Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026404411

AI and Cybersecurity: Transforming Digital Defense and Risk

Juncheng Zhu*

Dulwich International High School Suzhou, Suzhou, China

**Corresponding author: Juncheng Zhu.*

Abstract

As digital transformation accelerates and the complexity of cyberattacks continues to rise, cybersecurity has become a critical challenge for governments, businesses, and individuals. At the same time, the rapid development of Artificial Intelligence (AI) is profoundly transforming the field of cybersecurity. AI not only enhances the efficiency of threat detection, risk analysis, and incident response but also provides new technological tools for cyber defense. However, the widespread adoption of AI technologies also introduces new risks, including automated cyberattacks, deepfakes, privacy violations, and algorithmic bias. This study aims to explore how AI is transforming the field of cybersecurity and analyze its associated benefits, risks, and future challenges. Using a literature review approach, it comprehensively evaluates recent academic research, industry reports, and case studies on AI and cybersecurity to compare the distinct roles of AI in cyber defense and cyber-attacks. Research has found that AI can significantly enhance the speed and accuracy of detecting cyber threats, strengthen predictive defense capabilities, and reduce the workload for security personnel. However, attackers can also leverage AI to develop smarter and more covert attack methods, thereby increasing cybersecurity risks. Additionally, AI faces governance challenges in its application, including transparency, accountability, privacy protection, and inadequate regulation. This study holds that AI is both a crucial tool for enhancing cybersecurity capabilities and a potential source of new security threats. In the future, technological innovation, legal regulation, and international cooperation will be essential to ensure the safe, reliable, and responsible application of AI in cybersecurity, thereby promoting the long-term stable development of the digital society.

Keywords

threat avoidance, intrusion detection, large language models

1. Introduction

With the rapid advancement of digital transformation, cloud computing, and big data technologies, modern society's reliance on the network continues to grow. However, at the same time, the complexity and destructive power of cyberattacks are reaching an unprecedented rate. From ransomware attacks and data breaches to state-sponsored cyber warfare, cybersecurity has become a critical challenge for governments, businesses, and individuals alike.

Traditional cybersecurity systems primarily rely on rule-based detection mechanisms and manual analysis. Yet, as attack methods evolve continuously, these conventional approaches have revealed limitations such as slow response times, high false positive rates, and difficulties in handling massive amount of data. In this context, artificial intelligence is being increasingly applied in the field of cybersecurity. Leveraging technologies such as machine learning (ML), deep learning (DL), and large language models (LLMs), AI can automatically analyze vast amounts of network data, identify anomalous behaviors, and enable real-time threat detection and response.

Kaur, Gabrijelčič, and Klobučar [1] inform the public that AI is capable of helping cybersecurity teams streamline repetitive tasks, speeding up threat detection and response, and making their actions more reliable. This perspective reflects the core value of AI in modern cybersecurity. With its automated analytical capabilities, AI can complete tasks within seconds that previously required security analysts several hours or even days to accomplish. For example, AI-powered intrusion detection systems (IDS) can monitor network traffic in real time and identify potential attack patterns, while threat intelligence systems leverage machine learning to predict future attack trends, enabling organizations to adopt proactive defense measures.

However, AI is not merely a defensive tool. In recent years, generative AI and large language models—represented by technologies such as ChatGPT, GPT-4, Llama, and Gemini—have rapidly gained widespread adoption. While these technologies enhance cybersecurity efficiency, they also provide attackers with new tools. Malicious actors can use AI to generate more convincing phishing emails, automatically write malicious code, or even carry out social engineering attacks using deepfake technology.

Ferrag et al. point out that large language models (LLMs) are not only useful for threat detection and malware analysis but also introduce novel security risks such as “prompt injection” and “data poisoning” [2]. Research indicates that the capabilities of LLMs make them equally valuable as tools for both defense and offense.

Thus, AI in cybersecurity exhibits a clear double-edged sword characteristic. On one hand, it enhances threat detection speed, strengthens predictive capabilities, and reduces labor costs; on the other hand, it may be maliciously exploited, thereby amplifying the scale and impact of cyberattacks. This dual nature makes studying the impact of AI on cybersecurity highly significant both theoretically and practically.

Based on these findings, this paper addresses the following research questions: Research Question: How has AI transformed cybersecurity, and how are its benefits compared to the drawbacks?

