UDC 004.056.5 **DOI:10.30837/rt.2025.1.220.07**

Yu.L. GOLIKOV, Ye.V. OSTRIANSKA

RESEARCH AND CLASSIFICATION OF THE MAIN TYPES OF ATTACKS ON ARTIFICIAL INTELLIGENCE SYSTEMS IN CYBERSECURITY

Introduction

In today's world, artificial intelligence (AI) is increasingly being integrated into various areas of human activity, from finance and medicine to autonomous vehicles and cybersecurity. Along with the development of AI technologies, new challenges arise, including threats related to attacks on artificial intelligence. Such attacks can have serious consequences, including compromising data security, manipulating algorithms, and creating vulnerabilities in critical systems.

Machine learning, as the basis of most modern AI systems, allows them to find patterns in data, adapt to changes, and improve over time. However, since ML models rely heavily on the quality of the input data and the hypotheses they are based on, they are vulnerable to certain types of attacks. Attackers can manipulate training data, spoof input parameters, or exploit algorithmic weaknesses to achieve their goals.

In essence, the machine learning methodology used in modern AI systems is susceptible to attacks through publicly available APIs that expose the model and against the platforms on which they are deployed. For attacks on security models, attackers can compromise the privacy and data protection of both the model and the data simply by using publicly available interfaces and providing input data that is within an acceptable range. In this sense, the challenges faced by AML are similar to those faced by cryptography. Modern cryptography relies on secure algorithms in the theoretical sense of information. Thus, people should only focus on their reliable and secure implementation, which is the primary task for the cryptographic research community. Unlike cryptography, there are no information-theoretic security proofs for widely used machine learning algorithms. As a result, many advances in the development of mitigation tools for various classes of attacks are empirical and limited.

But many companies and organizations have been actively working in recent years to regulate the use of AI in their systems. For example, among a wide range of activities, NIST contributes to the research, standards, assessments, and data needed to develop, use, and ensure reliable artificial intelligence (AI). In 2024, NIST published a report [1] on machine learning threats, and in 2025 it updated it [25]. This report developed a taxonomy of concepts and defined terminology in the field of adversarial machine learning (AML).

Attacks on AI/ML systems can be divided into several categories depending on the attacker's goals, implementation methods, and the level of impact on the model. Among the most common types of attacks are poisoning attacks, evasion attacks, privacy attacks, and denial-of-service attacks. Each of these types of attacks uses different mechanisms of influence, from spoofing training data to exploiting vulnerabilities in already trained models.

The relevance of the study is due to the growing number of cases of hacking and manipulation of AI systems, which can lead to financial losses, security threats, and loss of trust in technology. Therefore, the purpose of this article is to study the threats and risks associated with the use of AI. The article identifies the types of attacks that can be directed at AI/ML models, the stages of the attack life cycle, the goals and objectives of the attacker, as well as the attacker's capabilities and knowledge of the learning process. The study allows us to better understand current risks in the field of artificial intelligence and identify measures to improve the security of such systems.

1. Classification of AI-based attacks

An AI-based attack can be classified according to many different parameters. For example, several attack classification systems were presented in [7, 8]. NIST also presented different types of

classification in its report on AI machine learning attacks [1]. In this section, we will consider some types of classifications.

1.1. Main types of attacks

Attacks based on machine learning and AI are usually classified according to the following general parameters [1]:

1) the training method and the stage of the training process when the attack is established;
2) goals and objectives of the attacker;
3) capabilities of the attacker;
4) the attacker's knowledge of the training process.

Fig. 1 shows a general classification of attacks aimed at AI/ML models.

