A Threat Monitoring Methodology for AI Based Applications

Romanos Kapsalis
Code4Thought PC
Patras, Greece
romanos@code4thought.eu

Christos Aridas
Code4Thought PC
Patras, Greece
chris@code4thought.eu

Yiannis Kanellopoulos
Code4Thought PC
Patras, Greece
yiannis@code4thought.eu

ABSTRACT

The rapid growth of Artificial Intelligence (AI) in recent years has raised some serious concerns regarding Security, Privacy and Transparency issues. Additionally, the increasing adoption of AI by organizations, industries, governments and individuals have made the standardization and certification of AI a necessity. In this paper, we propose a Threat-Monitoring methodology for image classification tasks, that includes 2 steps, in order to ensure the Verification and Validation of AI systems. In the first step an AI system is evaluated against integrity violation attacks, and more specifically, against evasion attacks which attract much of the recent attention of the literature [5] [13] [6] In the second step, two types of attacks are performed, exploratory and extraction to detect potential privacy threats. In both steps, State Of The Art techniques are being used. Finally, the expected benefits of applying the proposed methodology are being discussed.

1 INTRODUCTION

Threat-Monitoring is a key concept in cybersecurity as it helps minimizing security risks and enhances the robustness of a given system/network. In [1] is defined as "*Analysis, assessment, and review of audit trails and other information collected for the purpose of searching out system events that may constitute violations of system security.*"

Threat-Monitoring consists of real-time monitoring of threats and breaches and continually analyzing and auditing systems to detect potential security threats and issues, so as to safeguard the system.

In the field of Artificial Intelligence, Threat-Monitoring is one of the most common approaches to ensure the Verification and Validation of AI systems. There, Threat-Monitoring involves mainly the testing against adversarial attacks.

Adversarial attacks is a method to generate adversarial examples. According to ISO-29119 "*an adversarial example is where an extremely small change made to the input to a neural network produces an unexpected (and wrong) large change in the output (i.e. a completely different result than for the unchanged inputs)*".

There are many different ways to categorize adversarial attacks. One type of categorization is based on the adversary's knowledge.

More specifically, adversarial attacks can be divided to Black-box, White-box and Gray-box.

- In Black-box attacks, the attacker has limited knowledge of the architecture of the model and the adversarial examples are constructed based on queries. In this type of attack a common strategy, (which is applied in our proposed methodology), is to construct a surrogate model using the training dataset (or a part of it) as input. This model is trained to craft adversarial examples. This strategy, called transfer-based, relies on the transferability of adversarial examples between models. Another category of black-box attacks is decision-based attacks. These attacks rely solely on the last prediction of the model.
- On the other hand, in White-box attacks, the attacker has full knowledge of the AI system. Particularly, he has access to the model per se and usually has a thorough view of the model parameters and inner workings,
- In Gray-box attacks the attacker knows only of some parts of the AI system. He/she might know everything about the model (the architecture, the parameters, the hyperparameters), but training data will not be exposed to the attacker. Or on the contrary, the distribution of the training data might be visible to the attacker, but the model will be a black-box.

2 BACKGROUND WORK

Besides the Verification and Validation of AI based systems, Threat-Monitoring is an essential step towards the certification and standardization of them.

ISO standards and upcoming legislations in EU, clearly denote that AI-based systems, especially high risk systems as those used in aviation, cybersecurity, maritime/land domains, have to be resilient against adversarial attacks. Below we will present a series of extracts from articles proposed by the EU Artificial Intelligence Act [2] as well as the ISO/IEC TR 29119-11 and ISO/IEC DIS 5338 standards, regarding the robustness of AI systems.

In the Article 15 of the upcoming EU AI Act we read that:

"*High-risk AI systems shall be resilient as regards attempts by unauthorized third parties to alter their use or performance by exploiting the system vulnerabilities.*

The technical solutions aimed at ensuring the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent and control for attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws."

In the same spirit, ISO/IEC DIS 5338 in section 6.4.3.3 denotes that: "*...Security requirements: In case there is an additional attack surface resulting from the use of AI. Typically, this includes:*
— securing data that are used for either training or testing, or both, including protection against "poisoning attacks" when malicious actors inject data to influence the behaviour of machine learning models;
— protecting against input manipulation (e.g. a spam e-mail being classified as not spam);

[1] https://csrc.nist.gov/glossary/term/threat$_{monitoring}$

— protecting against "model inversion" when a malicious actor manages to deconstruct sensitive data that are used for training a model;
— protecting against "model theft" when a malicious actor aims to copy the behaviour of a model that is intellectual property."