2. Theoretical Background

In recent years, interdisciplinary research at the intersection of artificial intelligence and cybersecurity has grown rapidly, gradually becoming a key area of study in information security. Existing literature primarily focuses on three aspects: AI as a defensive tool (Defensive AI), applications of generative AI and large language models (Generative AI and LLMs), and ethical and governance concerns. To understand how AI is transforming modern cybersecurity, it is necessary to first introduce several foundational theories supporting this research. These theories not only form the core framework of modern cybersecurity systems but provide a theoretical basis for analyzing the advantages, risks, and future challenges of AI in cybersecurity.

To start with, the CIA Triad is the most classical and fundamental theoretical framework in information security, consisting of three core objectives. Confidentiality ensures that information is accessible only to authorized users. Integrity requires data to remain accurate and unchanged by unauthorized parties during storage, transmission, and processing; and availability mandates that systems and data remain continuously accessible when needed.

Secondly, this study adopts the Cyber Kill Chain as a key theoretical framework for analyzing the process of cyberattacks. Proposed by Lockheed Martin, this model divides a complete cyberattack into several stages: Reconnaissance, Weaponization, Delivery, Exploitation, Installation, Command and Control (C2), and Actions on Objectives [3]. This theory posits that cyberattacks are not isolated incidents but rather continuous processes, with each stage presenting distinct opportunities for defense.

Lastly, this study introduces the Dual-use Technology theory to explain the dual impact of AI in cybersecurity. Dual-use technology refers to a technology that can generate positive outcomes but may also be

exploited for destructive purposes, with its ultimate effect depending on the user and the manner of application. AI is currently one of the most typical examples of such dual-use technology.

In summary, the CIA Triad, Cyber Kill Chain, and Dual-use Technology together form the theoretical foundation of this paper. The CIA Triad is used to assess whether AI enhances the achievement of cybersecurity objectives; the Cyber Kill Chain helps analyze how AI transforms the methods of cyber-attacks and defenses; and Dual-use Technology explains why AI can simultaneously strengthen defensive capabilities while creating new security risks. These foundational theories provide a unified analytical framework for subsequent literature review and case analysis, and also facilitate a systematic understanding of the opportunities and challenges brought by AI in the field of cybersecurity.

3. Literature Review

3.1 AI as a Cybersecurity Defense Tool

Early research on AI and cybersecurity mainly focused on leveraging machine learning to improve threat detection efficiency. Traditional security systems rely on predefined rules to identify attacks, whereas AI can learn attack patterns from historical data and detect unknown threats.

Ezugwu et al. noted in their review that AI technologies have been widely applied in areas such as malware detection, anomaly detection, intrusion detection, and authentication [4]. Studies show that compared to traditional methods, machine learning models significantly enhance detection accuracy while reducing false positive rates.

In one of the most influential review studies to date, Kaur et al. systematically analyzed 236 core research papers and categorized AI applications according to the National Institute of Standards and Technology (NIST) Cybersecurity Framework [5]. The study found that AI has already covered five critical stages in the cybersecurity lifecycle, which are identify, protect, detect, respond, and recover.

The authors argue that AI can help organizations transition from reactive defense to proactive defense.

Meanwhile, Sarker pointed out that as the volume of data in cyberspace continues to grow, human analysts are increasingly unable to independently handle all security analysis tasks [6]; thus, AI-driven data analytics and automation have become essential directions for future cybersecurity development.

3.2 Applications of Generative AI and LLMs in Cybersecurity

Since 2022, generative AI—represented by ChatGPT—has advanced rapidly, propelling cybersecurity research into a new phase. Compared to traditional machine learning systems, large language models (LLMs) can not only analyze structured data but also process natural language text, code, and threat intelligence information.

Ferrag et al. highlighted that LLMs demonstrate significant potential across multiple cybersecurity tasks, including: Cyber Threat Intelligence, Malware Detection, Phishing Detection, Software Security Analysis, and Vulnerability Assessment [7].