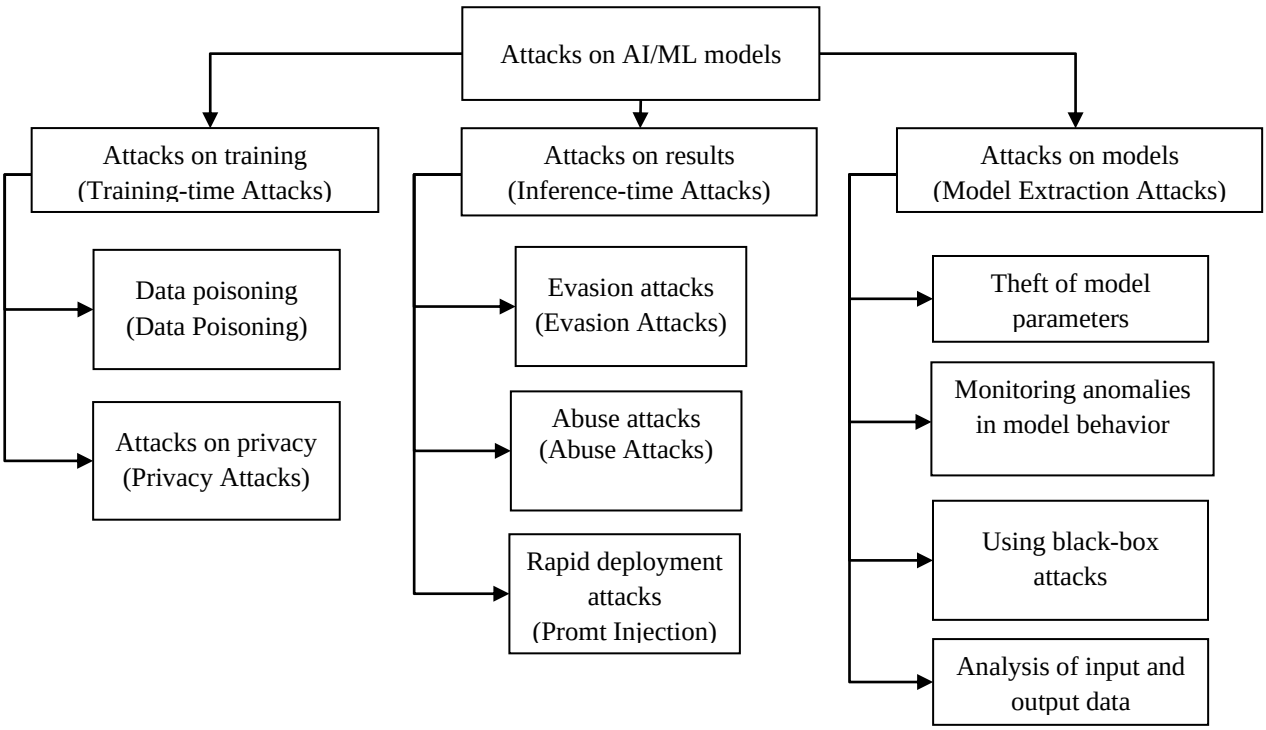

Fig. 1. Classification of attacks on AI systems

Fig. 1 shows a scheme of classification of attacks on AI systems. In turn, each of these attacks can be divided into the following main subgroups:

– Data Poisoning:
  - Introducing malicious data into the training set.
  - Label-flipping Attack.
  - Using hidden triggers (Backdoor Attack).
  - Manipulation of neural network weights.
– Privacy Attacks:
  - Extract data from the training set.
  - Model Inversion Attack.
  - Membership Inference Attack.
  - Hacking the model parameters.
– Evasion Attacks:
  - Manipulation of input data.
  - Image/text/audio attack.
  - Bypassing anti-virus systems.
  - Change the recognition parameters.

– Abuse Attacks:
  - Using generative AI to create fake content.
  - Deepfake (video, audio, images).
  - Social engineering attacks.
  - Creating malicious code.
– Prompt Injection attacks:
  - Influence on text models (ChatGPT, Bard, etc.).
  - Imposing unwanted answers.
  - Bypassing model limitations.
  - Generation of fake information.

1.2. Classification of attacks by attacker's goal

The attacker's goals are classified by three criteria in accordance with the three main types of security breaches [1] that are considered when analyzing system security: availability, integrity, and data confidentiality compromise. Accordingly, an attacker's success indicates the achievement of one or more of these goals.

An availability attack is an indiscriminate attack on machine learning (ML) in which an attacker attempts to disrupt the performance of a model during deployment. Availability attacks can be launched through data poisoning, where an attacker controls a portion of the training set, or through model poisoning, where an attacker controls the model parameters.

An integrity attack targets the integrity of the ML model's output, resulting in incorrect predictions made by the ML model. An attacker can cause an integrity breach by performing a bias attack during deployment or a poisoning attack during training. Evasion attacks require modifying test samples to create competitive examples that are misclassified by the model to a different class while remaining hidden and unnoticed by humans. Examples of such attacks can be found in [12, 13].

