Asian Journal of Technology and Applied Sciences
ISSN: 2301-xxxx (online) | IF: 10.01 | Vol: 1, issue 1
Available: http://asiansciencejournal.org

ARTIFICIAL INTELLIGENCE-BASED CYBERSECURITY: THREAT DETECTION CAPABILITIES, ADVERSARIAL THREATS, AND AN INTEGRATED MODEL FOR RISK MANAGEMENT

ABDUXAMIDOV AZIZBEK ADILBEKOVICH
Lieutenant Colonel, Head of the Department of Advanced Military Technologies, Faculty of the Central Military District, University of Military Security and Defense of the Republic of Uzbekistan

Abstract: This article analyzes the capabilities of artificial intelligence (AI) in the realm of cybersecurity, as well as the adversarial threats inherent to it. The objective of the research is to propose a conceptual model that integrates threat detection, model security verification, human oversight, and continuous risk management. Utilizing systemic, comparative, and functional analysis, along with methods of generalization, the study examines materials from ENISA, NIST, and MITRE ATLAS. As a result, a five-layered model has been developed, encompassing asset and data inventory, AI-driven detection, adversarial testing, human oversight, and ongoing monitoring and improvement. This model is a conceptual proposal; its effectiveness requires further validation through empirical testing.

Keywords: artificial intelligence, cybersecurity, threat detection, adversarial risks, model security, risk management.

Introduction

As digital services, cloud infrastructure, connected devices, and data exchange expand, the cyber threats faced by organizations are becoming increasingly complex. While traditional rule-based and signature-based security tools are essential for detecting certain types of attacks, additional automation is required to analyze large volumes of log entries and interconnected events. Artificial intelligence (AI) and machine learning enable the identification of anomalies within network traffic, user behavior, and system logs, while also facilitating the prioritization of incidents and supporting the work of security analysts.

However, AI does not automatically guarantee cybersecurity. Issues such as poor data quality, false positives, adversarial attempts to mislead models, and privacy concerns can negatively impact security outcomes. Furthermore, AI systems themselves can become targets—whether through data poisoning, prompt injection, model misuse, or vulnerabilities within the supply chain. Consequently, alongside using AI to detect threats, it is equally critical to secure the AI components themselves. The research problem stems from the need to integrate these two areas—detecting cyber threats via AI and managing risks inherent to AI systems—into a unified management model.

Research Goal and Objectives

The goal of the research is to propose a conceptual model that integrates technical, organizational, and human oversight layers to ensure the safe and responsible implementation of AI-based cybersecurity solutions.

Research objectives:

To identify AI-enabled defensive capabilities and their inherent limitations.

To classify adversarial risks specific to AI systems.

To compare international approaches based on established criteria.

To analyze system functions and link them to risk management.

To synthesize the analysis results and propose an integrated model along with its evaluation criteria.

Asian Journal of Technology and Applied Sciences
ISSN: 2301-xxxx (online) | IF: 10.01 | Vol: 1, issue 1
Available: http://asiansciencejournal.org

Research Object and Subject

The object of the research is the processes for ensuring cybersecurity within digital infrastructure and information systems. The subject of the research is the application of AI tools in cybersecurity, as well as the technical and organizational mechanisms for their self-protection and risk management.

Analysis of Literature and International Approaches

The ENISA report, Artificial Intelligence Cybersecurity Challenges, examines the AI ecosystem through the lens of its lifecycle, emphasizing the need to identify assets requiring protection and to systematically classify threats. This approach implies that AI security must be considered not only during the model's operational phase but also throughout the stages of requirements definition, data preparation, development, deployment, and updates.

The NIST AI Risk Management Framework (AI RMF 1.0) offers a cross-sectoral approach to managing risks associated with AI systems. Furthermore, the NIST profile for generative AI links the specific risks inherent to generative systems with practical management strategies.

MITRE ATLAS is an open knowledge base that documents attack tactics and techniques—whether directed against AI or executed via AI. It can be utilized to translate threats into practical testing scenarios and to assess the security of a model.

These resources complement one another: ENISA focuses on AI cybersecurity challenges and the AI lifecycle; NIST addresses risk management and trustworthiness; and MITRE ATLAS concentrates on systematizing attack techniques targeting AI systems. In this article, these approaches are synthesized within a unified five-layered conceptual model designed for the sequence of their practical implementation.