Researchers believe that large language models (LLMs) can assist security analysts in quickly retrieving information through natural language interaction, thereby improving the efficiency of threat analysis. Similarly, Zhang et al. analyzed over 300 related studies in “When LLMs Meet Cybersecurity” and found that large language models have been widely applied in areas such as vulnerability analysis, malware reverse engineering, and automated security operations [8]. The authors suggest that LLMs are poised to become a key auxiliary tool for future Security Operations Centers (SOCs). Xu et al. (2024) further demonstrated in their systematic review that LLM4Security has emerged as an emerging research hotspot, aiming to achieve higher levels of automated security analysis and decision support in complex threat environments [9].

However, these studies also highlight that large language models are not entirely reliable. Issues such as prompt injection, data poisoning, and hallucination may lead to incorrect analytical outcomes, thereby affecting security decisions.

3.3 Ethical and Governance Challenges

Beyond technical applications, an increasing number of studies are beginning to focus on ethical and governance challenges in AI-driven cybersecurity. Charmet et al. (2022), in their research on Explainable Artificial Intelligence (XAI), suggest that many AI-based security systems operate as "black box" models, lacking transparency in their decision-making processes [10]. This makes it difficult for organizations to understand why a system classifies a particular behavior as an attack, undermining user trust and accountability.

Meanwhile, the advancement of generative AI has further intensified governance challenges. Studies show that attackers can exploit AI to automatically generate phishing emails, malware, and disinformation, significantly lowering the cost of cyberattacks. Yigit et al. (2024) noted that generative AI can both enhance defensive capabilities and enable automated hacking, social engineering fraud, and malicious code generation [11]. Therefore, future research must establish more comprehensive ethical guidelines and governance mechanisms.

4. Key Finding

4.1 Benefits of AI in Cybersecurity

Artificial Intelligence (AI) has become an essential component of modern cybersecurity frameworks. As the cyber environment grows increasingly complex, traditional security defense models that rely on manual analysis and rule-based matching are gradually revealing shortcomings. In contrast, AI can leverage machine learning (ML), deep learning (DL), and large-scale data analytics to deliver more efficient solutions in real-life scenarios.

First, AI significantly enhances the speed and accuracy of detecting cyber threats. Modern enterprise networks generate vast amounts of log data, making it difficult for security analysts to manually monitor all activities in real time. AI systems can continuously analyze user behaviors and network traffic to identify anomalous patterns and quickly detect potential attacks.

This perspective indicates that AI not only enhances detection efficiency but also reduces the impact of human error through automated processing. Particularly when dealing with advanced persistent threats (APTs), AI can identify complex attack patterns that are difficult for traditional rule-based systems to detect.

Second, AI promotes automation in cybersecurity operations (Security Automation). Security Operations Centers (SOCs) typically handle a large volume of security alerts, many of which are false positives or low-risk incidents. AI-driven automation systems can classify, prioritize, and conduct preliminary analysis of alerts, thereby reducing the workload on security analysts.

In recent years, Security Orchestration, Automation, and Response (SOAR) platforms have widely leveraged AI technologies to enable automated response. For instance, when a system detects anomalous login behavior, AI can automatically trigger account lockout, multi-factor authentication verification, and incident investigation workflows without requiring human intervention. This capability significantly shortens response times and improves overall organizational defense efficiency.

Moreover, the emergence of generative AI and large language models (LLMs) has further advanced security automation. Ferrag et al. in "Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities" point out that LLMs are already being applied across multiple domains, including threat intelligence analysis and malware detection [2].

Third, AI strengthens predictive cybersecurity capabilities. Traditional cybersecurity often follows a reactive model—detecting and mitigating attacks after they occur. In contrast, AI can analyze historical attack data, vulnerability information, and attacker behavior patterns to forecast future threats. Machine learning models can identify attack trends and proactively uncover potential risks within systems. For example, by analyzing user behavior data, AI can detect early signs of insider threats; by studying the evolutionary patterns of malware families, AI can predict likely characteristics of emerging malware variants. This shift from "post-incident response" to "preemptive prevention" represents a critical direction in modern cybersecurity.

Overall, AI significantly enhances cybersecurity defenses by improving detection accuracy, automating security operations, strengthening predictive analytics, and boosting system scalability. As a result, an increasing number of organizations are recognizing AI as a core component of their future cybersecurity strategies.