In privacy attacks, attackers may be interested in obtaining information about the training data or the ML model (resulting in data and model privacy attacks, respectively). An attacker may have different goals for compromising the privacy of training data, such as data modification (extracting the content or features of training data), data injection [14, 15] (the ability to extract training data from generative models), and injection of properties regarding the distribution of training data [16].

1.3. Classification of attacks by attacker's knowledge

Another criterion for classifying attacks is the extent to which the attacker has knowledge of the machine learning system. There are three main types of attacks according to this criterion [1]: white box, black box, and gray box.

- White box attacks. They assume that an attacker operates with full knowledge of the machine learning system, including training data, model architecture, and additional model parameters. Although these attacks operate under very strong assumptions, the main reason for analyzing them is to test the vulnerability of the system against the worst attackers and evaluate potential mitigations.
- Black box attacks. These attacks involve minimal knowledge of the ML system. An attacker can gain access to query the model, but they have no other information about how the model is trained. These attacks are the most practical because they assume that the attacker does not know the AI system and uses system interfaces that are easily accessible for normal use.
- Gray box attacks. There are a number of gray box attacks that capture conflicting knowledge between black box and white box attacks. Paper [17] presents a framework for classifying gray box attacks. An attacker may know the model architecture but not its parameters, or an attacker may know the model and its parameters but not the training data. Other common assumptions for gray box attacks are that the attacker has access to data distributed identically to the training data and knows the function representation. The latter assumption is important in applications where feature extraction is used before training an ML model, such as cybersecurity, finance, and healthcare.

That is, generally speaking, from an information perspective, if an attacker has complete knowledge of the model, such as parameters, functions, and training data, we speak of a white box attack. Conversely, if the attacker has no knowledge of the model's inner workings and only has access to its predictions, we call it a black box attack. Everything in between falls into the gray box category [22]. This is shown schematically in Fig. 2.

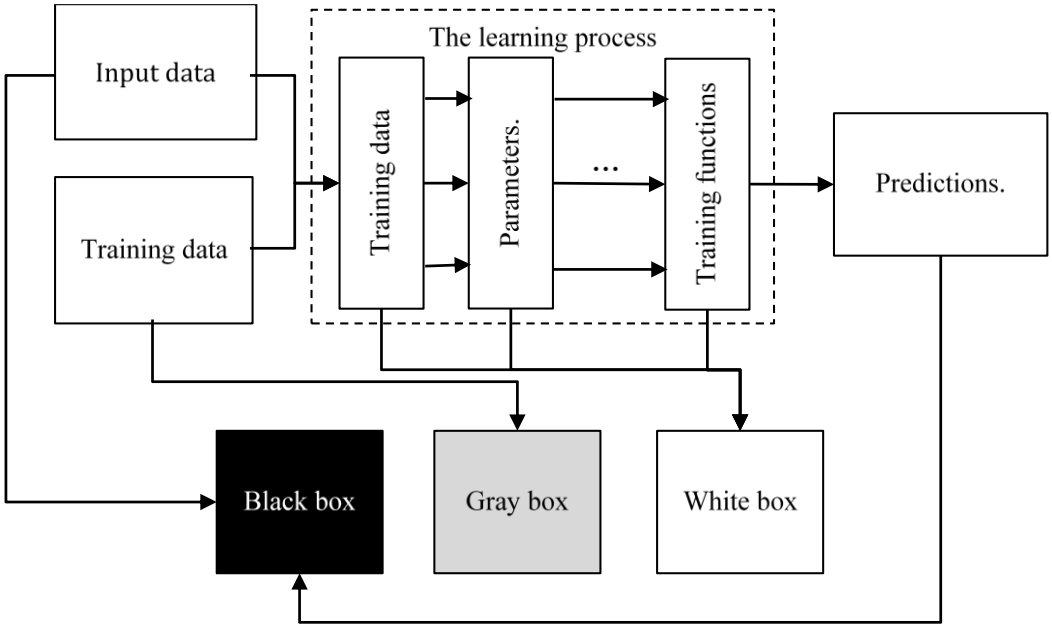

Fig. 2. Diagram of the degree of awareness of the attacker

In practice, the attacker often starts from a black box perspective and tries to increase his knowledge, for example by performing logical inference or oracle attacks, where the attacker queries the model to get clues about the model's internal elements or training data. Often, sensitive information about the target model can be obtained through more traditional means, such as open source intelligence (OSINT), social engineering, cyber espionage, etc.

