GSC Advanced Research and Reviews
eISSN: 2582-4597 CODEN (USA): GARRC2
Cross Ref DOI: 10.30574/gscarr
Journal homepage: https://gsconlinepress.com/journals/gscarr/

(REVIEW ARTICLE)

The Rise of AI in Cyber Defense: Opportunities and New Threats

Omowunmi Folashayo Makinde *

Department of Information Systems Security, University of the Cumberlands, Williamsburg, KY, USA.

GSC Advanced Research and Reviews, 2025, 25(01), 146-155

Publication history: Received on 14 September 2025; revised on 19 October 2025; accepted on 23 October 2025

Article DOI: https://doi.org/10.30574/gscarr.2025.25.1.0320

Abstract

Artificial intelligence has fundamentally transformed the cybersecurity landscape, introducing both unprecedented defensive capabilities and novel attack vectors. Organizations now leverage machine learning algorithms, neural networks, and automated response systems to detect and neutralize threats at speeds impossible for human analysts. However, this technological evolution has simultaneously empowered adversaries with sophisticated tools for launching adaptive attacks, evading detection systems, and exploiting vulnerabilities at scale. Key opportunities include real-time threat detection, predictive analytics, automated incident response, and behavioral anomaly identification. Conversely, emerging threats encompass adversarial machine learning, AI-powered social engineering, automated vulnerability exploitation, and algorithmic bias in security systems. This article examines the dual nature of AI in cybersecurity, analyzing how organizations can maximize defensive benefits while mitigating risks associated with AI-enabled attacks. A balanced approach combining human expertise with machine intelligence, continuous model validation, and ethical AI governance frameworks proves essential for maintaining security in an increasingly automated threat landscape.

Keywords: Artificial Intelligence; Machine Learning; Cyber Defense; Threat Detection; Adversarial AI; Automated Security

1. Introduction

The integration of artificial intelligence into cybersecurity represents one of the most significant technological shifts in digital defense strategies over the past decade. As cyber threats grow in sophistication and volume, traditional security approaches that rely primarily on human analysis and signature-based detection have become increasingly inadequate. Organizations now face attack surfaces that expand exponentially with cloud adoption, Internet of Things devices, and remote work infrastructures, creating security challenges that exceed human capacity to monitor and respond effectively (Ofusori et al., 2024).

The cybersecurity industry has witnessed remarkable growth in AI adoption over recent years. Security teams now deploy machine learning algorithms to analyze network traffic patterns, identify malicious behavior, and predict potential vulnerabilities before attackers can exploit them. These systems process terabytes of security data daily, correlating events across distributed infrastructure and detecting threats that would remain invisible to human analysts working with traditional tools. Natural language processing enables automated analysis of threat intelligence feeds, while computer vision techniques identify visual indicators of compromise in security camera footage and user interface manipulations. The promise of AI in cybersecurity extends beyond mere automation to fundamentally enhanced capabilities that transform how organizations approach security challenges (Salem et al., 2024).

However, this technological revolution presents a paradox. The same AI capabilities that strengthen defensive postures also empower adversaries with unprecedented attack tools. Cybercriminals now deploy machine learning algorithms

* Corresponding author: Omowunmi Folashayo Makinde

to identify vulnerabilities, craft personalized phishing campaigns, and develop malware that adapts to evade detection systems. Nation-state actors leverage AI to conduct reconnaissance at scale, automate social engineering attacks, and create deepfakes for disinformation campaigns. The democratization of AI tools has lowered technical barriers to sophisticated attacks, enabling less skilled adversaries to launch operations previously requiring expert knowledge (Mohamed, 2025).

The cybersecurity community now confronts a fundamental question: how can organizations harness AI's defensive potential while protecting against AI-enabled threats? This challenge extends beyond technical implementation to encompass strategic considerations about resource allocation, workforce development, and ethical frameworks for autonomous security systems. The transformation of cybersecurity through AI represents not merely an incremental improvement but a fundamental shift in how organizations conceptualize and implement digital defense, demanding corresponding changes in security architectures, operational processes, and organizational cultures. (Achuthan et al., 2024).

