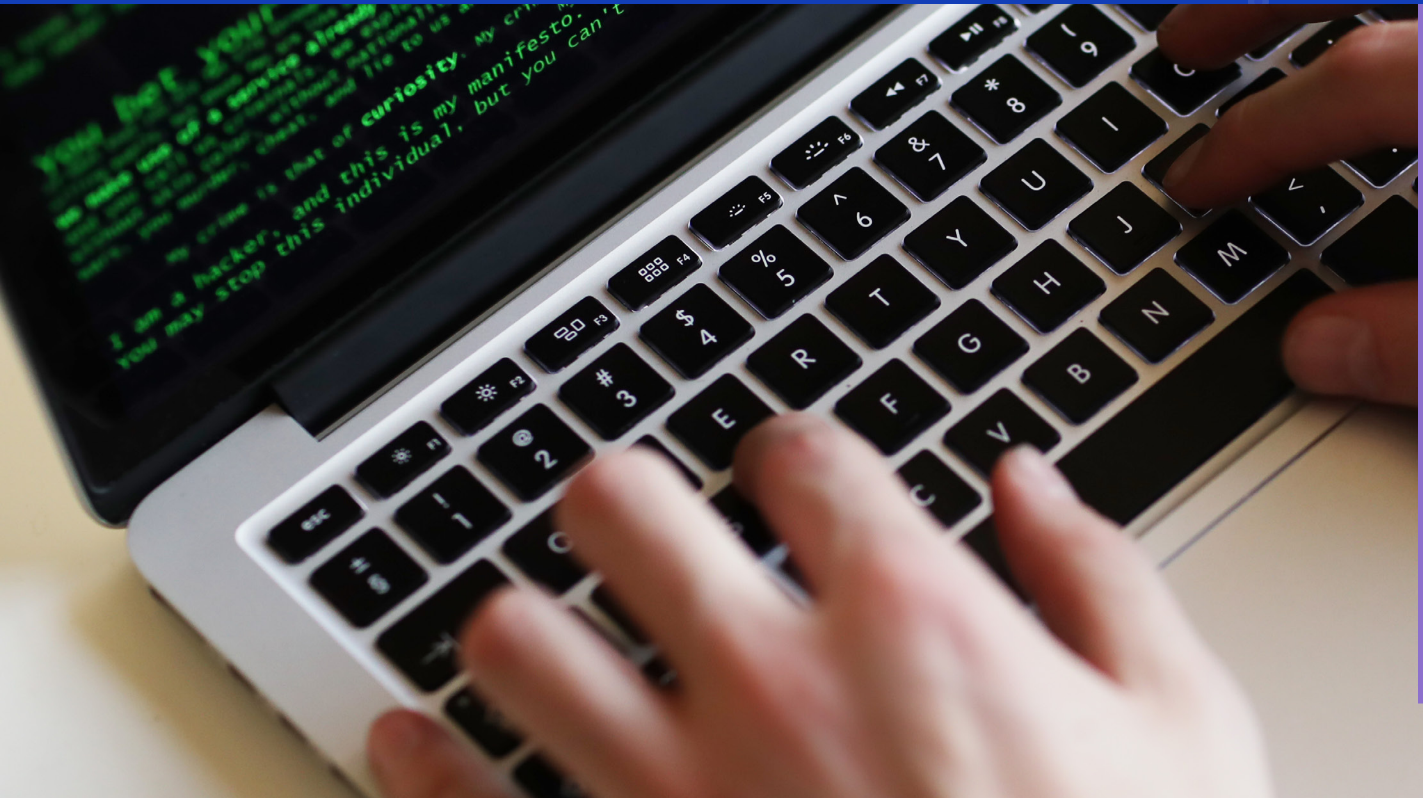
White paper

Adversarial AI in the cyber domain



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

February 2023

TNO innovation for life

Adversarial AI in the cyber domain

Contents

Introduction

> “Artificial Intelligence is identified in all international trend reports as the key disruptive technology of the next few years. AI will have an impact on all Defence capabilities.”
>
> Ministry of Defence, 2020, p. 36

What threats are associated with the use of AI? This is a question that TNO seeks to answer through its recent research into the vulnerabilities of AI applications in the cyber domain.