2. Data poisoning attack

Attacks at the ML training stage are called poisoning attacks [1, 9]. In a data poisoning attack [5, 9], an attacker controls a subset of the training data by inserting or modifying training samples. In a model poisoning attack [10], the attacker controls the model and its parameters. Data poisoning attacks can be applied to all learning paradigms, while model poisoning attacks are most common in federated learning [11], where clients send local model updates to the server that processes the input, and in supply chain attacks, where malicious code can be added to the model by the model technology providers. Here, federated learning refers to a machine learning method focused on settings in which multiple entities (often called clients) jointly train a model, with the data used for training distributed in a decentralized manner. This distinguishes it from machine learning, in which data is stored centrally.

The first poisoning attacks detected in cybersecurity applications were availability attacks against the generation of worm profiles and spam classifiers that indiscriminately affect the entire machine learning model and essentially cause a denial-of-service attack for users of the AI system.

Poisoning attacks are considered to be one of the most dangerous among AI attacks and can cause either availability or integrity violations. In particular, availability attacks cause degradation of the machine learning model at all stages, while targeted and backdoor poisoning attacks are more stealthy and cause integrity violations on a small dataset. Poisoning attacks utilize a wide range of competitive capabilities such as data poisoning, model poisoning, label control, source code control, and test data control, resulting in several subcategories of poisoning attacks. According to the threat

model, they can be used in both white-box and black-box scenarios [18], which were discussed in Section 1.3 of this article.

Among the methods of preventing data poisoning attacks, there are two most effective ones [1]:

- Cleaning the training data. These methods exploit the fact that poisoned sets are usually different from normal training sets that are not controlled by attackers. Thus, data cleaning techniques are designed to clean the training set and remove the poisoned sets before performing machine learning training.
- Robust training. An alternative approach to mitigating availability attacks is to modify the ML learning algorithm and perform robust learning instead of regular learning. Several articles have identified robust optimization methods, such as using a reduced loss function [20] or random smoothing to add noise during training [21].

3. Evasion attack

Evasion Attacks are a type of cyberattack in which attackers modify input data to bypass a machine learning system and cause misclassification or false decisions. Such attacks usually occur at the stage of using the model, when it is already trained and deployed. Attackers make minor, often unnoticeable changes to the input data, which, however, significantly affect the model's performance.

The types of evasion attacks according to the NIST AI 100-2e2025 classification [25] include the following:

1. Gradient-based attacks: Attackers use information about the gradients of the model's loss function to determine the most effective input changes that will lead to misclassification [22].
2. Score-based attacks: Attackers gain access to the model's confidence scores and use optimization techniques to create fake examples that the model misclassifies.
3. Decision-based attacks: Attackers have access only to the final decisions of the model (e.g., class labels) and use optimization techniques to create fake examples that cause the model to make mistakes.
4. Transfer attacks [23]: Attackers train a replacement model, generate fake examples on it, and transfer these attacks to the target model by exploiting similarities between the models.

Evasion attacks pose a serious threat to cybersecurity systems. Attackers can change the characteristics of malware to bypass antivirus systems or modify network traffic to avoid detection by intrusion systems.

Currently, the main methods of protection [1] against evasion attacks are:

1. Adversarial training: Incorporating fake examples into the model training process to increase its resistance to attacks.
2. Use of ensembles of models (Ensemble methods) [24]: Combining multiple models to reduce the likelihood of a successful attack on all models at once.
3. Continuous monitoring and updating: Regularly updating detection models and systems to adapt to new attack strategies and improve resilience.

However, these methods have various limitations, such as reduced accuracy for adversarial learning and random smoothing, and computational complexity for formal methods. Therefore, a trade-off between robustness and accuracy should always be sought. Understanding and implementing these security techniques is critical to ensuring the security and reliability of machine learning systems in the face of increasing threats from evasion attacks.

4. Attack on privacy

Privacy attacks are a type of cyberattack aimed at obtaining sensitive information from artificial intelligence (AI) models, their training data, or output. These attacks can be used to steal personal data, compromise commercial information, or hack machine learning models.

Below, we'll look at the main types of privacy attacks:

1. Data Reconstruction Attacks: Attackers attempt to recreate the original data by accessing the AI model or its output. This can happen when the model produces overly detailed answers or has vulnerabilities that allow parts of the training data to be recovered.

