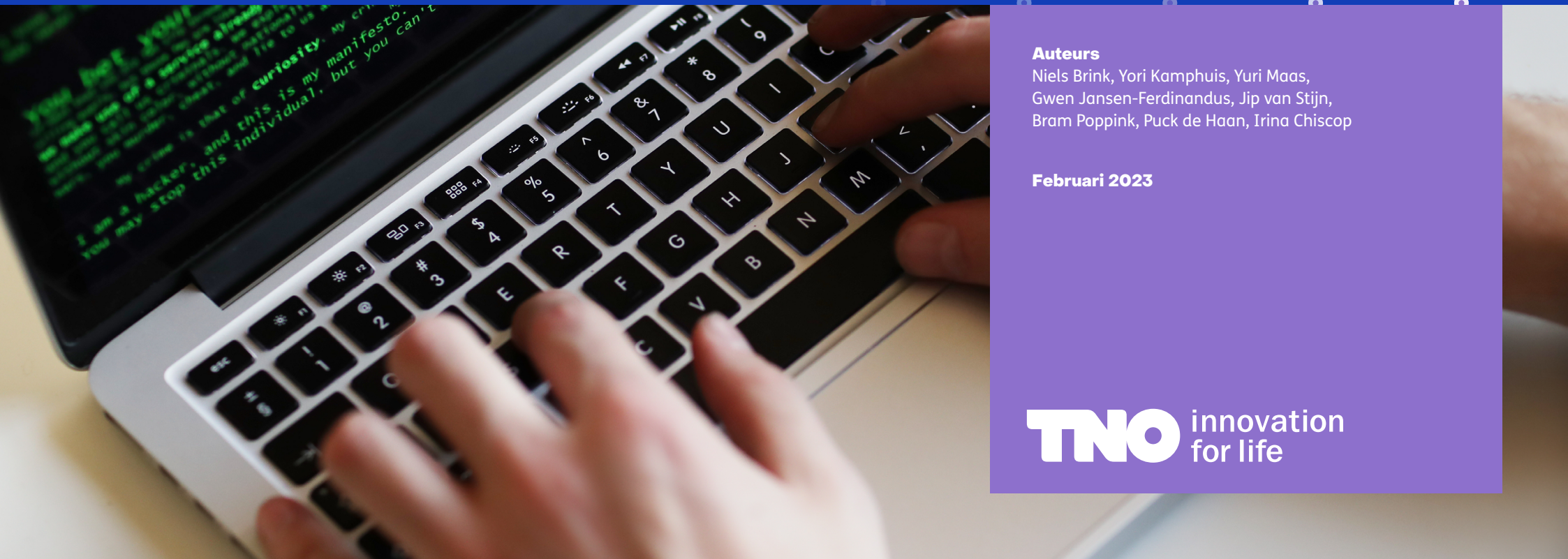
Whitepaper

Adversarial AI in het cyberdomein



Auteurs
Niels Brink, Yori Kamphuis, Yuri Maas,
Gwen Jansen-Ferdinandus, Jip van Stijn,
Bram Poppink, Puck de Haan, Irina Chiscop

Februari 2023

TNO innovation for life

Adversarial AI in het cyberdomein

Inhoud

Introductie

> “Kunstmatige Intelligentie wordt in alle internationale trendrapporten gezien als de belangrijkste disruptieve technologie van de komende jaren. Voor Defensie zal AI effect hebben op alle capaciteiten.”
>
> Ministerie van Defensie, 2020, p. 36

Welke dreigingen zijn er verbonden aan het gebruik van AI? Dat is waar TNO inzicht in geeft, aan de hand van het recente onderzoek naar de kwetsbaarheden van AI-toepassingen in het cyberdomein.

Artificial Intelligence (AI)-systemen gebruiken grote hoeveelheden data om beslissingen te maken in een complex systeem (AI HLEG, 2020). Wanneer de AI zelf leert van data, maakt het gebruik van Machine Learning (ML). Dat zijn computerprogramma’s die automatisch en efficiënt leren door ervaring op te doen (Mitchell, 1997).