Whereas ISO/IEC TR 29119-11 in section 7.8, emphasizes more on the benefits gained from adversarial testing:

"*Adversarial testing is often referred to as performing adversarial attacks. By performing these attacks and identifying vulnerabilities during testing, measures can be taken to protect against future failures and so the robustness of the neural network is improved.*"

The aforementioned legislation and ISO standards highlight the importance of establishing a methodology to verify the robustness of an AI based system, in all the steps of the life cycle. The proposed methodology investigates the resilience of AI systems against **integrity violations** of the model and of the most common **privacy threats**.

The rationale behind the selection of those types of attacks is a) are prescribed in all the above mentioned legislations and standards and b) a considerable amount of literature has been published on this, denoting its criticality. For instance, a very recent publication [14] of the National Institute of Standards and Technology (NIST) of the US Department of Commerce, has heightened the need to examine those types of attacks. The purpose of this paper is to provide a concrete and thorough analysis of AI systems, that will not only help organizations to be more compliant to existing and upcoming legislations, but it will also help them mitigate potential security risks and detect them in real time.

3 PROPOSED METHODOLOGY

The following methodology validates the robustness of a model against 3 of the most common attack scenarios: evasion, exploratory and extraction, for image classification tasks.

The first step is to check how easily a model can be fooled. For this reason, an Evasion attack scenario is investigated.

In **Evasion attack** [12] scenario the objective is to feed adversarial examples into the input of a model/classifier, so as to be misclassified. The attack is applied in a way that the difference between the original and the modified input is not recognizable by a human. Unlike poisoning attacks, the attack happens in the test phase and training data are not modified. The goal of the analysis is to identify the smallest perturbation level [18] that the model is not susceptible to. Three distance metrics are being used to quantify the difference between the original pixel of an image and the pixel as modified by an adversarial example. Those metrics are L_0, L_2, L_∞ as defined in [6].

(1) L_0: number of pixels that have been modified in an image
(2) L_2: its value remains low when there are many minor changes to many pixels
(3) L_∞: each pixel of an image is allowed to be changed up to a threshold value, with no limit in the number of pixels that are changed.

The risk in this scenario is that the model gives false predictions, which can be quite costly, especially in AI systems used for defense.

The described methodology applies to two main types of evasion attacks based on target:

- **Untargeted attacks:** where the goal is to change the original prediction of the model to an arbitrary class
- **Targeted attacks:** change the original prediction of the model to a specific class
  - On least likely target: change predicted class to the class that the model has the lowest confidence.
  - On most likely target: change predicted class to the class that the model has the next best confidence, besides the predicted class.

As an evaluation metric the *attack-success rate per class* is used, which measures the percentage of successful attacks per class. In Figure 1 an example of an targeted attack on least likely targeted is presented, on the in CIFAR10 public dataset [11].

After investigating how susceptible a model is to an evasion attack, the next step is to check how resilient it is against privacy attacks and answer questions like:

- How easily can an attacker steal our model?
- How easily can an attacker reconstruct our training data? How accurate can this reconstruction be?

Exploratory attacks as denoted in [7] "*do not modify the training set but instead try to gain information about the state by probing the learner*". This scenario can be further divided into a number of categories. Among those, Membership Inference Attacks (MIA) [17] and Attribute Inference (or Model Inversion [9]) Attacks are the most common and severe categories that need to be examined.

The objective of Membership Inference attack is to infer if a data instance was used to train a model. In a Model Inversion (or Attribute Inference) attack, the attacker is able to reconstruct the training dataset, having access only to the output of the model. Consequently, Model Inversion could be a major threat to data privacy, as sensitive private data could be stolen/leaked.

Another type of adversarial attack that is considered more difficult to be implemented, is copying/stealing a model by executing random queries at the target model. This type of attack is called **Extraction attack (or Model Theft)**. Several techniques have been defined in the literature to implement this attack scenario. One technique is via crafting a copycat network. This attack is performed as follows, as described in [8]. First, the target model is probed with random unlabelled data. Then, the predicted labels by the model and the input data are concatenated to create a fake dataset, this fake dataset is used for training the copycat network. Another technique that is being examined is defined in [10] and tries to extract a functionally-equivalent model to the target model with similar predictions in all inputs. Finally another approach in extraction scenario is to use either reinforcement [15] or active learning to create more efficient queries to the target model.

The proposed methodology applies all these types of attacks to ensure the robustness of the model and is applicable to both Black-box and White-box attacks.

