AI-ASsisted cybersecurity platform empowering SMEs to defend against adversarial AI attacks

Grant agreement ID: 101131292

FAQ

[Question] **Which are the main types of attacks that an adversary can perform on ML/DL models?**

[Answer] There are two major types of attacks that an adversary can perform on ML/DL models:

- **The poisoning attack** is the attack type that targets the system's learning process during the training phase.
- **An evasion attack** targets the testing phase of already trained ML/DL classifiers by manipulating the samples without having access to the training data aiming at violating the system's integrity

[Question] **What is a digital twin and how can it enhance deception mechanisms?**

[Answer] A Digital Twin (DT) is a virtual replica of a real system that updates continuously with real-time data, allowing for monitoring and simulation of what-if scenarios. Unlike static models, DTs create a dynamic link between the physical and digital world.

In cybersecurity, DTs improve deception mechanisms by creating highly realistic environments that mimic actual systems. Unlike traditional honeypots, DTs can adapt dynamically to evolving threats by generating valuable intelligence on attacker behaviour.

[Question] **Which are the detection mechanisms in ML/DL models?**

[Answer]

Detection mechanisms in ML/DL models can be categorized into supervised/semi-supervised learning, statistical methods, and distance-based techniques. Supervised learning approaches use auxiliary models trained to differentiate between adversarial and benign inputs, relying on labeled datasets of adversarial examples. Statistical methods identify adversarial attacks by analyzing deviations in feature distributions

HORIZON EUROPE

This project has received funding from the European Union under HORIZON-TMA-MSCA-SE, Topic HORIZON-MSCA-2022-SE-01-01, Grant Agreement No 101131292.

caused by perturbations. Distance-based techniques measure the similarity between inputs and known valid samples, rejecting anomalous inputs that significantly deviate from expected patterns.

[Question] **What are honeypots, and which is the difference between low-, mid- and high-interaction honeypots?**

[Answer]

A honeypot is a computer system or resource designed to attract and deceive potential attackers to analyze their tactics, gather threat intelligence, or divert attacks from real systems. Honeypots can simulate vulnerabilities so that attackers interact with them while security administrators collect valuable data on intrusion attempts.

Low-Interaction Honeypots provide the least amount of interaction for an adversary connecting to the system. They offer no real interaction capacities (also known as depth) to the attacker.

- Simulate basic services or systems but do not allow real interaction with the attacker.
- Easy to deploy and require minimal resources.
- Mainly used to detect automated attacks and port scans.

Mid-Interaction Honeypots still do not provide advanced interaction capacities to the attacker (there is not a deep exchange system behind) but can simulate a system shell to run commands on. Hence, these honeypots try to present a more attractive target and catch a greater scope of attacks, at the same time that collect malware sample uploaded by the adversary.

- Offer more functionalities than low-interaction honeypots but still limit what the attacker can do.
- Simulate some services and provide more realistic responses to gather information on attack techniques.
- Useful for studying malware and more advanced attack methods.

High-Interaction Honeypots have the highest odds of trapping a human adversary. On this interaction level, the honeypot has a real depth in interaction with the attacker, providing a full fledge environment that apparently might look valid, appearing as vulnerable to the outside world. The high-interaction possibilities allow for insight into attacker movement and activities on the system.

- Fully functional systems designed to allow attackers to interact with them without obvious restrictions.
- Provide detailed insights into advanced attacks and cybercriminal techniques.
- More challenging to set up and maintain, requiring extra security measures to prevent attackers from using them as a pivot point.

[Question] **What are web crawlers? What does it mean to explore AI-based web crawlers for the deep web with cyber-security purposes?**

[Answer]

This project has received funding from the European Union under HORIZON-TMA-MSCA-SE, Topic HORIZON-MSCA-2022-SE-01-01, Grant Agreement No 101131292.

Web crawlers are systematic web crawling tools to collect information from the web pages found; they are based on the sequential tracking of links from one page to another according to predefined bases (e.g. page titles, websites URL, metatags and web page contents). Normally, web crawlers are used as search engines, but it is also very common to use them in conjunction with a scraper, which allows to extract specific information from web pages, for example to cybersecurity purposes.

The deep web refers to parts of the internet not indexed by search engines, including private databases, password-protected pages, and hidden services. AI-based web crawlers are advanced tools that use artificial intelligence and machine learning techniques to automatically explore and extract information from these hidden parts of the web. In that way, cybersecurity engineers use this information to: Identify Threats, Automate Threat Intelligence, Behavioral Analysis, Data Breach Detection, etc.

[Question] **What kind of white-box adversarial attacks can target models working with tabular data (e.g., Modbus/TCP network flow statistics)?**

[Answer] White-box attacks against models working with tabular data, such as Modbus/TCP network flow statistics, exploit the known architecture and parameters of the target model. Commonly used evasion techniques are gradient-based (FGSM, PGD) or optimization-based (C&W, DeepFool). These attacks aim to create imperceptible adversarial examples by perturbing input features while key constraints are kept in valid ranges. Furthermore, other types of attacks, like model extraction, membership inference, or poisoning attacks, can target models working with tabular data. Model extraction attacks aim to copy the target model's functionality, while membership inference attacks aim to identify whether a given input was part of the target model's training dataset. Finally, poisoning attacks, similarly to evasion attacks, want to degrade the target model's performance, but they do so during its training process.

