# ARTIFICIAL INTELLIGENCE IN CYBERSECURITY: OPPORTUNITIES, RISKS, AND RESPONSIBLE IMPLEMENTATION

**Baliyev Firdavs**
**student, faculty of Cybersecurity Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi**
**bolievfirdavs0@gmail.com**

https://doi.org/10.5281/zenodo.22075279

***Abstract:*** This article examines the dual role of artificial intelligence in cybersecurity. AI can help security teams analyze large event streams, detect suspicious behavior, prioritize vulnerabilities, and respond to incidents more quickly. At the same time, AI systems create new risks involving poisoned data, adversarial inputs, model theft, privacy leakage, prompt injection, and unsafe automation. The paper proposes a responsible implementation model that combines secure data, validated models, human oversight, continuous monitoring, and measurable governance.

***Keywords:*** artificial intelligence, cybersecurity, machine learning, threat detection, adversarial machine learning, AI governance.

**Introduction.** Modern organizations generate more security data than analysts can review manually. Authentication logs, endpoint events, network traffic, cloud activity, email, and vulnerability reports arrive continuously. Rule-based tools remain necessary, but they may miss changing behavior or produce too many alerts. Artificial intelligence and machine learning can identify patterns across this data, estimate risk, and direct attention to the events most likely to require investigation. AI is not an independent replacement for cybersecurity professionals. A model may be inaccurate, biased by incomplete data, manipulated by an attacker, or applied outside the conditions for which it was tested. The NIST AI Risk Management Framework therefore treats AI risk as a continuous process organized around the functions Govern, Map, Measure, and Manage [1]. In cybersecurity, this means evaluating both sides of the technology: how AI can strengthen defense and how the AI system itself must be protected.

**Main Body.** Applications of AI in cyber defense. AI is most useful when it supports clearly defined security tasks. Classification models can help identify phishing messages, malicious files, and suspicious domains. Behavioral analytics can compare current account or device activity with an established baseline. Natural-language systems can summarize alerts, extract indicators from reports, and help analysts search technical evidence. Predictive models can also rank vulnerabilities by combining severity, exposure, asset value, and evidence of exploitation. The principal advantage is prioritization rather than perfect prediction. A security operations center may receive thousands of alerts, while only a small number represent serious incidents. A well-tested model can score these events so analysts examine the highest-risk cases first. However, the original evidence must remain available, and important actions such as disabling accounts or isolating critical servers should follow approved rules and human review when the possible impact is high.

AI-supported detection and response workflow. A practical system begins with controlled data collection. Security events are normalized, time synchronized, labeled where possible, and protected from unauthorized changes. The model then produces a score or classification, but the decision layer also considers asset importance, identity, current threat intelligence, and

confidence. The final output should explain which signals influenced the result and route uncertain cases to an analyst. For example, a login from a new device is not automatically malicious. Risk increases if the event also uses an unusual location, occurs after repeated failures, accesses sensitive data, and differs from the user's normal behavior. The system can request stronger authentication, restrict the session, or open an investigation. Analyst feedback then becomes part of controlled evaluation data, allowing the team to measure false positives, missed detections, and changes in attacker behavior without allowing unreviewed feedback to corrupt the model. AI must operate beside conventional controls. Multi-factor authentication, patching, network segmentation, backups, least privilege, encryption, and incident-response procedures remain essential. If the AI service fails, the organization should continue operating through safe defaults and documented manual procedures. This layered design prevents one model error from becoming a complete security failure.

Cybersecurity risks of AI systems. Attackers can target the data, model, software pipeline, interface, or human operator. Data poisoning changes training or feedback records to influence future decisions. Evasion uses carefully modified input to avoid detection. Model extraction attempts to reproduce a system through repeated queries, while privacy attacks seek information about training data. NIST's adversarial machine learning taxonomy groups such attacks by lifecycle stage, attacker goals, knowledge, and capabilities [2].

Generative AI introduces additional concerns. Prompt injection may cause a tool to ignore intended instructions, reveal protected context, or invoke connected functions incorrectly. Confident but false output can mislead an analyst, and sensitive information may be exposed when users paste logs, source code, or credentials into an unauthorized service. NIST's Generative AI Profile emphasizes risk management across the lifecycle rather than relying on a single technical control [3]. Access to models, prompts, retrieval stores, tools, and output logs should therefore be separated and monitored. Security controls should include data provenance, integrity checks, model and dependency inventories, restricted administrative access, version control, adversarial testing, rate limits, output validation, and rollback capability. Models should be tested with realistic malicious inputs before deployment and after significant updates. ENISA identifies poisoning, adversarial attacks, and data exfiltration among the important threats to machine-learning systems and recommends controls across their critical stages [4].

Governance and responsible implementation. Each AI security use case should have a named owner, a documented purpose, approved data sources, acceptable error thresholds, and a procedure for appeal or override. Privacy impact, retention, access rights, and legal requirements must be reviewed before training or connecting operational logs. Performance should be measured separately across relevant environments because an accurate laboratory model may behave differently in production. Deployment should follow secure-by-design principles. Joint guidance published through CISA recommends protecting infrastructure and models, applying secure authentication and access controls, continuously monitoring system behavior, and maintaining incident-management processes for externally developed AI systems [5]. Procurement must therefore examine the provider's update process, dependency security, logging, data handling, vulnerability disclosure, and support for independent evaluation.

Phased adoption model. The first phase selects one low-impact, measurable task, such as phishing triage or alert grouping. The second establishes a baseline using historical data and analyst decisions. The third tests accuracy, false-positive rate, detection coverage, latency, resilience to manipulation, and operational cost. The fourth runs the model in advisory mode before limited automation is allowed. Expansion should occur only when monitoring, rollback, staff training, and accountability are proven.

Conclusion. Artificial intelligence can improve cybersecurity by converting large volumes of technical evidence into prioritized and timely information. Its value is strongest in repetitive analysis, anomaly detection, classification, summarization, and decision support. These capabilities can reduce response time and help limited security teams focus on complex investigations. The same technology creates a new attack surface and cannot be treated as automatically trustworthy. Responsible adoption requires secure data, tested models, conventional security controls, human oversight, transparent metrics, and continuous governance. Organizations that begin with narrow use cases and expand only after measurable validation can gain the benefits of AI without transferring uncontrolled risk into their cybersecurity operations.

## Adabiyotlar, References, Литературы:

1. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1
2. Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2025. https://doi.org/10.6028/NIST.AI.100-2e2025
3. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1
4. European Union Agency for Cybersecurity (ENISA). (2021). Securing Machine Learning Algorithms. https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms
5. Yesbosinova, N. (2026). The pedagogical and psychological essence of developing teachers' writing skills. в international conference on health & technology (Т. 2, Выпуск 1, сс. 10–12). Zenodo. https://doi.org/10.5281/zenodo.18194399