2. Model Inversion Attacks: In this case, attackers use the output of the model to recreate the inputs or characteristics used in training. This can reveal sensitive information about the parties represented in the training data.

3. Membership Inference Attacks: Attackers try to determine whether specific records were included in the model's training dataset. This can be used to reveal a person's participation in certain events or membership in certain groups.

4. Metadata-based attacks: Even if the data itself remains secure, attackers can use metadata (e.g., access times, file sizes) to obtain sensitive information or establish patterns that can be used in further attacks.

5. Side-channel attacks: Attackers can analyze the behavior of an AI system, such as response time or energy consumption, to gain information about internal processes or model data.

It is recommended to protect against such attacks:

- Increasing anonymity and aggregating data. Use methods that reduce the risk of identifying individual records in training data.
- Federated Learning – training models on local devices without transferring data to the server.
- Differential privacy. Adding controlled noise to the data or model results to prevent the original data from being recovered without significantly affecting the model's accuracy.
- Access restriction and monitoring: Control access to models and data, and continuously monitor usage to detect suspicious activity.
- Vulnerability assessment: Regularly testing models for resistance to privacy attacks and implementing appropriate security measures.

Privacy attacks pose a serious threat to AI systems, including in the field of cybersecurity. Therefore, protecting privacy in AI systems is critical to maintaining user trust and regulatory compliance. Attackers use various methods to gain access to sensitive data or model parameters. Protecting such systems requires a comprehensive approach, including modern cryptographic methods, activity monitoring, and raising user awareness of risks.

5. Attack of abuse

Another type of attack on AI systems is abuse attacks [25]. These attacks are aimed at abusing or manipulating artificial intelligence (AI) systems to produce undesirable or harmful results. These attacks exploit vulnerabilities in the structure or implementation of AI models to cause the system to behave inappropriately.

Examples of abuse attacks:

1. Bias Exploitation: An attack in which an attacker exploits existing biases or weaknesses in an AI model to produce certain results or amplify discriminatory tendencies.

2. Functionality Misuse: Manipulating an AI system to perform actions that were not intended by the developers, such as using a chatbot to generate unwanted content or spam.

3. Prompt Injection Attacks: Injecting specially crafted queries or commands that cause the AI model to generate unwanted or malicious content.

To prevent such attacks, NIST recommends the following security measures:

- Improvement of anomaly detection algorithms – development of mechanisms that can recognize suspicious interactions with AI.
- Stricter data verification policies – analyzing incoming data to identify possible manipulations.
- Increase the transparency of AI systems by improving the documentation of decision-making processes in models.

- Mechanisms to counteract malicious intrusion, such as the introduction of additional levels of verification in security models.
- Development of standards for the ethical use of AI – active regulation and control over the implementation of such technologies.

Thus, misuse attacks pose a serious threat to AI systems, as they allow attackers to use them in unexpected ways. The use of AI to automate fraud, manipulate public opinion, or circumvent restrictions creates new challenges for cybersecurity. Preventing such attacks requires a comprehensive approach, including improving security algorithms, developing policies for the responsible use of AI, and continuously monitoring threats.

6. General summary of attacks on AI systems

In Sections 2-5 of this article, we have discussed the 4 most influential types of attacks on ML/AI systems, namely:

Data poisoning attacks – are carried out during the training phase and can have a long-term impact on the model's accuracy, causing it to draw incorrect conclusions.

Evasion attacks are runtime attacks where attackers try to trick the model by injecting specially created data into it.

Privacy attacks – aimed at extracting sensitive data from the model, which jeopardizes the security of users' personal information.

Abuse attacks are associated with the misuse of AI capabilities, for example, to create fake videos, malicious code, or manipulate social media.

The results of the analysis of the considered attacks are summarized in Table 1 below.

Table 1

Comparative characteristics of attacks on AI/ML systems

Type of attack	The purpose of the attack	The attack phase	Methods	Consequences
Poisoning Attack (data poisoning attack)	Impact on the quality of education	Model training	Adding malicious data to the training set	Deterioration in the quality of decisions, false alarms, and reduced security for further attacks
Evasion Attack (evasion attack)	Bypassing model security mechanisms	Execution of the model	Manipulation of source data, generation of special malicious data	Bypassing defense mechanisms, reducing model accuracy
Privacy Attack (privacy attack)	Theft of data used to train the model	Execution of the model	Analyzing model responses, data recovery	Confidential data leakage
Abuse Attacks (misuse attacks)	Using AI to create malicious data	After deploying the model	Generative models for attacks, manipulation	Information manipulation, fraud automation

Conclusions

1. AI is a good tool for automating the processes of detecting and responding to attacks and threats, and significantly increases the efficiency of protecting systems and companies. The use of machine learning algorithms helps to quickly analyze large amounts of data and identify anomalies in the behavior of users and systems.