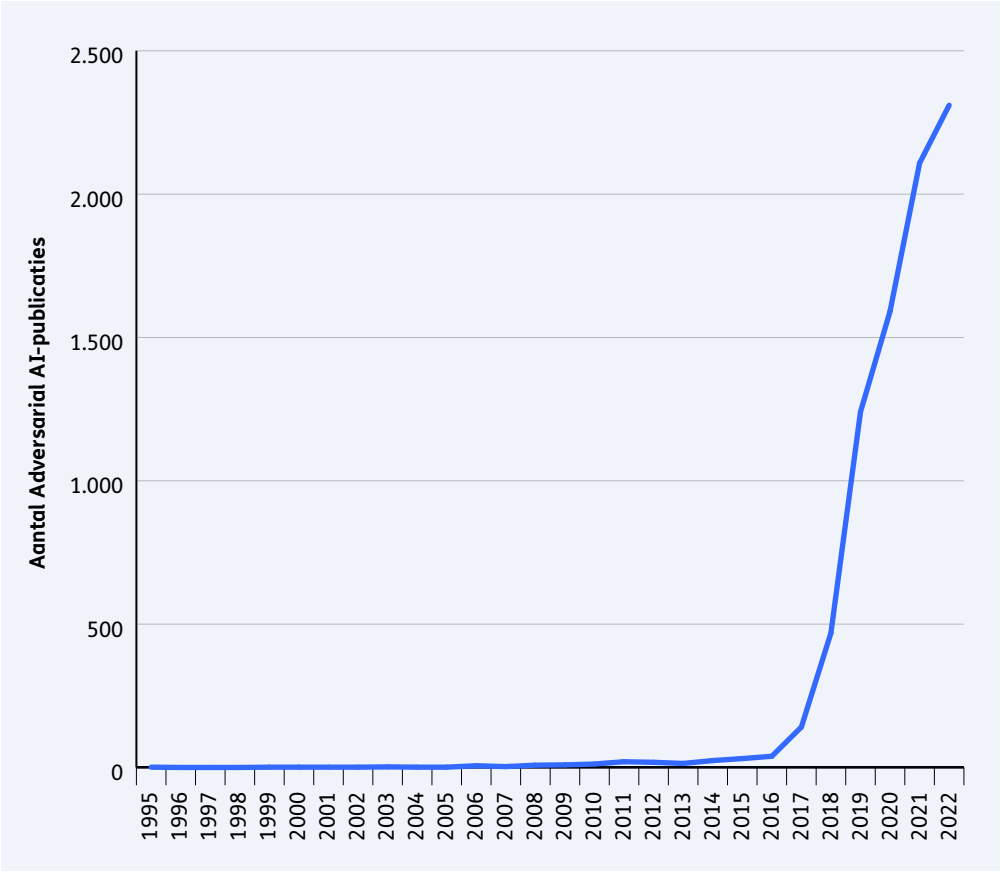
Naast civiele toepassingen heeft AI ook veel potentie in het veiligheidsdomein, aangezien een significant onderdeel van de activiteiten daar afhangen van het maken van beslissingen op basis van de juiste informatie (Swillens, 2022). Dit maakt de mogelijkheid van AI-toepassingen binnen defensie een relevante optie. Tegenwoordig heeft AI bijvoorbeeld al een belangrijke rol binnen het verzamelen van informatie, net als bij het besturen van (semi-)autonome voertuigen, zoals drones (Xue, Yuan, Wu, Zhang & Liu, 2020; Araya & King, 2022). Maar hoe robuust die AI-systemen tegen externe dreigingen zijn, dient zorgvuldig in kaart te worden gebracht alvorens ze grootschalig ingezet kunnen worden. TNO helpt hieraan mee door middel van onderzoek naar de stand van zaken op gebied van de robuustheid van AI-systemen. Dit artikel geeft een overzicht van de conclusies van dat onderzoek.

1 Wat is *Adversarial AI*?

Het *Adversarial AI*-onderzoeksveld houdt zich bezig met het bestuderen van de kwetsbaarheden in AI-systemen (Huang, Joseph, Nelson, Rubinstein & Tygar, 2011). *Adversarial AI* is al langer onderwerp van onderzoek, en dat heeft significante resultaten opgeleverd. Binnen het gebied van computervisie, dat computers aanleert om visuele beelden te herkennen (IBM, z.d.), is het onderzoekers bijvoorbeeld gelukt om AI-systemen die afbeeldingen classificeren te misleiden (Kurakin, Goodfellow & Bengio, 2017; Anley, 2022). Zodoende waren ze in staat om een schildpad te laten classificeren als een geweer (Athalye, Engstrom, Ilyas & Kwok, 2018).

