# Artificial Intelligence

Cybersecurity – challenges for political Switzerland

## State of the art

The terms machine learning and Artificial Intelligence (AI) cover a broad field originating from statistics and operations research. At its core, AI is a method for predicting outcomes from sets of data. It does this by creating models that automatically infer patterns in the data and use those patterns to make decisions. AI is quickly becoming a critical technology for the digital society and industry. We increasingly depend on AI's ability to learn from past experiences, to reason, to discover meaning, or classify complex data to make critical decisions, and to automate processes and decision making.

AI pervasiveness gives rise to «**Adversarial Artificial intelligence (AAI)**» where attackers (A) exploit AI to craft attacks to compromise AI models in use, and (B) use AI to scale and automate elements of attacks that previously were simply impossible (DeepFakes) or relied heavily on manual processes.

AAI causes machine learning models to misinterpret inputs and behave in a way that's favourable to the attacker. To compromise a models' behaviour, attackers create «adversarial examples / data» that often resemble normal inputs, but instead break the model's performance. AI models then misclassify adversarial examples to output incorrect answers with high confidence.

## Recommendations

1. Organizations need to include their AI models and AI driven automation and decision making in their risk assessment.
2. Build an understanding of where adapting AAI driven threats will challenge current postures and where to focus limited resources.
3. Include testing and assessment of robustness of AI models in decision making.
4. Government, business and education systems need to ensure that they have and develop the requisite skills base and talent pipeline.

## Challenges

Cheap computational power and abundance of collected data have allowed modellers and attackers to develop increasingly complex AI models at low cost. As the accuracy and complexity of AI models continued to grow, many behaviours they capture defy any comprehensive human understanding. Most of these AI models have become black boxes. If an attacker can determine a particular behaviour in an AI model that is unknown to its developers, they can exploit that behaviour for potential gain.

Several AI models, including state-of-the-art neural networks, are vulnerable to adversarial examples. That is, these models misclassify examples that are only slightly different (imperceptibly different for humans) from correctly classified examples.

The vulnerability to AAI becomes one of the major risks for applying AI in safety or security critical environments. Attacks against core technologies like computer vision, optical character recognition (OCS), natural language processing (NLP), voice and video (DeepFakes), and malware detection have already been demonstrated.

AAI threat examples include:

1. Face-swapping in videos using machine-learning algorithms. Deep-fake services are already offered online for a few dollars.
2. Image recognition/classification in the autonomous driving domain, e.g. mis-interpretation of street signs or obstacles.
3. Manipulation of text recognition in automated document or payment processing service lines.
4. Ability to learn and industrialize bypassing of AI driven fraud detection and control mechanisms.
5. Malicious biasing of AI models to favour or miscredit certain groups.

## Need for action

AAI targets areas of the attack surface we never previously secured, the AI models themselves. Organizations need to include their AI models and AI driven automation and decision making in their risk assessment. Defending against AAI encompasses *proactive* and *reactive* strategies. Proactive strategies make AI models more robust against adversarial examples while reactive strategies aim to detect adversarial examples when the AI model is in use.

We need to appreciate and develop an understanding of this new and evolving threat environment and increasingly challenge processes driven by automated AI decision making.

## References

- Explaining and Harnessing Adversarial Examples: https://arxiv.org/pdf/1412.6572.pdf
- Ai Is The New Attack Surface: https://www.accenture.com/_acnmedia/Accenture/Redesign-Assets/DotCom/Documents/Global/1/Accenture-Trustworthy-AI-POV-Updated.pdf
- What is adversarial artificial intelligence and why does it matter?: https://www.weforum.org/agenda/2018/11/what-is-adversarial-artificial-intelligence-is-and-why-does-it-matter/
- Deepfakes web β: https://deepfakesweb.com

## Contact

Nicole Wettstein
Head of priority programme Cybersecurity
+41 44 226 50 13

https://www.satw.ch/cybersecurity-challenges

**Impressum**
Swiss Academy of Engineering Sciences SATW
**Expert contributions**
Karl Aberer, EPFL | Umberto Annino, InfoGuard | Alain Beuchat, Banque Lombard Odier & Cie SA | Matthias Bossardt, KPMG | Adolf Doerig, Doerig & Partner | Stefan Frei, ETH Zürich | Roger Halbheer, Microsoft | Pascal Lamia, MELANI | Martin Leuthold, Switch | Hannes Lubich, Board of Directors and Advisor | Adrian Perrig, ETH Zürich | Raphael Reischuk, Zühlke Engineering AG | Riccardo Sibilia, VBS | Bernhard Tellenbach, ZHAW | Daniel Walther, Swatch Group Services | Andreas Wespi, IBM Research Lab
**Editing and graphics**
Beatrice Huber, Claude Naville, Adrian Sulzer, Nicole Wettstein

The views expressed here are those of the members of the SATW Cyber Security Advisory Board and do not necessarily reflect the official position of SATW and its members.

www.satw.ch
September 2020