2. However, AI not only provides effective protection, but also creates new threats due to its possible use by malicious actors. That is why the issue of detecting and counteracting attacks is a very important issue in the modern cyberspace. Therefore, in this article, we have reviewed and comprehensively analyzed attacks on modern ML/AI models.

3. The reliability of an AI system depends on all the attributes that characterize it. For example, an AI system that is accurate but easily susceptible to aggressive actions is unlikely to be trustworthy. Likewise, an AI system that produces harmfully biased or unfair results is unlikely to be trusted, even if it is reliable. There are also trade-offs between transparency and competitiveness. Unfortunately, it is not possible to maximize the performance of an AI system with respect to these attributes simultaneously. For example, AI systems optimized only for accuracy tend to underperform in terms of competitive reliability and fairness. Conversely, an AI system optimized for competitiveness may demonstrate lower accuracy and poorer reliability results.

4. Data poisoning attacks are the most dangerous type of AI-based attacks in the long run, as they affect the model itself and can remain undetected for a long time.

5. Privacy attacks and abuse attacks are particularly dangerous because of the possibility of leaking sensitive data and creating malicious content.

6. The study revealed the general consequences of AI-based attacks:

- Undermining trust in AI – constant attacks and manipulations can make AI a less reliable decision-making tool.
- Confidential information leakage – privacy attacks lead to large-scale losses of personal and corporate data.
- Defense bypassing – model evasion and poisoning jeopardize modern cybersecurity systems, reducing their effectiveness.
- Scalability of attacks – attackers can use AI to automate and accelerate attacks, which increases their impact.
- State-level risks – AI-based attacks can threaten national security, the economy, and critical infrastructure.

7. Therefore, the main ways to protect against AI-based attacks are as follows:

- Developing resilient AI models that are less vulnerable to poisoning or evasion.
- Implementation of attack detection mechanisms, such as monitoring changes in training data and algorithms.
- Increasing the transparency of AI by creating explainable models that can be checked for manipulations.
- Strengthening regulation and standards – governments and organizations should set rules for the use of AI.

8. Each type of attack on AI systems has different mechanisms of influence, but all of them can significantly reduce the efficiency and security of machine learning models. Protection against such attacks requires a comprehensive approach. In addition to the above methods of counteracting attacks, the human factor remains important – training of specialists and users, as many attacks are based on social engineering.

9. In most cases, organizations will have to make a trade-off between desirable attributes and decide which to prioritize depending on the AI system, use case, and potentially many other considerations regarding the economic, environmental, social, cultural, political, and global implications of AI technology.

10. It is important to note that with the development of AI technologies, new types of attacks are emerging, so constant monitoring and updating of knowledge in this area is essential to ensure cybersecurity.

11. To summarize, it is important to note that the safe use of AI in cybersecurity is a balance between technological innovations and threats arising from their development. The development of adaptive security methods and the improvement of machine learning models will help reduce the risks associated with attacks and ensure a reliable level of cyber defense in the future.

References:

1. Vassilev A., Oprea A., Fordyce A., Anderson H (2024) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Ar-

tifcial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2023. Access mode: https://doi.org/10.6028/NIST.AI.100-2e2023.

2. Booth H., Souppaya M., Vassilev A., Ogata M., Stanley M., Scarfone K. (2024) Secure Development Practices for Generative AI and Dual-Use Foundation AI Models: An SSDF Community Profile. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) NIST SP 800-218A. Access mode: https://doi.org/10.6028/NIST.SP.800-218A.

3. Oprea A., Singhal A. and Vassilev A. Poisoning Attacks Against Machine Learning: Can Machine Learning Be Trustworthy? // Computer. 2022. Vol. 55, no. 11. P. 94–99. doi: 10.1109/MC.2022.3190787.

4. Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, Zheng Wang. Physical Adversarial Attack Meets Computer Vision: A Decade Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. Vol. 46, no. 12. P. 9797–9817.