Artificial Intelligence (AI) systems use large amounts of data to make decisions in a complex system (AI HLEG, 2020). In order for an AI system to learn specialized tasks, such as discrimination of the different elements within the complex system it operates in (also known as classification), Machine Learning (ML) is applied. These are computer programmes that learn automatically and efficiently through experience (Mitchell, 1997).

Besides civilian applications, AI also has a lot of potential in the security domain, as a significant proportion of activities in that area depend on making decisions based on the right information (Swillens, 2022). AI systems are therefore a relevant option to consider for Defence sector applications. For instance, AI currently already plays an important role in information gathering, as well as in driving autonomous and semi-autonomous vehicles such as drones (Xue, Yuan, Wu, Zhang & Liu, 2020; Araya & King, 2022). However, proper analysis of the ability of these AI systems to withstand external threats is essential before they can be deployed on a large scale. TNO is contributing by researching the state of the art when it comes to AI system robustness. This article summarises the conclusions of that research.

1 What is Adversarial AI?

The Adversarial AI field of research studies the vulnerabilities in AI systems (Huang, Joseph, Nelson, Rubinstein & Tygar, 2011). Adversarial AI is a subject of research that has acquired a lot of traction and scientific attention over the past years, with significant results. For example, in the field of computer vision, which teaches computers to recognise visual images (IBM, n.d.), researchers have succeeded in misleading AI systems that classify images (Kurakin, Goodfellow & Bengio, 2017; Anley, 2022). This enabled them to have a turtle classified - or better put, misclassified - as a gun (Athalye, Engstrom, Ilyas & Kwok, 2018).

But despite these significant developments, most Adversarial AI research remains focused on the computer vision and text domains. That is why TNO's research has focused on Adversarial AI applications in the cyber security domain. Within that domain, the number of published scientific papers per year has increased dramatically. As shown in Figure 1, the number of publications on Adversarial AI in the cyber domain has risen from 24 in 2014 to more than 2,400 in 2021. And in 2022, there were on average an astonishing amount of nine publications per day: a further 30% increase compared to 2021[1].

Though the number of scientific publications has increased immensely, there are many subareas to explore within this field of research. Moreover, probably caused by the enormous collection of scientific works, there is still a lack of structure, especially for defensive measures. This has prompted TNO to produce an overview of Adversarial AI applications and defensive measures within the cyber security domain. This article provides clarification, in that order, on the current state of the Adversarial AI field of research.