Research Methodology and Methods

The research is conceptual and analytical in nature. Open-access documents from organizations such as ENISA, NIST, and MITRE, along with materials concerning AI and cybersecurity, were selected as sources. The selection criteria included the official status of the documents, their direct relevance to the subject matter, their coverage of threat or risk management issues, and the inclusion of references that allow for independent verification.

1. Systems Analysis Method

Systems analysis was employed to examine AI-based cybersecurity as an integrated system composed of interconnected components. The analysis identified key components: data sources, threat detection, model security, risk assessment, human oversight, and monitoring. Their interdependencies were examined: data quality influences detection outcomes; detection outcomes inform risk assessment; and risk assessments guide the selection of protective measures. Furthermore, monitoring results serve as a basis for re-evaluating data, models, and management processes.

This analysis demonstrates that the accuracy of an AI model cannot be considered the sole criterion for overall security. Comprehensive security also depends on data reliability, model robustness, the appropriateness of protective measures, and the capacity to respond to incidents.

2. Functional Analysis Method

Functional analysis was employed to examine the functions of an AI-based cybersecurity system and the contribution of each function to overall protection. The following functions were analyzed:

Threat detection: identifying suspicious indicators within network traffic, authentication logs, and endpoint data.

Asian Journal of Technology and Applied Sciences
ISSN: 2301-xxxx (online) | IF: 10.01 | Vol: 1, issue 1
Available: http://asiansciencejournal.org

Classification and prioritization: conducting an initial assessment of the severity level of detected incidents.

Model protection: mitigating risks such as data poisoning, adversarial attacks, prompt injection, and unauthorized use.

Decision support: assisting analysts by synthesizing evidence and reviewing potential response options.

Monitoring and auditing: tracking incorrect outputs, model drift, and the status of security controls.

Recovery and improvement: analyzing the root causes of incidents and addressing identified shortcomings.

The functional analysis revealed that if detection, risk assessment, decision validation, and monitoring are performed in isolation, gaps may emerge within the security process. For instance, if a threat is detected but its impact and priority are not assessed, the response may be delayed. Therefore, these functions must be interconnected within a unified management cycle.

4. Method of Generalization

The method of generalization was employed to synthesize conclusions derived from systemic, comparative, and functional analyses, and to formulate a conceptual model developed by the author. This process yielded three primary conclusions: first, AI enhances the capability to detect threats, though the outcomes depend on data quality and the operational conditions of the model; second, adversarial risks can emerge at various stages of an AI system's lifecycle; and third, technical detection tools must be integrated with human oversight, risk assessment, and incident response.

Based on these conclusions, a five-layered integrated model is proposed. This model is not an empirically tested solution, but rather a conceptual framework that synthesizes approaches found in international sources into a practical management sequence.

Capabilities of AI-based Cybersecurity

Anomaly Detection and Incident Prioritization

Machine learning models can assist in identifying deviations from the norm within network traffic, authentication logs, endpoint telemetry, and other observational data. They help analysts group large volumes of alerts and prioritize them according to the level of potential risk. However, an anomaly does not always signify an attack: new, yet legitimate, activity can also appear unusual. Therefore, the model's output should be interpreted as a signal for further investigation.

Threat Intelligence and Analyst Support

AI can accelerate the sorting of large volumes of technical reports, indicators, incident logs, and open-source data. Generative AI can draft initial incident summaries or investigative questions; however, any generated links, indicators, or causal explanations must be verified against reliable sources. Before sending sensitive data to external AI services, the organization's information security policy must be consulted.

Response Support

AI systems can assist in classifying incidents, assessing their impact on critical assets, and suggesting response options. However, high-impact actions—such as blocking accounts, isolating network segments, or halting services—must be executed in accordance with risk levels, authorization protocols, and pre-approved policies. The level of automation should be proportionate to the potential consequences of an erroneous decision.

Cyber Risks Specific to AI Systems

Data Poisoning and Adversarial Attacks on Models

If the data used to train or update a model is intentionally corrupted, the model may learn incorrect patterns. Adversarial influences, meanwhile, aim to mislead a model's decision by altering input data. Defensive measures may include verifying the provenance of data, implementing version control, restricting access to the training process, and conducting authorized adversarial testing.