4 EXPECTED BENEFITS

In this section the main benefits of the proposed methodology will be presented. First and foremost, it helps organizations in being compliant to existing and upcoming legislations in EU (EU AI Act)

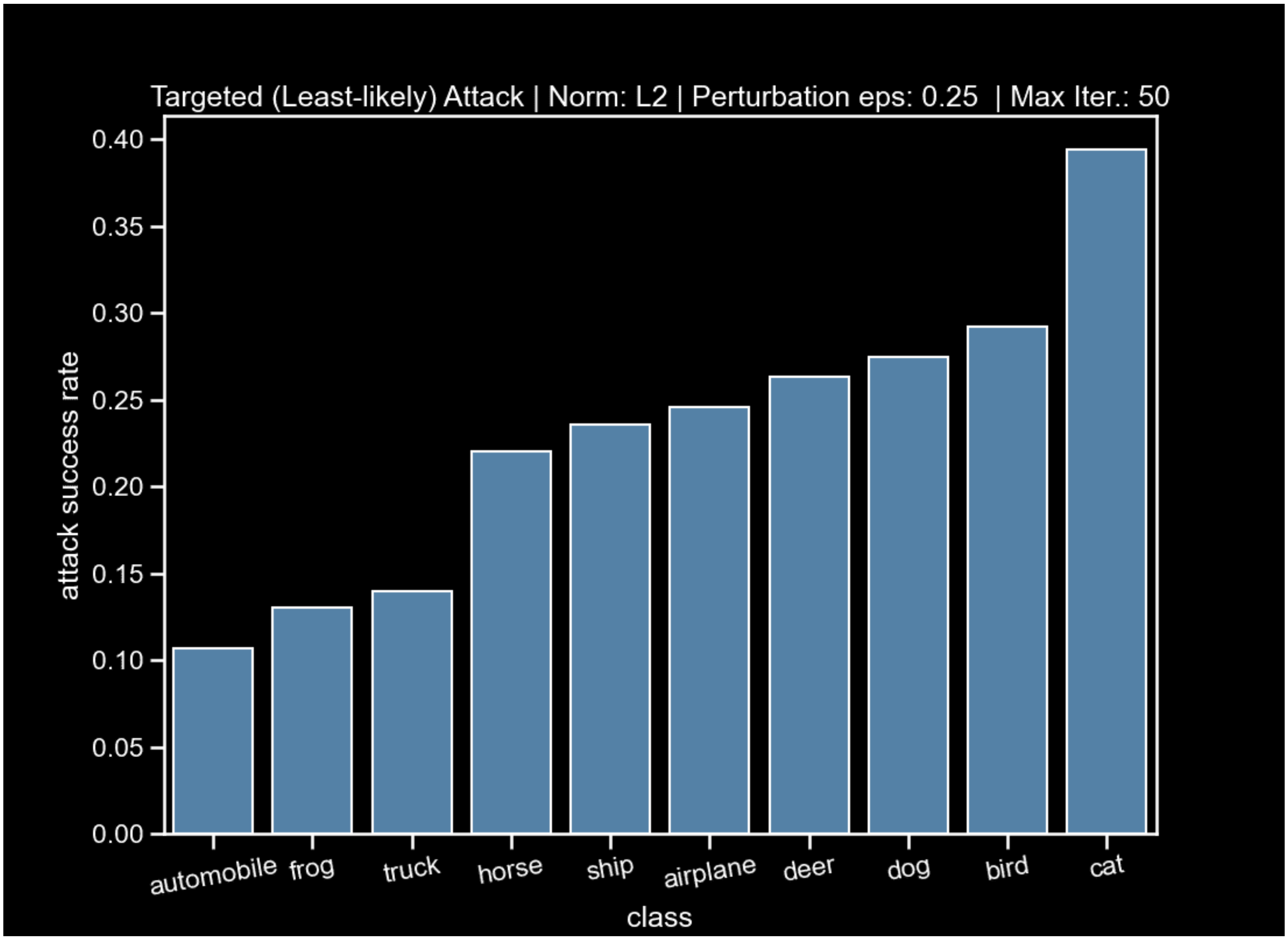

Figure 1: An example of targeted attack on least likely target

and US (US AI Bill of Rights), frameworks, regulations and the ISO 29119 and 5338 standards. The steps described above and the selected types of attacks are based on what these standards and legislations dictate. More specifically:

- Legislations:
  - **EU AI Act**:
    - *".. They should be resilient against risks connected to the limitations of the system (e.g. errors, faults, inconsistencies, unexpected situations) as well as against malicious actions that may compromise the security of the AI system and result in harmful or otherwise undesirable behaviour."*
  - **Blueprint for an AI Bill of Rights** [4] of the White House of US:
    - *"...Systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing monitoring that demonstrate they are safe and effective based on their intended use, mitigation of unsafe outcomes including those beyond the intended use, and adherence to domain-specific standards."*
- Framework:
  - **Artificial Intelligence Risk Management Framework (AI RMF 1.0)** of NIST [3]
    - *"...Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data, or other intellectual property through AI system endpoints."*
- Regulation:
  - **General Data Protection Regulation (GDPR)** [1] - Article 5.1 (f)
    - *"Personal data shall be: ... processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures ('integrity and confidentiality')."*
- ISO standards:
  - **ISO 5338** in section 6.4.3.3 includes explicitly Model Inversion and Model Theft attacks and implicitly evasion attacks at the security requirements:

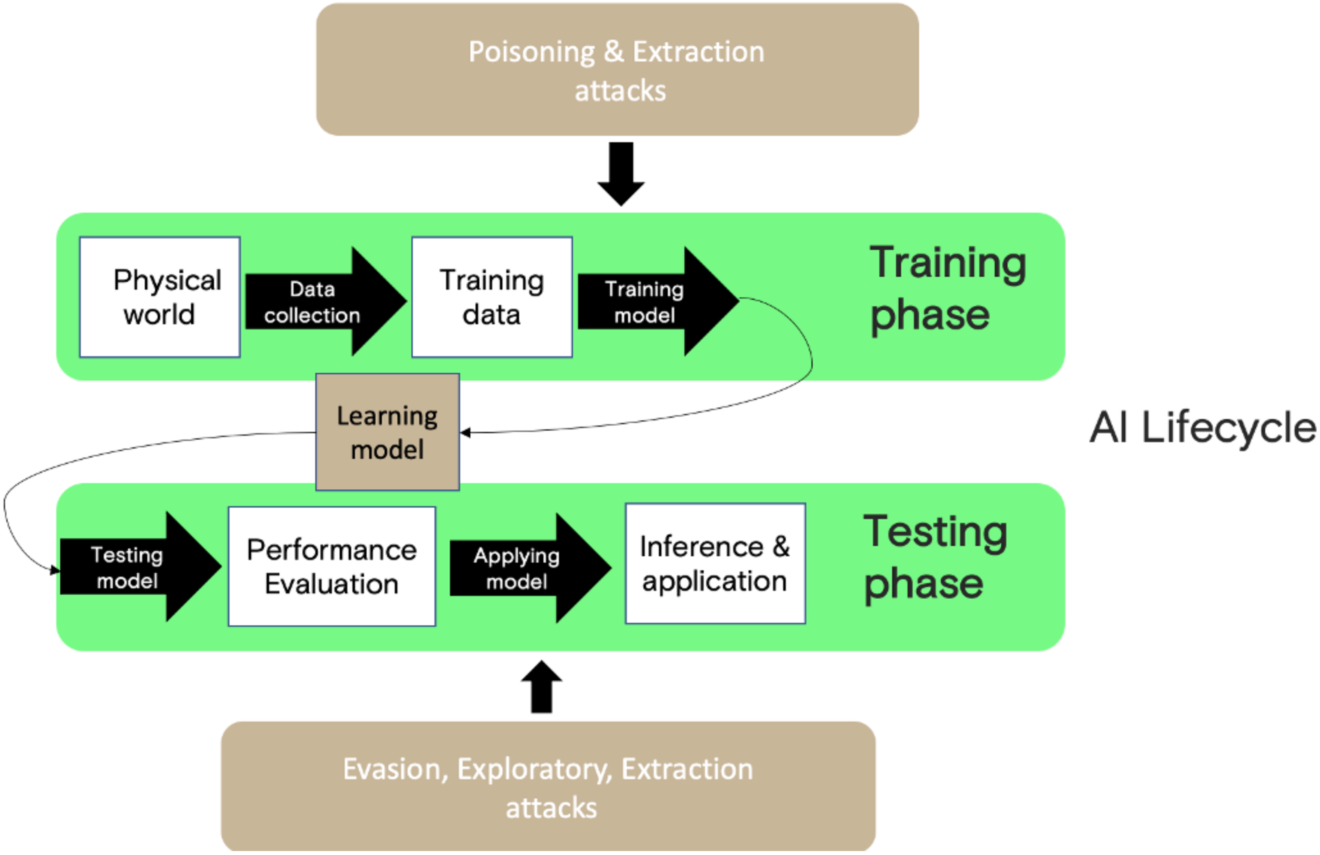

Figure 2: The attacks of the proposed methodology in AI lifecycle

  - ∗ "Security requirements: In case there is an additional attack surface resulting from the use of AI.Typically, this includes:— securing data that are used for either training or testing, or both, including protection against "poisoning attacks" when malicious actors inject data to influence the behavior of machine learning models;— protecting against input manipulation (e.g. a spam e-mail being classified as not spam);— protecting against "model inversion" when a malicious actor manages to deconstruct sensitive data that are used for training a model; — protecting against "model theft" when a malicious actor aims to copy the behavior of a model that is intellectual property."
- – In a different manner, **ISO-29119-11** in section 7.8 splits the types of attacks in 2 categories:
  - ∗ "Attacks can be made when training the model and then on the trained model (neural network) itself.", as depicted in Figure 2 and further describes each one of them: " Attacks during training can include corrupting the training data (e.g. modifying labels), adding bad data to the training set (e.g. unwanted features) and corrupting the learning algorithm."