[Question] What kind of defenses can be used to mitigate the effects of white-box adversarial attacks against models working with tabular data (e.g., Modbus/TCP network flow statistics)?

[Answer] Commonly used defenses against adversarial attacks involve adversarial training, where adversarial examples are used combined with clean data to train the machine learning models. Differential privacy masks the identity of individuals in a dataset by increasing its randomness. Another method is adding noise and clustering feature values to the model's inputs to stabilize them against perturbations. Furthermore, many models limit the computational resources, such as the number of queries per user, provided to the attackers. Anomaly detection is also exploited to identify potentially altered inputs or patterns that may harm the model's performance. Some models output hard labels without probabilities in order to make them more difficult to use for attackers. Finally, in some cases, the model's outputs are also perturbed to hide valuable information from the attackers.

[Question] **Which are the mitigation techniques in ML/DL models?**

[Answer] Mitigation techniques in ML/DL models focus on enhancing their robustness against adversarial attacks. These techniques include adversarial training, defensive distillation, ensemble learning, and pre-processing methods. Adversarial training improves model resilience by incorporating adversarial examples into the training process. Defensive distillation strengthens model predictions by training a second model with softened probability outputs. Ensemble learning increases robustness by combining multiple models to reduce vulnerability to specific attacks. Pre-processing techniques, such as feature reduction and input

HORIZON EUROPE

This project has received funding from the European Union under HORIZON-TMA-MSCA-SE, Topic HORIZON-MSCA-2022-SE-01-01, Grant Agreement No 101131292.

transformation, modify data before feeding it into the model, making adversarial perturbations less effective.

[Question] **What kind of blackbox adversarial attacks can target models working with tabular data (e.g., Modbus/TCP network flow statistics)?**

[Answer] Black-box attacks do not have any knowledge of the target model's architecture or inner parameters. So, it is quite common to use the target model as an oracle to train attack models that copy its behavior. These are the transferability attacks, which rely on crafting adversarial examples using the known attributes of the attack models that are later used against the target model. Another type of attack is the query-based attack, where the attackers query the target model with a high number of queries to infer decision boundaries and craft adversarial examples. Furthermore, the gradient-based attacks adapt to the black-box setting by using gradient estimations through queries that approximate the target model's gradients. Finally, a different type of attack starts with a highly perturbed input and iteratively decreases the perturbation to create imperceptible adversarial examples that remain misclassified.

[Question] **What kind of defenses can be used to mitigate the effects of blackbox adversarial attacks against models working with tabular data (e.g., Modbus/TCP network flow statistics)?**

[Answer] The most used defense against adversarial attacks is adversarial training, where, aside from clean data, adversarial examples are also used to train a model to increase its robustness. Also, many models use feature engineering, such as feature selection or noise reduction, to minimize exploitable patterns. Furthermore, synthetic datasets generated by GANs are also useful to avoid using public or easily accessed datasets. Reducing the available computational resources of the attackers or increasing their computational cost is another way to defend against adversarial attacks. Anomaly detection to identify adversarial examples is also exploited, while returning hard labels without probabilities is always helpful. Finally, wavelet transform analysis to identify perturbations in high-frequency regions of tabular data can highly increase a system's robustness.

[Question] **Which is the major type of adversarial attack against an image-based neural network?**

[Answer] The major type of adversarial attack against image-based neural networks is the **one-pixel attack**, where an attacker can modify one pixel of an image such that the neural network fails to predict the content of the image, or it classifies the image in the wrong family (e.g. a picture of a dog is classified as a cat).

[Question] **What is a virtual persona and how can it be utilized in cybersecurity?**

[Answer] A **virtual persona** is a digitally constructed identity that replicates real-world behaviors, interactions and attributes of a human user or system component. It serves as an advanced cybersecurity mechanism, strategically designed to mislead adversaries, gather intelligence, and reinforce security measures. Virtual personas can be dynamically generated and adapted to diverse scenarios, allowing organizations to strengthen their overall security framework.
These entities play a strategic role in cybersecurity, enhancing threat defense, facilitating intelligence acquisition, and strengthening AI-driven security mechanisms. When deployed within networks, they mislead attackers by diverting malicious activities into controlled environments, enabling real-time monitoring and forensic analysis. By actively engaging with cybercriminals, these personas provide

This project has received funding from the European Union under HORIZON-TMA-MSCA-SE, Topic HORIZON-MSCA-2022-SE-01-01, Grant Agreement No 101131292.

valuable intelligence on attack tactics, techniques, and procedures, enhancing proactive threat detection and incident response. Furthermore, they enhance intelligent cybersecurity frameworks by simulating adversarial behavior, refining anomaly detection models, and distinguishing between legitimate users and malicious entities.

This project has received funding from the European Union under HORIZON-TMA-MSCA-SE, Topic HORIZON-MSCA-2022-SE-01-01, Grant Agreement No 101131292.