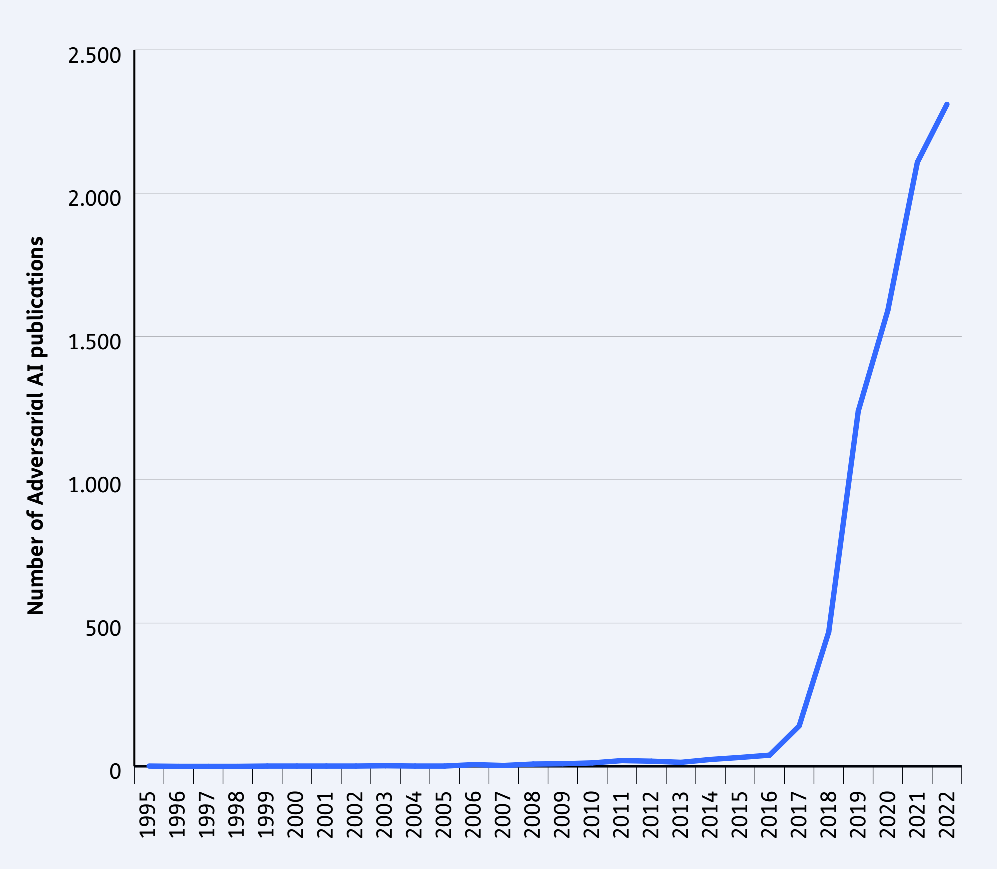


Figure 1: Adversarial AI publications in the cyber security domain per year[2].

2 What are Adversarial AI attacks

Previous research has proven one thing: AI systems are already vulnerable to existing attacks (Huang, Joseph, Nelson, Rubinstein & Tygar, 2011). However, the success of an attack depends on the attacker having a certain amount of knowledge about the AI system. Based on that level of knowledge, attacks can be divided into four categories of increasing complexity (Rosenberg, Shabtai, Elovici & Rokach, 2021):

- **Black box:** when the attacker has little or no information about the architecture, model, and data of the ML model.
- **Gray box:** when the attacker has limited knowledge of the ML model or training data.
- **White box:** when the attacker has full knowledge of the ML model, including its architecture and training data.
- **Transparent box:** when the attacker has full knowledge of the model, architecture and data, plus knowledge of the defensive measures taken to improve the robustness of the model.

3 How can attackers target AI systems?

The possibility of attacks on AI systems has been proven (Huang, Joseph, Nelson, Rubinstein & Tygar, 2011), however the scientific community is yet to reach consensus on how these methods of attack should be classified. Based on TNO's analysis of key publications, it is possible to identify five different types of attack: poisoning, backdoor, input/evasion, membership inference, and model stealing (Figure 2).

The Figure shows the phases of an ML model, as defined in ETSI (2020). In this model, TNO has added the five types of attacks where they target the ML model. These are discussed in this chapter.

