CYBERSECURITY NATIONAL LABORATORY

sAIfer Lab
Joint lab on Safety and Security of AI

Pitfalls in using AI for Cybersecurity

GIORGIO GIACINTO
giorgio.giacinto@unica.it

Sicurezza Cibernetica e Aerospazio
Roma, 7 aprile 2025

UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

Università di Genova

Progress in Artificial Intelligence

- Research in AI has made extraordinary progress in areas such as
  - Image/Video Understanding
  - Machine and Robot Vision (Computer Vision)
  - Text and Speech comprehension and translation (Natural Language Processing, Speech Recognition)
  - Autonomous and Decision Support Systems

- Key drivers
  - increasingly sophisticated models of data-driven learning (*machine, deep learning, large models*),
  - from large masses of data (*big data*),
  - with continuous advances in related technologies such as IoT (*Internet of Things*) with distributed sensory devices
  - with the large availability of computational resources, ranging from High Performance Computing (*HPC*), cloud-based services, mobile and embedded platforms.

Machine Learning and (Cyber)Security

Attack

Identification, tracking and **modelling** static and dynamic characteristics of the target: source- and byte-code, execution behavior, network traffic, etc.

Discovering vulnerabilities in the target networks and systems, both in the physical and virtual domains to create **automatic**, **targeted** attacks.

Defence

Identify, track and **model** the **behaviour** of attackers to design and implement effective **adaptive strategies** for protection and defence.

Early detection of vulnerabilities and weaknesses in any of the employed systems and technologies.

Large Language Models and Cybersecurity

- Large Language Models (LLMs) are extensively used in **software development**
  - Software development assistants
  - Code completion
  - Generation of software documentation
- Threat actors can use LLMs to develop **malicious code**
- LLMs help **find vulnerabilities** in source code
- LLMs can **assist in collecting security information** from multiple document sources

Limitation: the knowledge base behind LLMs includes trusted and untrusted information

LLM for attack deployment

https://threatresearch.ext.hp.com/hp-wolf-security-threat-insights-report-september-2024/

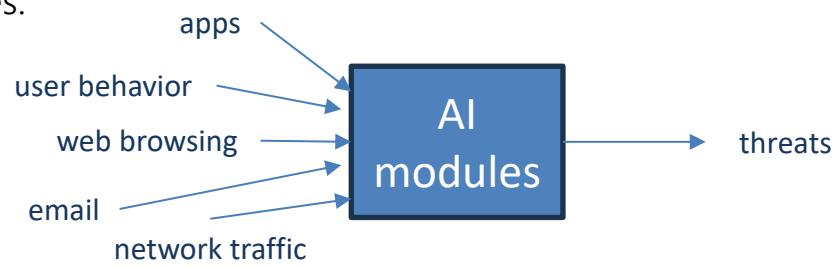
- HP Wolf Security detected a suspicious email attachment posing as an invoice.
- The attachment contained malicious 'script' code
- *Based on the scripts' structure, consistent comments for each function and the choice of function names and variables, it's highly likely that the attacker used GenAI*

```
// Arrête un processus PowerShell en cours d'exécution
function arreterProcessusAvecPowerShell() {
    // Exécution de PowerShell
    shellWsh.Run(cheminPowerShell, 2);

    // Obtenir la collection des processus en cours via WMI
    var serviceWMI = obtenirServiceWMI();
    var requeteProcessus = "SELECT * FROM Win32_Process";
    var collectionProcessus = serviceWMI.ExecQuery(requeteProcessus);
    var enumerateur = new Enumerator(collectionProcessus);

    // Parcours des processus en cours
    for (; !enumerateur.atEnd(); enumerateur.moveNext()) {
        var processus = enumerateur.item();
```

