U.D.C 004.056.57

PROBLEMS RELATED TO THE DETECTION OF ATTACKS ON ARTIFICIAL INTELLIGENCE (AI) SYSTEMS

Sivkov D.I. (ITMO University)
Scientific Supervisor – Associate Professor (Qualification Category "Ordinary Associate Professor"), PhD Vorobeva A.A.
(ITMO University)

Introduction. Artificial intelligence (AI) has revolutionized the way information is processed and analyzed, making it possible to automate complex tasks and extract valuable information from huge amounts of data. However, as the use of AI systems becomes more widespread, it becomes increasingly important to address the security and privacy issues associated with these systems. One of the main problems is the ability of attackers to exploit vulnerabilities in AI systems to steal, manipulate or destroy confidential information. To reduce these risks, it is extremely important to develop effective methods for detecting and preventing threats to the information security of AI systems [1].

The main part. There are several problems associated with detecting attacks on AI systems [2], including:

1. Adversarial attacks: adversarial attacks aim to manipulate input data in a way that is difficult to notice, by using traditional security measures.
2. Complexity of AI systems: AI systems often have complex architectures and workflows, which make it difficult to identify the source of a security breach.
3. The changing nature of adversarial attacks: as AI systems become more widely used, attackers are constantly developing new and more sophisticated attack methods, making it difficult to track the changing threat landscape.
4. Limitations of existing protective mechanisms [3]: existing defense mechanisms are often designed to detect certain types of attacks, but they may not be sufficient to detect new or unknown attack methods.
5. Lack of transparency in AI systems: AI systems are often black boxes with little or no understanding of how they make decisions. The lack of transparency makes it difficult to detect and diagnose security breaches and assess the effectiveness of protective mechanisms.
6. The complexity of collecting tagged data: to detect attacks on AI systems, it is often necessary to have tagged data representing both normal and abnormal behavior. However, collecting tagged data can be difficult, especially for security-related events, since such events occur rarely and are often poorly documented.
7. Difficulty distinguishing attacks and normal behavior: it can be difficult to distinguish normal behavior from attacks, especially when attacks are designed to mimic normal behavior. This can lead to false alarms or missed detections.
8. The problem of the balance between false alarms and missed detections: to detect attacks on AI systems, it is necessary to find a balance between false alarms and missed detections. A high level of false positives can lead to a investigation which can take a

lot of time and money. Equally, a high level of missed detection can lead to missed security breaches, which can have serious consequences.

Conclusions. This article provides an analysis of current problems related to the detection of attacks on AI systems. The methods for solving the problems found are also given.

Sources:

1. Challenges in Adversarial Machine Learning // arxiv.org : site. – URL: https://arxiv.org/abs/1703.06857 (date of application: 14.02.2023)
2. A Comprehensive Survey on Safe and Secure Machine Learning // arxiv.org : site. – URL: https://arxiv.org/abs/1902.06861 (date of application: 14.02.2023)
3. SoK: Security and Privacy in Machine Learning // ieeexplore.ieee.org : site. – URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8406613 (date of application: 14.02.2023)