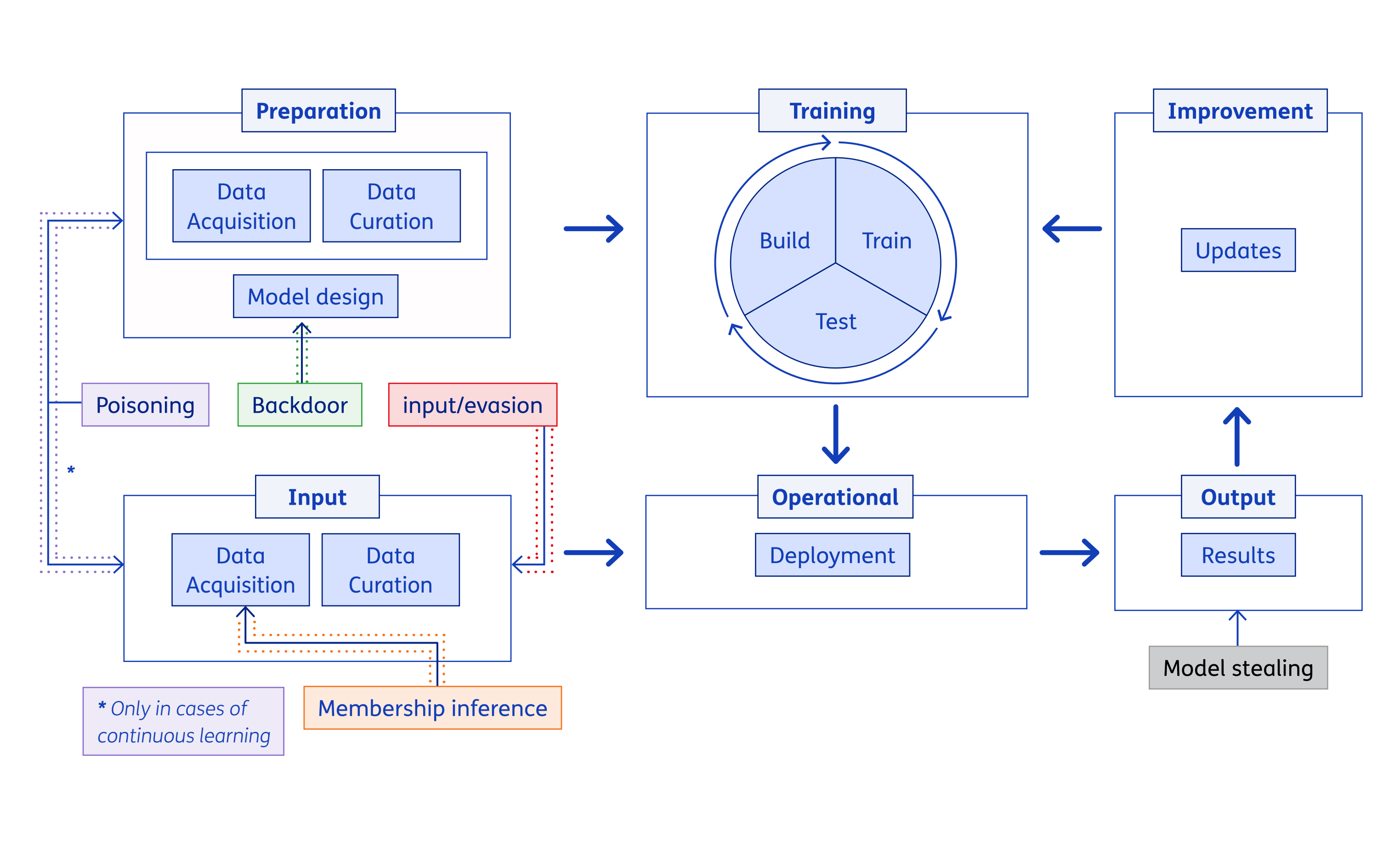


Figure 2: Possibilities to attack the ML life cycle (own work, supplement to ETSI, 2020, p. 11).

Poisoning attacks, shown in purple in Figure 2, involve manipulating (adding, removing or changing) training data to increase the likelihood of misclassifications by the model (Biggio & Roli, 2018). Attackers can carry out this kind of attack by adjusting the boundaries between object types in the training data so that certain objects will fall into the wrong category when presented to the model. In the case of the turtle attack, the attacker could manipulate the training data so that the model picks up certain characteristics as features of a gun. When the attackers then provide a turtle with those characteristics as the input, there is a greater likelihood that the model will classify it as a gun.

In a **backdoor attack**, attackers add a piece of code to the model that keeps it functioning normally until an input is received that meets the set of characteristics specified by the attackers. When this happens, the backdoor is triggered and the model will produce the attackers' desired output. Although this type of attack is currently mainly limited to theoretical applications, recent research has proven the possibility (under certain conditions) of invisible backdoors (Goldwasser, Kim, Vaikuntanathan & Zamir, 2022). This type of attack is shown in green in Figure 2. In the case of the turtle attack, the attackers could build a backdoor into the model that is triggered when a number of specific characteristics (a signature) in the input data (a picture of a turtle) is presented to the model. The model would then classify this particular turtle as a gun, but continue to function normally when other turtles are presented.