4.2 Risks and Negative Consequences

Although AI brings significant advantages to cybersecurity, its development has also created new security risks. AI is inherently a "dual-use technology," capable of enhancing defenders' capabilities while also being exploited by attackers to improve the efficiency of their attacks. Thus, the widespread adoption of AI in cybersecurity does not mean reduced risk; instead, it fuels a new technological arms race between offense and defense.

To start with, generative AI has dramatically lowered the technical barrier for cyberattacks. In the past, launching high-quality phishing attacks typically required attackers to possess strong language skills and social engineering expertise. However, modern large language models (LLMs) can now automatically generate grammatically correct, logically coherent, and highly personalized phishing emails.

Secondly, the advancement of deepfake technology poses serious challenges to digital identity security. Deepfakes leverage generative adversarial networks (GANs) to produce highly realistic audio and video content, enabling attackers to impersonate individuals convincingly. AI can assist attackers in developing more sophisticated malware. Traditional malware often relies on fixed code and known attack patterns, making it relatively easy for security software to detect. In contrast, AI-powered malware can dynamically adapt its behavior based on the environment and even learn ways to evade detection systems.

Malatji and Tolah's research on adversarial AI and offensive AI highlights that AI can be used to automate vulnerability discovery, plan attack paths, and optimize malicious code, thereby increasing the level of attack automation [12]. In the future, attackers may employ autonomous AI agents to carry out entire attack cycles—from reconnaissance to infiltration—making cyberattacks more scalable and intelligent. Related studies also emphasize the challenges posed by offensive AI systems to existing security frameworks.

Therefore, AI is not merely a defensive tool—it has also become a new attack surface. As both attackers and defenders increasingly adopt AI technologies, the cybersecurity landscape is evolving into a continuously escalating "AI arms race." Future competition will no longer be solely between human experts, but rather a contest among AI systems themselves.

4.3 Ethical, Legal, and Governance Challenges

Beyond technical risks, the application of AI in cybersecurity raises a series of ethical, legal, and governance issues. If these challenges are not properly addressed, AI could enhance security while simultaneously undermining privacy, fairness, and social trust.

Firstly, privacy protection is particularly prominent. AI models typically require vast amounts of data for training and continuous optimization, including user behavior data, network traffic information, and personally identifiable information. To improve threat detection capabilities, organizations often collect and analyze extensive records of user activities. However, such practices may infringe upon individual privacy rights and increase the risk of data misuse.

Another aspect is that AI models may suffer from bias and false positives. The quality of machine learning systems' decisions heavily depends on their training data. If training data is imbalanced or contains historical biases, AI may produce unfair outcomes. For example, certain legitimate user behaviors might be incorrectly flagged as malicious, resulting in account suspension or access denial.

In summary, the development of AI in cybersecurity involves not only technological challenges but also comprehensive ethical, legal, and governance concerns. Only through synchronized progress in technological innovation and institutional development can AI achieve truly secure, trustworthy, and sustainable growth.

5. Discussion

5.1 Application

From a practical application perspective, AI has been widely applied in various fields such as enterprise Security Operations Center (SOC), Threat Intelligence, malware analysis, and automated security operations. For large organizations, AI can significantly reduce security operation costs, improve the efficiency of handling massive security incidents, and alleviate the problem of the global shortage of cybersecurity professionals. However, there are significant differences in data quality, computing resources, and professional talent among different industries and organizations of different sizes. Small and medium-sized enterprises may find it difficult to bear the deployment and maintenance costs of high-performance AI models. Therefore, the large-scale application of AI in the field of cybersecurity in the future not only requires continuous technological improvement but also needs more economical, scalable, and easy-to-deploy solutions.

5.2 Research Limitation

At the same time, this study also has certain limitations. This paper mainly adopts the literature review method to conduct a comprehensive analysis of the representative studies on AI and cybersecurity in recent years. However, it lacks support from actual case studies, experimental data, or quantitative research. Therefore, this paper mainly reflects the current development trends in the academic and industrial fields, but cannot directly verify the specific effects of different AI technologies in real network environments. Moreover, due to the extremely rapid development of artificial intelligence technology, especially the continuous rapid evolution of Generative AI, Frontier AI, and Autonomous AI Agents, some research conclusions may change with the development of new technologies. Therefore, the analysis of this study has a certain time limitation.