Maar ondanks deze significante ontwikkelingen blijft het merendeel van het *Adversarial AI*-onderzoek gericht op de computervisie- en tekstdomeinen. Daarom heeft TNO's onderzoek zich gefocust op de toepassingen van *Adversarial AI* binnen het cybersecurity-domein. Binnen dat domein is het aantal gepubliceerde wetenschappelijke publicaties per jaar enorm toegenomen. Zoals Figuur 1 laat zien, is het aantal publicaties over *Adversarial AI* in het cyberdomein gegroeid van 24 in 2014 naar meer dan 2.400 in 2021. En in 2022 verschenen er zelfs zo'n negen publicaties per dag, wat weer een dertig procent toename is vergeleken met 2021[1].

Echter bestaan er rond de huidige onderzoeken nog relatief veel onzekerheden. Dit was aanleiding voor TNO om een overzicht van de toepassingen van *Adversarial AI*, en de verdedigingsmaatregelen, binnen het cybersecurity-domein te documenteren. Dit artikel maakt in die volgorde duidelijk wat de staat is van het *Adversarial AI*-onderzoeksveld.



Figuur 1. *Adversarial AI*-publicaties binnen het cybersecurity-domein per jaar[2].

2 Wat zijn *Adversarial AI*-aanvallen?

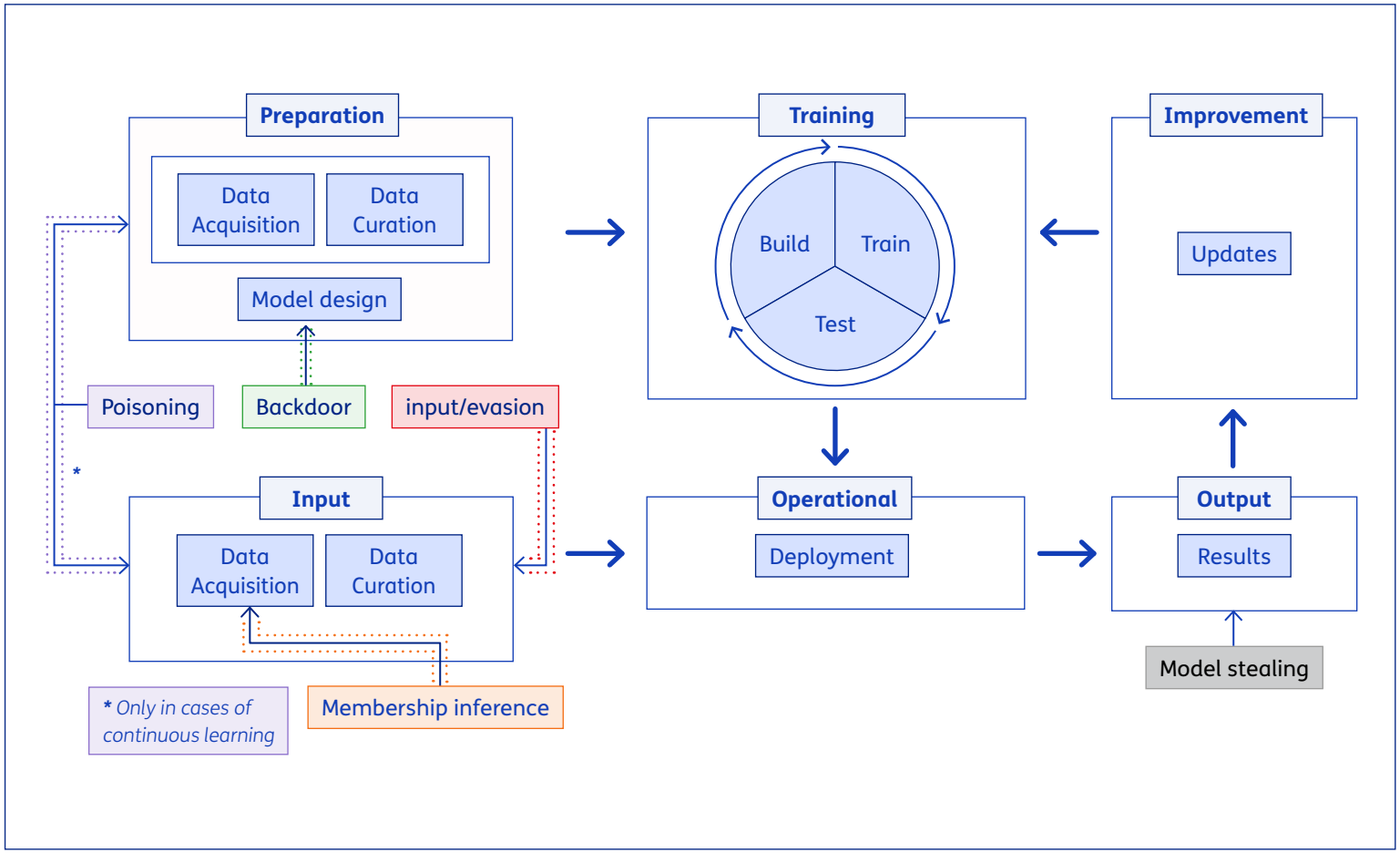
Onderzoek maakt één ding duidelijk: AI-systemen zijn al kwetsbaar voor bestaande aanvallen (Huang, Joseph, Nelson, Rubinstein & Tygar, 2011). Maar daarvoor heeft de aanvaller wel een bepaalde hoeveelheid kennis over het AI-systeem nodig. Aan de hand van dat kennisniveau zijn aanvallen onder te verdelen in vier categorieën met toenemende complexiteit (Rosenberg, Shabtai, Elovici & Rokach, 2021):