5. Anjan K. Koundinya S. S. Patil, Chandu B. R. Data Poisoning Attacks in Cognitive Computing // IEEE 9th International Conference for Convergence in Technology (I2CT). 2024. P.1–4.

6. National Institute of Standards and Technology. Artifcial Intelligence Risk Management Framework. 2023. (AI RMF 1.0). Access mode: https://doi.org/10.6028/NIST.AI.100-1.

7. Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84:317–331, December 2018.

8. Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning FAIL? generalized transferability for evasion and poisoning attacks // 27th USENIX Security Symposium (USENIX Security 18). 2018. P. 1299–1316.

9. Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines // Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML, 2012.

10. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks // NDSS. The Internet Society, 2018.

11. Kairouz, Peter; McMahan, H. Brendan; Avent, Brendan; Bellet, Aurélien; Bennis, Mehdi; Bhagoji, Arjun Nitin; Bonawitz, Kallista; Charles, Zachary; Cormode, Graham (June 22, 2021). Advances and Open Problems in Federated Learning // Foundations and Trends in Machine Learning 14 (1-2): doi:10.1561/2200000083. ISSN 1935-8237.

12. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks // NDSS. The Internet Society, 2018.

13. Kairouz, Peter, McMahan, H. Brendan, Avent Brendan, Bellet Aurélien, Bennis Mehdi, Bhagoji Arjun Nitin, Bonawitz Kallista, Charles Zachary, Cormode Graham (June 22, 2021). Advances and Open Problems in Federated Learning // Foundations and Trends in Machine Learning 14 (1-2). doi:10.1561/2200000083. ISSN 1935-8237.

14. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks // International Conference on Learning Representations, 2014.

15. Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples // International Conference on Learning Representations, 2015.

16. Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and testing unintended memorization in neural networks // USENIX Security Symposium, USENIX 19). 2019. P. 267–284. Access mode: https://arxiv. org/abs/1802.08232.

17. Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert - Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models // 30th USENIX Security Symposium (USENIX Security 21). 2021. P. 2633–2650. USENIX Association, August 2021.

18. Karan Ganju, Qi Wang, Wei Yang, Carl A. Property inference attacks on fully connected neural networks using permutation invariant representations // Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS '18, pages 619-633, New York, NY, USA, 2018. Association for Computing Machinery.

19. Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning FAIL? generalized transferability for evasion and poisoning attacks // 27th USENIX Security Symposium (USENIX Security 18). 2018. P. 1299–1316.

20. Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines // Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML, 2012.

21. Nihad Hassan. What is data poisoning (AI poisoning) and how does it work? Search Enterprise AI, TechTarget, 2024. Access mode: https://www.techtarget.com/searchenterpriseai/definition/data-poisoning-AI-poisoning.

22. Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In International Conference on Machine Learning. PMLR, 2019. P.1596–1606.

23. Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-fipping attacks via randomized smoothing. In International Conference on Machine Learning. PMLR, 2020. P. 8230–8241.

24. The Tactics & Techniques of Adversarial Machine Learning. HiddenLayer. 2022. Access mode: https://hiddenlayer.com/innovation-hub/the-tactics-and-techniques-of-adversarial-ml.

25. Chi Zhang, Zifan Wang, Ravi Mangal, Matt Fredrikson, Limin Jia, Corina Pasareanu. Transfer Attacks and Defenses for Large Language Models on Coding Tasks. November 22, 2023. Access mode: https://doi.org/10.48550/arXiv.2311.13445.
26. D. Li and Q. Li. Adversarial Deep Ensemble: Evasion Attacks and Defenses for Malware Detection // IEEE Transactions on Information Forensics and Security. June 30, 2022. Access mode: https://doi.org/10.48550/arXiv.2006.16545.
27. Vassilev A., Oprea A., Fordyce A., Anderson H. (2025) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artifcial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2025. Access mode: https://doi.org/10.6028/NIST.AI.100-2e2025.

Received 11.01.2025

Information about the authors:
Yuriy Golikov – CEO and Founder of DevBrother tech company, USA; e-mail: yuriy@devbrother.com; ORCID: https://orcid.org/0009-0008-7946-4663
Yelyzaveta Vadymivna Ostrianska – V.N. Karazin Kharkiv National University, junior researcher, JSC "IIT", information security systems analyst, Ukraine; e-mail: antelizza@gmail.com; ORCID: https://orcid.org/0000-0003-1412-8470