Attackers could also mislead the model by manipulating their input, known as an **input/evasion attack**, shown in red in Figure 2 (Sadeghi, Banerjee & Gupta, 2020; Athalye, Engstrom, Ilyas & Kwok, 2018). By doing so, they could prevent a model from functioning properly and let their potentially malicious input through. The attackers could do this by modifying their own input with small changes (perturbations). Examples include a minimal change in the colour composition of the turtle or an added layer of noise over an image. While a human could easily see that the turtle is still a turtle, such modifications can be enough for an ML model to classify it as a gun.

In the previous attack techniques, the attackers' aim is to force a specific outcome. Another goal may be to learn certain aspects about the training data used to train an ML model. This type of attack is known as a **membership inference attack**, shown in orange in Figure 2. In this type of attack, the attackers use techniques to identify characteristics of the data used to train the model (NIST, 2019; Microsoft, 2021; AI HLEG, 2020). This can have significant implications if the training data contains sensitive information, as it could potentially allow attackers to access that sensitive information. For example, a model could be trained to recognise certain medical conditions. If that training data contains information about real patients, it would be imperative to prevent a potential breach of that data.

Finally, attackers could build **their own copy of the ML model** by abusing their access to the original model (ETSI, 2020; Microsoft, 2021; NIST, 2019). This type of attack is shown in grey in Figure 2. If attackers succeed in reproducing an important model that is not freely available, this could have a significant impact on the security of the original model because the attackers could use their own version to determine what type of attack would be most effective against the original model (Anley, 2022). If the original model is used in a defence context, this could have massive implications because those models are generally deployed for high-impact decisions. An example would be a model that helps a drone identify enemy positions. The stakes are therefore high when it comes to the deployment of an AI system in a defence context.

What these methods of attack clearly illustrate is that an ML model is vulnerable at every stage of its lifecycle. During the development of the model, attackers can manipulate the training data or build in a backdoor. What's more, once the model has been released, access to it constitutes yet another vulnerability that attackers can exploit by misleading or even reproducing the model.

4 What defensive measures can be taken?

Now that AI systems are starting to become more prominent in the security domain, it is all the more important that the security of these systems can also be guaranteed as far as possible. Despite the large volume of research published on Adversarial AI, the presence, design and effectiveness of defensive measures lag behind in theory and therefore also in practice. And while TNO was able to produce an overview of methods of attack, this proved difficult for the defensive side of Adversarial AI because existing publications are often inconclusive or use differing terminology.

Nevertheless, we can divide defensive measures into technical and tactical measures. These measures coalesce in the counter AI field of research, which focuses on how AI vulnerabilities can be used to exploit underlying systems. After all, although the Adversarial AI methods of attack are largely technical, defensive measures can be both technical and non-technical. Examples include the use of humans alongside AI systems or improving collaboration in the design phase of a model to better secure the model's building blocks (Hoffman, 2021). A good defensive strategy needs to include both technical and tactical measures (Thomas, 2020).

The difference in development progress between research on methods of attack and defensive measures may be due to the fact that Adversarial AI is a relatively recent phenomenon, which means that the scientific community has only recently been able to focus on the relevant defensive measures. Another reason for this difference could be that research on defensive measures follows on from research into methods of attack. Since that overview is not yet complete, however, research into defensive measures is also hampered.

5 How can Adversarial AI be applied in cyberspace?

As the description of the methods of attack clearly shows, Adversarial AI attacks are already possible in the computer vision and text domains. However, those domains differ substantially from the cyber security domain (Rosenberg, Shabtai, Elovici & Rokach, 2021), which raises the question of the extent to which Adversarial AI attack methods from those domains are also applicable in the cyber security domain.

To answer that question, the differences between the computer vision and text domains and the cyber security domain need to be explained. First and foremost, it is important to remember that cyberspace operates on the basis of protocols, or rules that govern how computers communicate, such as the IP protocol. This means that when attackers want to carry out an attack, they must ensure that the payload itself complies with those protocols. This is different compared to attacks in the computer vision domain, where it is possible to make minimal changes to a pixel without affecting the function of the image itself. As a result, attackers need to develop ways of ensuring that their attack complies with the protocols as well as keeping the payload intact. The variety of data types and data sources in the cyber security domain also means that it is harder for algorithms to perform their function, which works to the attackers' advantage. This is because this diversity broadens the attack surface. It also complicates the development of attacks because an attack targeting one model cannot simply be used to attack other models, due to the different data types used by the models. Because of these differences, attack methods developed for other domains cannot automatically be applied to the cyber security domain.