- **Black-box:** wanneer de aanvaller weinig tot geen informatie heeft over de architectuur, het model, en de data van het ML-model.
- **Gray-box:** wanneer de aanvaller beperkte kennis heeft over het ML-model of de trainingsdata.
- **White-box:** wanneer de aanvaller volledige kennis heeft over het ML-model, inclusief de architectuur en de trainingsdata.
- **Transparent-box:** wanneer de aanvaller volledige kennis heeft over het model, de architectuur en de data, en daarbovenop ook kennis heeft over de defensieve maatregelen die aanwezig zijn om de robuustheid van het model te verbeteren.

3 Op welke manieren kunnen aanvallers AI-systemen bestoken?

Dat aanvallen op AI-systemen mogelijk zijn is bewezen (Huang, Joseph, Nelson, Rubinstein & Tygar, 2011), maar over hoe deze aanvalsmethoden ingedeeld moeten worden is de wetenschappelijke gemeenschap het nog niet helemaal eens. Op basis van een door TNO uitgevoerde analyse van de belangrijkste publicaties kan een onderscheid worden gemaakt tussen vijf soorten aanvallen (Figuur 2).

Daarin worden de fases van een ML-model weergegeven, zoals gedefinieerd in ETSI (2020). Daarbovenop heeft TNO de vijf soorten aanvallen gepositioneerd op de plek waar ze het ML-model bestoken. Deze worden in dit hoofdstuk behandeld.



Figuur 2. Aanvalsmogelijkheden op de *ML cycle* (eigen werk, aanvulling op ETSI, 2020, p. 11).

***Poisoning*-aanvallen**, aangegeven in het paars in Figuur 2, omvatten het manipuleren (toevoegen, weghalen of veranderen) van trainingsdata om zo de kans op misclassificaties door het model te verhogen (Biggio & Roli, 2018). Aanvallers kunnen dit soort aanvallen uitvoeren door in de trainingsdata de grenzen tussen de verschillende soorten objecten zo aan te passen dat bepaalde objecten in een verkeerde categorie zullen vallen wanneer ze aan het model gepresenteerd worden. In het geval van de schildpadaanval zou de aanvaller de trainingsdata zo kunnen manipuleren dat het model bepaalde kenmerken oppikt als eigenschappen van een geweer. Wanneer de aanvallers een schildpad met die kenmerken vervolgens als input geven, is er een grotere kans dat het model het als een geweer classificeert.

Wanneer aanvallers een ***backdoor*-aanval** uitvoeren voegen ze een stuk code toe aan het model waardoor die normaal blijft functioneren totdat er een input binnenkomt die aan de door de aanvallers gespecificeerde set kenmerken voldoet. In dat geval wordt de *backdoor* geactiveerd en zal het model een output geven die de aanvallers wensen. Hoewel dit aanvalstype momenteel vooral gelimiteerd is tot theoretische toepassingen, heeft recent onderzoek de mogelijkheid van (onder bepaalde condities) onzichtbare *backdoors* bewezen (Goldwasser, Kim, Vaikuntanathan & Zamir, 2022). Dit aanvalstype staat aangegeven in het groen in Figuur 2. In het geval van de schildpadaanval zouden de aanvallers een *backdoor* kunnen inbouwen in de trainingsdata die activeert als een aantal specifieke kenmerken van hun schildpad gepresenteerd worden aan het model. Dan zou het model deze specifieke schildpad classificeren als een geweer, maar blijft het model normaal functioneren wanneer andere schildpadden worden gepresenteerd.