Prompt Injection and Misuse of Generative Systems

External documents, messages, or user prompts can induce a generative system to execute instructions that contradict its intended purpose or policies. To mitigate this risk, controls are required—such as avoiding the treatment of untrusted text as commands, separating system instructions from data, granting only minimal permissions to tools, requiring confirmation for external actions, and validating system outputs.

Privacy and Data Leakage

AI systems may inadvertently disclose personal, commercial, or technical secrets. A security program should incorporate measures such as data minimization, the identification of sensitive information, the limitation of data retention periods, the management of access rights, and the protection of system logs.

Supply Chain and Third-Party Risks

Reliance on external models, libraries, datasets, APIs, or cloud services introduces an additional layer of risk. Organizations should maintain an inventory of components, evaluate the security practices of their vendors, monitor updates, and develop contingency plans for potential service disruptions.

False Alerts and Model Drift

Network and user behaviors change over time. If the model's initial training data does not reflect new conditions, detection performance may decline. Therefore, it is essential to regularly measure false positives and false negatives, monitor for model drift, and establish procedures for re-evaluation.

Proposed Approach: A Five-Layer Integrated Model

Based on a synthesis of the results from systemic, comparative, and functional analyses, a five-layer conceptual model for AI-driven cybersecurity management is proposed. This model integrates AI-based protection with the protection of the AI components themselves into a unified process.

Analysis shows that AI serves a dual role in cybersecurity: on one hand, it enhances capabilities for detecting and analyzing incidents; on the other, it can itself become a target or a tool for an attacker. Therefore, focusing solely on the accuracy of a detection model is insufficient. Data provenance, adversarial testing, integration with external systems, as well as defined authorities and responsibilities, are all integral components of the overall security outcome.

A comparative analysis revealed that various international resources serve distinct functions, yet when applied together, they form a complementary methodological framework. Functional analysis underscored the necessity of integrating threat detection, risk assessment, decision validation, and monitoring into a continuous cycle. Systems analysis clarified the interdependencies between these components, while a synthesis of these findings enabled their consolidation into a five-layer model.

The advantage of the proposed model lies in its integration of protection and management into a unified cycle. Its limitation stems from its conceptual nature: the model has not undergone empirical testing across different organizations, and there are no quantitative findings regarding its performance metrics. In the future, it will be necessary to

test the model in organizations of various sizes, compare it with traditional protection processes, and evaluate the costs of implementation.

Conclusion

This article has systematically examined the capabilities of artificial intelligence (AI) in cybersecurity, as well as the adversarial risks inherent to it. A systematic analysis revealed the interdependencies between data, detection algorithms, model security, and organizational oversight. A comparative analysis identified the distinct yet complementary roles of the approaches taken by ENISA, NIST, and MITRE ATLAS. A functional analysis highlighted the necessity of integrating detection, risk assessment, human oversight, and monitoring into a unified process. Based on these findings, a five-layered integrated conceptual model is proposed.

This model consolidates asset and data inventory, AI-driven detection, adversarial testing, human oversight, and continuous improvement into a single management cycle. Its scientific and practical contribution lies in synthesizing international approaches into a sequence for practical implementation and in systematizing evaluation criteria.

Organizations are advised to maintain an inventory of assets and data before deploying AI systems, conduct security testing on models and their integrations, ensure human oversight for high-impact decisions, regularly measure outcomes, and establish an improvement cycle to address identified shortcomings. Future empirical research should focus on determining the model's effectiveness, limitations, and costs within various organizational settings.

USED LITERATURE

1. European Union Agency for Cybersecurity (ENISA). (2020). Artificial Intelligence Cybersecurity Challenges. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges
2. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
3. Autio, C., Schwartz, R., Dunietz, J., et al. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1
4. MITER. (n.d.). MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.https://atlas.mitre.org/
5. National Institute of Standards and Technology (NIST). (n.d.). Cybersecurity, Privacy, and AI. https://www.nist.gov/itl/applied-cybersecurity/cybersecurity-privacy-and-ai