5.3 Future Research Direction

Given these limitations, future research can be further conducted in multiple directions. Firstly, it can conduct empirical comparisons of the performance of different AI models in tasks such as Threat Detection, Malware Analysis, and Incident Response by integrating real enterprise network environments, in order to verify the practical application effects of different technical solutions. Secondly, as LLMs gradually become important tools in cybersecurity, future research should pay more attention to security issues targeting AI models themselves, such as Prompt Injection, Data Poisoning, Model Hallucination, and Adversarial Attacks, and explore more reliable model protection mechanisms. Moreover, AI governance (AI Governance) remains an important research direction in the future, including establishing a unified international regulatory framework, improving the liability determination mechanism, and promoting the development of mechanisms such as Explainable Artificial Intelligence (XAI) and Human-in-the-Loop, in order to enhance the transparency, credibility, and social acceptance of AI systems.

Overall, AI is driving the field of cybersecurity into a more intelligent and automated development stage. However, the technological advantages can only truly be transformed into long-term, stable and trustworthy cybersecurity capabilities with the improvement of governance mechanisms, continuous human supervision, and the joint support of international cooperation. This also means that the development of cybersecurity in the future not only depends on the progress of AI technology itself, but also on the coordinated development of technology, systems and ethics.

6. Conclusion

In conclusion, artificial intelligence has profoundly transformed the operation mode of modern cybersecurity and has become an important technological force driving the enhancement of network defense capabilities. Through automated threat detection, predictive analytics, real-time response, and security operations automation, AI significantly improves the efficiency, accuracy, and scalability of cybersecurity systems.

Regarding the research question of this paper – “How does AI change cybersecurity and how to compare its advantages and disadvantages”, the research results show that the overall benefits brought by AI to cybersecurity outweigh its potential risks. AI significantly enhances threat detection and response capabilities

and effectively alleviates the problem of a shortage of cybersecurity professionals. However, these advantages can only be fully utilized under reasonable regulation and effective supervision. In other words, AI is not the ultimate solution to cybersecurity problems but a powerful tool that needs to be managed carefully and continuously optimized.

In the future, cybersecurity will increasingly rely on AI-driven technological systems. But the key to future success does not lie in simply pursuing more powerful AI models, but in achieving a balance between technological innovation and governance frameworks. Only by adhering to the principle of "equal emphasis on technology development and human supervision" can we ensure that artificial intelligence enhances cybersecurity capabilities while maintaining social trust, protecting individual rights, and promoting the long-term stable development of the digital society.

References

[1] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions," Information Fusion, vol. 97, Art. no. 101804, Sep. 2023. doi: 10.1016/j.inffus.2023.101804. (sciencedirect.com)

[2] M. A. Ferrag, N. Battah, N. Tihanyi, L. Csendes, M. Debbah, and T. Lestable, "Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities"

[3] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," in Proceedings of the 6th International Conference on Information Warfare and Security (ICIW 2011), Washington, DC, USA, 2011, pp. 113–125.

[4] A. E. Ezugwu, L. Abualigah, J. O. Agushaka, et al., "The Impacts of Artificial Intelligence Techniques in Augmentation of Cybersecurity: A Comprehensive Review," Complex & Intelligent Systems, vol. 8, pp. 1769–1808, 2022.

[5] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions," Information Fusion, vol. 97, Art. no. 101804, Sep. 2023.

[6] I. H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues toward Automation, Intelligent and Smart Systems," SN Computer Science, vol. 3, Art. 158, 2022.

[7] M. A. Ferrag, N. Battah, N. Tihanyi, L. Csendes, M. Debbah, and T. Lestable, "Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities"

[8] C. Zhang, Y. Wang, X. Liu, et al., "When LLMs Meet Cybersecurity: A Systematic Literature Review"

[9] X. Xu, Y. Zhang, Z. Li, et al., "Large Language Models for Cyber Security: A Systematic Literature Review"

[10] F. Charmet, G. Srivastava, R. H. Jhaveri, et al., "XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions"

[11] Y. Yigit, W. J. Buchanan, M. G. Tehrani, and L. Maglaras, "Review of Generative AI Methods in Cybersecurity"

[12] M. Malatji and N. Tolah, "Artificial Intelligence (AI) Cybersecurity Dimensions: A Comprehensive Framework for Understanding Adversarial and Offensive AI," Discover Artificial Intelligence, vol. 4, 2024.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.