Aanvallers zouden ook het model kunnen misleiden door hun input te manipuleren, wat bekend staat als een ***input/evasion*-aanval**, aangegeven in het rood in Figuur 2 (Sadeghi, Banerjee & Gupta, 2020; Athalye, Engstrom, Ilyas & Kwok, 2018). Zodoende zouden ze ervoor kunnen zorgen dat een model niet correct functioneert en hun potentieel kwaadaardige input doorlaat. De aanvaller zouden dit kunnen doen door hun eigen input met kleine verstoringen (perturbaties) aan te passen. Denk hierbij aan een minieme verandering in de kleursamenstelling van de schildpad of een toegevoegde laag ruis over een foto. Hoewel een mens gemakkelijk zou kunnen zien dat de schildpad nog steeds een schildpad is, kan dit al genoeg zijn voor een ML-model om het te classificeren als een geweer.

Bij de voorgaande aanvalstechnieken is het doel van de aanvallers om een bepaalde uitkomst te forceren. Een ander doel kan zijn om bepaalde aspecten te leren over een ML-model. Deze aanvalssoort staat bekend als een ***membership inference-aanval***, aangegeven in het oranje in Figuur 2. Bij dit type aanval gebruiken de aanvallers technieken om te bepalen op welke data het model getraind is (NIST, 2019; Microsoft, 2021; AI HLEG, 2020). Dit kan enorme gevolgen hebben wanneer die trainingsdata gevoelige informatie bevat, omdat de aanvallers zo potentieel toegang zouden kunnen krijgen tot die gevoelige informatie. Een model zou bijvoorbeeld getraind kunnen zijn om bepaalde medische aandoeningen te herkennen. Als die trainingsdata informatie bevat over echte patiënten, zou het van groot belang zijn om te voorkomen dat die data zou kunnen lekken.

Als laatste zouden aanvallers een **eigen kopie van het ML-model** kunnen bouwen door hun toegang tot het originele model te misbruiken (ETSI, 2020; Microsoft, 2021; NIST, 2019). Dit aanvalstype is aangegeven in het grijs in Figuur 2. Als de aanvallers een belangrijk model dat niet vrij beschikbaar is weten te reproduceren, zou dat een significante impact kunnen hebben op de veiligheid van het originele model omdat de aanvallers hun eigen versie zouden kunnen gebruiken om te bepalen welke aanval het meest effectief zou zijn tegen het originele model (Anley, 2022). Wanneer het originele model in een defensiecontext wordt gebruikt, zou dit enorme gevolgen kunnen hebben omdat die modellen over het algemeen ingezet worden voor zeer impactvolle beslissingen. Denk hierbij bijvoorbeeld aan een model dat een drone helpt om vijandelijke posities te identificeren. De belangen rondom de inzet van een AI-systeem in een defensiecontext wegen dus zwaar.

Wat deze aanvalsmethodes duidelijk maken is dat een ML-model in iedere fase van de levensduur kwetsbaar is. Tijdens de ontwikkeling van het model kunnen aanvallers de trainingsdata manipuleren of een *backdoor* inbouwen. En wanneer het model is uitgebracht vormt de toegang ertoe weer een andere kwetsbaarheid die de aanvallers kunnen exploiteren door het model te misleiden of zelfs te reproduceren.

4 Welke verdedigingsmaatregelen zijn er?

Nu AI-systemen een steeds prominentere rol gaan spelen in het veiligheidsdomein, is het des te belangrijker dat de veiligheid daarvan ook voor zover mogelijk gegarandeerd kan worden. Ondanks het vele onderzoek dat gepubliceerd wordt over *Adversarial AI*, blijven de aanwezigheid, inrichting en effectiviteit van verdedigingsmaatregelen in de theorie en daarmee ook de praktijk achter. En hoewel er voor de aanvalsmethodes nog een overzicht geproduceerd kon worden, is dit lastig voor de defensieve zijde van *Adversarial AI* omdat bestaande openbare publicaties vaak onduidelijk zijn of uiteenlopende terminologie gebruiken.