A specific example of a type of ML algorithm in the cyber security domain that is vulnerable to adversarial attacks is a detector that recognises malicious internet domains used by attackers to control malware attacks. Using Adversarial AI techniques, attackers are able to rapidly generate domain names that can bypass these detectors (Sivaguru, Choudhary, Yu & Tymchenko, 2018). These attack algorithms are called Domain Generation Algorithms (DGAs). TNO's recent research on improving the robustness of algorithms against these Adversarial AI DGAs found that the detection algorithm can be improved by adding malicious samples to the training data. The idea behind this is that a model will be better able to recognise malicious domains if it is trained on malicious samples.

It is important to consider which data will be used for this purpose, since an algorithm trained to recognise one type of malicious domain is generally less successful in recognizing other types of malicious domains (Anley, 2022). A key element to consider when selecting suitable training data is therefore that the dataset must be representative of the context in which the algorithm will ultimately be deployed. A comparison can be made to a Formula 1 car, which can fly around a track but is useless in everyday life. One problem is that a method has not yet been developed for selecting malicious samples. Future research therefore needs to focus on this area.

6 What conclusions can we draw?

Adversarial AI is an exceptionally active and rapidly developing threat. Due to the increasing use of AI, more and more products are therefore potentially becoming vulnerable to such attacks. However, there is a general lack of awareness of these disturbing developments, and defensive measures against Adversarial AI lag behind attack capabilities. This imbalance between attack capabilities on the one hand and defensive measures on the other needs to be resolved so that new AI systems can be securely developed and deployed, particularly within vital sectors. It is therefore important to improve the robustness of AI systems now, to ensure that the potential of AI can be safely realised.

In the research domain, substantial progress is being made in the development of, and defence against, Adversarial AI attacks. While this is positive, it also makes it more difficult to determine the level of knowledge as it is hard to select the most up-to-date and relevant literature. Such difficulties are reflected in the literature, where authors often propose overlapping classifications while certain aspects tend to be neglected. TNO's research into improving model robustness has also shown that improvements are indeed possible. One such improvement is training the model on malicious samples, which does pose the challenge of carefully selecting these malicious samples.

One area in which the Adversarial AI field of research could further develop is in the creation of a clear overview of defensive measures. As we have shown, it is possible to create a broadly supported overview of attack techniques. The same does not apply to defensive measures, however, where research inherently still lags behind recent developments in the field of attacks. It is therefore important to place a greater focus on this essential component in order to make ML models more resilient to Adversarial AI.

The results of TNO's research show that Adversarial AI can be applied to the cyber security domain and that experiments with Adversarial AI techniques within this domain can yield relevant insights for applications within other domains.

Bibliography

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.

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.

Endnotes

1 Source: www.scopus.com with the search term “adversarial machine learning”

2 Source:
1) Scopus with the search term: “(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 with the search term: ‘adversarial machine learning’ AND ‘cyber’, from 2014.

Authors

Niels Brink

✉ niels.brink@tno.nl

Yuri Maas

✉ yuri.maas@tno.nl

Jip van Stijn

✉ jip.vanstijn@tno.nl

Puck de Haan

✉ puck.dehaan@tno.nl

Yori Kamphuis

✉ yori.kamphuis@tno.nl

Gwen Jansen-Ferdinandus

✉ gwen.ferdinandus@tno.nl

Bram Poppink

✉ bram.poppink@tno.nl

Irina Chiscop

No longer working at TNO

innovation
for life

Context

This publication stems from counter-AI research within the TNO Cyber & Electronic Warfare cluster and ties in with REAIM 2023, the first global summit on responsible artificial intelligence in the military domain, organised by the Dutch government.

Data and images from this publication may be used provided TNO is quoted as the source.

tno.nl/en/safe