Apart from the compliance, another important benefit is that the proposed methodology helps preserve the integrity and trustworthiness of AI outputs and decisions. An attack on AI systems in defense applications, either to their privacy or integrity, can be very costly. An integrity violation attack could leak highly confidential data, like potential targets of an army, potential vulnerabilities and details about their equipment (weapons, ammunitation, vehicles). Or even worse an evasion attack could lead to human casualties in a war condition.

Also, another aspect that this methodology might be found useful is the fact that the robustness of the AI systems must be tested and evaluated in a systematic and comprehensive manner. This enables detecting and preventing malicious activities that could compromise security or accuracy by capturing the full range of scenarios that the AI system might be applied.

Finally, by conducting exploratory attacks as the ones included in our methodology, we facilitate preserving data privacy for users, which is essential for:

- Increasing the trustworthiness of AI systems,
- Minimising data leakages/ data breaches
- Mitigating the risks associated with the difficulty of control and accountability.

5 NEXT STEPS

In conclusion, the proposed methodology is comprehensive, model-agnostic and covers both targeted and untargeted attacks. However, despite the fact that the methodology examines attacks related to privacy threats and integrity violation, it could be further extended with other types of attacks, such as poisoning attacks [16]. Furthermore, our team is developing a tool whose purpose is to implement and automate the steps of the proposed methodology and suggest potential mitigation and defense techniques.

REFERENCES

[1] 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). http://data.europa.eu/eli/reg/2016/679/2016-05-04/eng Legislative Body: OP_DATPRO.

[2] 2021. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206

[3] 2023. *AI Risk Management Framework: AI RMF (1.0)*. Technical Report NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg, MD. 48 pages. https://doi.org/10.6028/NIST.AI.100-1

[4] 2023-03-15. Blueprint for an AI Bill of Rights | OSTP. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[5] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion Attacks against Machine Learning at Test Time. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 387–402.

[6] Nicholas Carlini and David Wagner. 2016. Towards Evaluating the Robustness of Neural Networks. https://doi.org/10.48550/ARXIV.1608.04644

[7] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial Attacks and Defences: A Survey. https://doi.org/10.48550/ARXIV.1810.00069

[8] Jacson Rodrigues Correia-Silva, Rodrigo F. Berriel, Claudine Badue, Alberto F. de Souza, and Thiago Oliveira-Santos. 2018. Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. https://doi.org/10.1109/ijcnn.2018.8489592

[9] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1322–1333. https://doi.org/10.1145/2810103.2813677

[10] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. 2019. High Accuracy and High Fidelity Extraction of Neural Networks. https://doi.org/10.48550/ARXIV.1909.01838

[11] Alex Krizhevsky et al. 2009. Learning multiple layers of features from tiny images. (2009).

[12] Daniel Lowd and Christopher Meek. 2005. Adversarial Learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (Chicago, Illinois, USA) *(KDD '05)*. Association for Computing Machinery, New York, NY, USA, 641–647. https://doi.org/10.1145/1081870.1081950

[13] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Alina Oprea and Apostol Vassilev. 2023. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (Draft)*. Technical Report NIST AI 100-2e2023 ipd. National Institute of Standards and Technology. https://csrc.nist.gov/publications/detail/white-paper/2023/03/08/adversarial-machine-learning-taxonomy-and-terminology/draft

[15] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2018. Knockoff Nets: Stealing Functionality of Black-Box Models. https://doi.org/10.48550/ARXIV.1812.02766

[16] Miguel A. Ramirez, Song-Kyoo Kim, Hussam Al Hamadi, Ernesto Damiani, Young-Ji Byon, Tae-Yeon Kim, Chung-Suk Cho, and Chan Yeob Yeun. 2022. https://doi.org/10.48550/ARXIV.2202.10276

[17] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-Preserving Deep Learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1310–1321. https://doi.org/10.1145/2810103.2813687

[18] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. https://doi.org/10.48550/ARXIV.1312.6199