Desondanks kunnen we verdedigingsmaatregelen wel opsplitsen in technische en tactische maatregelen, die samenvallen onder het *counter AI*-onderzoeksveld, dat focust op hoe AI-kwetsbaarheden gebruikt kunnen worden om de onderliggende systemen te exploiteren. Want hoewel de *Adversarial AI*-aanvalsmethoden grotendeels technisch zijn, bestaan er naast technische ook niet-technische verdedigingsmaatregelen. Voorbeelden hiervan zijn om mensen in te zetten naast AI-systemen of om samenwerking in de ontwerpfase van een model te verbeteren om de bouwstenen van het model beter te beveiligen (Hoffman, 2021). Voor een goede verdedigingsstrategie moeten technische en tactische maatregelen allebei meegenomen worden (Thomas, 2020).

Het verschil in de staat van de ontwikkelingen tussen het onderzoek naar aanvalsmethodes en verdedigingsmaatregelen is mogelijk veroorzaakt doordat *Adversarial AI* pas relatief recentelijk is opgekomen waardoor de wetenschappelijke gemeenschap zich ook pas recentelijk heeft kunnen focussen op de relevante verdedigingsmaatregelen. Daarnaast kan dit verschil komen doordat het onderzoek naar verdedigingsmaatregelen volgt op het onderzoek naar aanvalsmethodes. Maar aangezien dat overzicht nog niet compleet is, hindert dit ook het onderzoek naar verdedigingsmaatregelen.

5 Hoe kan *Adversarial AI* toegepast worden in cyberspace?

Zoals de beschrijving van de aanvalsmethodes duidelijk maakt, zijn *Adversarial AI*-aanvallen al mogelijk in de computervisie- en tekstdomeinen. Echter verschillen die domeinen substantieel van het cybersecurity-domein (Rosenberg, Shabtai, Elovici & Rokach, 2021), wat de vraag oproept hoezeer *Adversarial AI*-aanvalsmethodieken uit die domeinen ook toepasbaar zijn in het cybersecurity-domein.

Om die vraag te beantwoorden moeten de verschillen tussen de computervisie- en tekstdomeinen en het cybersecurity-domein uitgelegd worden. Ten eerste is het belangrijk om te onthouden dat cyberspace functioneert door middel van protocollen, oftewel de regels die bepalen hoe computers communiceren, zoals het IP-protocol. Wanneer een aanvaller een aanval wil uitvoeren, moeten ze er dus voor zorgen dat de *payload*, het pakketje data dat de aanval omvat, zelf ook voldoet aan die protocollen.

Dit staat in contrast met aanvallen in het computervisiedomein, waarin het mogelijk is om een pixel op een minimale manier te modificeren zonder dat dit de functie van de foto zelf beïnvloedt. Dat betekent dat aanvallers manieren moeten ontwikkelen waarmee hun aanval voldoet aan de protocollen én waarmee ze de *payload* intact kunnen houden. De verscheidenheid aan datatypen en databronnen in het cybersecurity-domein betekent ook dat het lastiger is voor algoritmes om hun functie te vervullen, wat in het voordeel van de aanvallers werkt. Deze diversiteit verbreedt namelijk het aanvalslandschap. Ook compliceert dit het ontwikkelen van aanvallen omdat een aanval die gericht is op één model niet zondermeer gebruikt kan worden om andere modellen aan te vallen, vanwege de verschillende datatypen die de modellen gebruiken. Door deze verschillen zijn aanvalsmethodes die zijn ontwikkeld voor andere domeinen niet automatisch toepasbaar in het cybersecurity-domein.

Een concreet voorbeeld van een type ML-algoritme in het cybersecurity-domein dat kwetsbaar is voor *adversarial*-aanvallen is een detector die kwaadaardige internetdomeinen herkent die aanvallers gebruiken om *malware*-aanvallen te besturen. Met behulp van *Adversarial AI*-technieken kunnen aanvallers in hoog tempo domeinnamen genereren die deze detectoren kunnen omzeilen (Sivaguru, Choudhary, Yu & Tymchenko, 2018). Deze aanvalsalgoritmen worden Domein Generatie Algoritmes (DGA's) genoemd. TNO's recente onderzoek naar het verbeteren van de robuustheid van algoritmes tegen deze *Adversarial AI* DGA's heeft uitgewezen dat het detectiealgoritme verbeterd kan worden door kwaadaardige voorbeelden toe te voegen aan de trainingsdata. Het idee daarachter is dat een model beter in staat is om kwaadaardige domeinen te herkennen als het getraind wordt op kwaadaardige voorbeelden.