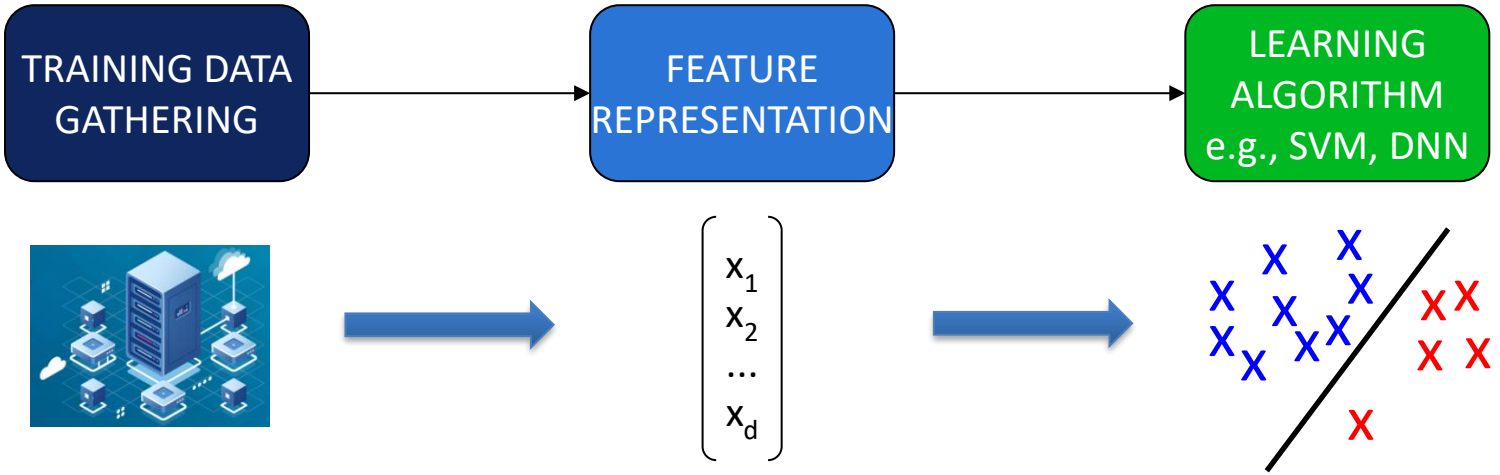
Some uses of ML and AI in cybersecurity

- **Sophistication** and **obfuscation of attacks**
  - Made up of multiple components
  - Executed in multiple stages
  - Similarity with legitimate activities
  - Target hidden software vulnerabilities

- The detection process requires the **extraction** and **correlation** of a **large number of features** to spot any weak signals of malicious activities.
  - Different categories of malware
  - Spam
  - Phishing
  - Malicious network traffic
  - Attacks against web applications



Machine Learning for Cybersecurity



Reliable, representative, trustworthy data sources

Suitable representation of source code, binary code, network traffic, etc.

Learning algorithms optimise an **objective function in the feature space**

I. Corona, G. Giacinto, F. Roli, Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues, Information Sciences, 2013.

Pitfalls in AI for Cybersecurity

- Choice of the data set to train the model
  - Is it representative of a real-world scenario?
  - Can we trust the labels (malicious, legitimate) assigned to the samples?

- Design of the learning process
  - Which methodology has been used to select the parameters?
  - Is there a process to check if the learned correlations agree with the experts' experience?

- Evaluation of performances
  - Choice of metrics
  - Choice of the test sets

Adversaries can leverage the above weaknesses in the design and deployment processes

Adversarial Machine Learning

	Attacker's Goal		
	Misclassifications that do not compromise normal system operation	Misclassifications that compromise normal system operation	Querying strategies that reveal confidential information on the learning model or its users
Attacker's Capability	**Integrity**	**Availability**	**Privacy / Confidentiality**
Test Data	**Evasion** (a.k.a. adversarial examples)	**Sponge attacks** to increase power consumption or response latency	**Model extraction** / **stealing** and model **inversion** (a.k.a. hill-climbing attacks)
Training Data	**Poisoning** to allow subsequent intrusions) - e.g., backdoors or trojans	**Poisoning** (to maximise classification error)	**Poisoning** aimed at extracting training data at test time

Adapted from: Biggio B., Roli F., "Wild patterns: Ten years after the rise of adversarial machine learning", 2018

Safe adoption of AI in Cybersecurity

- Artificial Intelligence **opportunities**
  - Empower **autonomous** system
  - Produce human-actionable **information from** large quantities of **raw data**
  - Increasing role in **security** and **cybersecurity** tasks

- **Challenges** for a strong implementation of AI in Cybersecurity
  - Align the methodologies and technologies to **cybersecurity goals**
  - Education and **Training** on the correct design and use of AI and ML tools
  - Thorough **tests** in real contexts and scenarios, facing adversaries

Adaptation and Evolution should characterise the next generation of Artificial Intelligence approaches to security to be prepared for new emerging cybersecurity threats