Daarbij is het van belang om een goede afweging te maken welke data hiervoor gebruikt wordt. Een algoritme dat namelijk wordt getraind om één type kwaadaardige domeinen te herkennen kan over het algemeen namelijk minder goed andere soorten kwaadaardige domeinen herkennen (Anley, 2022). Een essentieel onderdeel van het selecteren van geschikte trainingsdata is daarom dat de dataset een representatief voorbeeld is van de context waarin het algoritme uiteindelijk wordt ingezet. Dit is te vergelijken met een Formule 1-auto, die vliegensvlug rond een circuit kan rijden maar nutteloos is in het dagelijks leven. Voor het selecteren van kwaadaardige voorbeelden is alleen nog geen methode ontwikkeld. Dit is waar toekomstig onderzoek zich dan ook op moet focussen.

6 Wat kunnen we concluderen?

Adversarial AI is een dreiging die zich buitengewoon actief en snel aan het ontwikkelen is. Gezien het toenemende gebruik van AI betekent dit ook dat steeds meer producten potentieel kwetsbaar worden voor dergelijke aanvallen.
Echter mist bij velen het besef van deze verontrustende ontwikkelingen en loopt de verdediging tegen *Adversarial AI* achter op de aanvalsmogelijkheden. Deze onbalans tussen de aanvalsmogelijkheden enerzijds en de verdedigingsmaatregelen anderzijds dient opgelost te worden zodat nieuwe AI-systemen veilig uitgerold kunnen worden, zeker binnen de vitale sectoren. Daarom verdient het verbeteren van de robuustheid van AI-systemen nu de aandacht. Dit zorgt ervoor dat de mogelijkheden van AI op een veilige manier gerealiseerd kunnen worden.

Op het onderzoeksgebied wordt veel voortgang geboekt op het gebied van het ontwikkelen van, en verdedigen tegen, *Adversarial AI*-aanvallen. Hoewel dit positief is, maakt dit het ook lastiger om het kennisniveau te bepalen doordat het lastig is om een selectie te maken van de meest actuele en relevante literatuur. Dit is terug te zien in de literatuur, waarin de auteurs veelal overlappende classificaties voorstellen terwijl bepaalde aspecten dikwijls missen. Daarnaast heeft TNO's onderzoek richting het verbeteren van de robuustheid van modellen ook uitgewezen dat het verbeteren inderdaad mogelijk is. Dit kan bijvoorbeeld door het model te trainen op kwaadaardige voorbeelden, waarbij de selectie van de kwaadaardige voorbeelden zorgvuldig gedaan moet worden.

Waar het *Adversarial AI*-onderzoeksveld zich verder kan ontwikkelen is het scheppen van een eenduidig overzicht van verdedigingsmaatregelen. Zoals we hebben laten zien, is het mogelijk om een breed gedragen overzicht te creëren voor de aanvalstechnieken. Maar dit geldt niet voor de verdedigingsmaatregelen, waar het onderzoek, van nature, nog achterloopt op de recente ontwikkelingen op het aanvalsgebied. Daarom is het van belang dat dit essentiële onderdeel meer aandacht verdient zodat ML-modellen beter bestand worden gemaakt tegen *Adversarial AI*.

De resultaten van TNO's onderzoek laten zien dat *Adversarial AI* van toepassing is in het cybersecurity-domein en dat experimenten met *Adversarial AI*-technieken binnen dit domein relevante inzichten kunnen produceren voor toepassingen binnen andere domeinen.

Literatuurlijst

AI HLEG. (2020). High-Level Expert Group on Artificial Intelligence - The Assessment List For Trustworthy Artificial Intelligence (ALTAI). Brussels: European Commission. doi: https://doi.org/10.2759/002360.

Anley, C. (2022, July 6). Whitepaper – Practical Attacks on Machine Learning Systems. Retrieved from NCC Group: https://research.nccgroup.com/2022/07/06/whitepaper-practical-attacks-on-machine-learning-systems/.

Araya, D. & King, M. (2022). The Impact of Artificial Intelligence on Military Defence and Security. Waterloo, ON, Canada: Centre for International Governance Innovation (CIGI).

Athalye, A., Engstrom, L., Ilyas, A. & Kwok, K. (2018). Synthesizing Robust Adversarial Examples. The 35th International Conference on Machine Learning, (pp. 1-19). Stockholm. Retrieved from arXiv.

Biggio, B. & Roli, F. (2018, December). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, pp. 317-331.

Dalvi, N., Domingos, P., Mausam, Sanghai, S. & Verma, D. (2004). Adversarial Classification. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 99-108). Seattle: ACM Press.

ETSI. (2020). Securing Artificial Intelligence (SAI): Problem Statement. Sophia Antipolis Cedex, France: ETSI.

Goldwasser, S., Kim, M.P., Vaikuntanathan, V. & Zamir, O. (2022). Planting Undetectable Backdoors in Machine Learning Models. arXiv preprint arXiv: 2204.06974.

Hoffman, W. (2021). Making AI Work for Cyber Defense: The Accuracy-Robustness Tradeoff. Center for Security and Emerging Technology: CSET.

Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I. & Tygar, J.D. (2011). Adversarial Machine Learning. Proceedings of 4th ACM Workshop on Artificial Intelligence and Security, (pp. 43-58).

IBM. (z.d.). What is computer vision? Retrieved from https://www.ibm.com/topics/computer-vision.

Kurakin, A., Goodfellow, I.J. & Bengio, S. (2017). Adversarial examples in the physical world. 5th International Conference on Learning Representations, (pp. 1-14). Toulon.

Microsoft. (2021, October). Digital Defense Report October 2021.

Ministerie van Defensie. (2020, november 27). Strategische Kennis- en Innovatieagenda (SKIA) 2021-2025. Retrieved from Ministerie van Defensie: https://www.defensie.nl/downloads/publicaties/2020/11/25/strategische-kennis--en-innovatieagenda-2021-2025.

Mitchell, T. (1997). Machine Learning. New York: McGraw-hill.

NIST. (2019). Draft NISTIR 8269 - A taxonomy and terminology of Adversarial Machine Learning. NIST.

Rosenberg, I., Shabtai, A., Elovici, Y. & Rokach, L. (2021). Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. ACM Computing Surveys (CSUR), 54(5), 1-36. doi: https://doi.org/10.1145/3453158.

Eindnoten

Sadeghi, K., Banerjee, A. & Gupta, S.K. (2020). A system-driven taxonomy of attacks and defenses in adversarial machine learning. IEEE transactions on emerging topics in computational intelligence, 450-467.

Sivaguru, R., Choudhary, C., Yu, B. & Tymchenko, V. (2018). An Evaluation of DGA Classifiers. 2018 IEEE International Conference on Big Data (pp. 5058-5067). Seattle: IEEE.

Swillens, J. (2022). Speech. MIVD-seminar: Fog of War 2.0 (p. 11). Campus Wijnhaven, Den Haag: MIVD.

Thomas, M. (2020). Time for a Counter-AI Strategy. Strategic Studies Quarterly, 14, 3-8.

Xue, M., Yuan, C., Wu, H., Zhang, Y. & Liu, W. (2020). Machine Learning Security: Threats, Countermeasures, and Evaluations. IEEE Access, 8, pp 74720-74742. doi: https://doi.org/10.1109/ACCESS.2020.2987435.

1 Bron: www.scopus.com met de zoekterm “adversarial machine learning”

2 Bron:
   1) Scopus met de zoekterm: “(ALL (adversarial AND machine AND learning) AND (cyber) AND (LIMIT-TO (DOCTYPE), ‘ar’) OR LIMIT-TO (DOCTYPE), ‘cp’) OR LIMIT-TO (DOCTYPE), ‘re’)” en
   2) Google Scholar met de zoekterm: ‘adversarial machine learning’ AND ‘cyber’, vanaf 2014.

Auteurs

Niels Brink
niels.brink@tno.nl

Yori Kamphuis
yori.kamphuis@tno.nl

Yuri Maas
yuri.maas@tno.nl

Gwen Jansen-Ferdinandus
gwen.ferdinandus@tno.nl

Jip van Stijn
jip.vanstijn@tno.nl

Bram Poppink
bram.poppink@tno.nl

Puck de Haan
puck.dehaan@tno.nl

Irina Chiscop
Niet meer werkzaam bij TNO

TNO innovation for life

Context

Deze publicatie komt voort uit het *counter-AI*-onderzoek binnen het TNO cluster Cyber & Electronic Warfare en sluit aan op REAIM 2023, de eerste wereldwijde top over verantwoorde kunstmatige intelligentie in het militaire domein met de Nederlandse regering als organisator.

tno.nl/digitaleweerbaarheid