1.1. The Evolution of AI in Cybersecurity

The application of artificial intelligence to cybersecurity has progressed through several distinct phases, each characterized by increasing sophistication and broader implementation. Early applications focused primarily on automating routine tasks such as log analysis and signature matching, providing incremental efficiency improvements over manual processes. These initial systems operated within narrowly defined parameters, executing predetermined responses to recognized patterns without genuine learning capabilities (Achuthan et al., 2024).

The emergence of machine learning marked a significant advancement, enabling security systems to identify patterns across vast datasets and detect anomalies that deviated from established baselines. Organizations began deploying supervised learning models trained on labeled datasets of malicious and benign activities, achieving detection rates that surpassed traditional signature-based approaches. These systems demonstrated particular effectiveness in identifying variants of known threats, recognizing malicious patterns even when specific signatures had been modified to evade detection (Mohamed, 2025).

Contemporary AI security systems incorporate deep learning architectures capable of processing unstructured data, including network traffic patterns, user behaviors, and system logs, to identify sophisticated threats. Neural networks with multiple hidden layers can extract complex features from raw data, identifying subtle indicators of compromise that simpler algorithms would miss. Natural language processing enables analysis of textual data from emails, chat messages, and social media to detect social engineering attempts and insider threats (Ofori et al., 2025).

Recent developments have introduced reinforcement learning approaches that enable security systems to improve through interaction with their environments. These systems learn optimal response strategies by receiving feedback on the effectiveness of their actions, gradually developing sophisticated decision-making capabilities. The current frontier involves autonomous security systems capable of detecting, analyzing, and responding to threats with minimal human intervention. However, this increasing autonomy raises important questions about accountability, transparency, and the appropriate balance between machine decision-making and human oversight in security operations (Adawadkar and Kulkarni, 2022).

2. Opportunities in AI-Powered Cyber Defense

2.1. Advanced Threat Detection and Analysis

Artificial intelligence has revolutionized threat detection by enabling security systems to process and analyze data at scales and speeds impossible for human analysts. Modern AI-powered security platforms can monitor millions of endpoints simultaneously, analyzing network traffic, system logs, and user behaviors in real time to identify potential threats. Machine learning algorithms excel at recognizing patterns across vast datasets, detecting subtle anomalies that might indicate compromise even when individual indicators appear benign in isolation (Ofusori et al., 2024).

Behavioral analytics powered by AI have proven particularly effective at identifying insider threats and compromised accounts. By establishing baseline patterns of normal user behavior, these systems can detect deviations that suggest malicious activity or account takeover. An employee who suddenly accesses sensitive files outside their normal responsibilities, logs in from unusual locations, or exhibits atypical data transfer patterns triggers alerts for security teams to investigate. This capability addresses one of the most challenging aspects of cybersecurity, as insider threats and compromised credentials often evade traditional perimeter defenses (Ye et al., 2025).

Deep learning models have demonstrated remarkable effectiveness in malware detection, achieving accuracy rates that exceed traditional signature-based approaches. These neural networks analyze file characteristics, behavioral patterns, and code structures to identify malicious software, including previously unknown variants. Unlike signature-based systems that require updates for each new threat, AI models can recognize malware families based on structural and behavioral similarities, providing protection against zero-day threats (Song et al., 2025).

Network traffic analysis has similarly benefited from AI integration, with machine learning algorithms identifying command-and-control communications, data exfiltration attempts, and lateral movement within networks. These systems establish normal traffic patterns and detect anomalies that might indicate compromise, such as unusual data volumes, connections to suspicious domains, or encrypted traffic to unexpected destinations. Threat intelligence platforms enhanced with AI capabilities can aggregate and analyze information from diverse sources, identifying emerging threats and providing actionable insights to security teams (Abbasi et al., 2021).

2.2. Automated Incident Response

The speed advantage provided by AI-powered automated response systems represents one of the most significant benefits in modern cybersecurity. Traditional incident response processes that require human analysis and decision-making can take hours or days, during which attackers may establish persistence, move laterally through networks, and exfiltrate sensitive data. AI-enabled systems can detect threats and initiate containment measures within milliseconds, dramatically reducing the window of opportunity for attackers (Ofusori et al., 2024).

Security orchestration, automation, and response platforms integrate AI capabilities to coordinate responses across multiple security tools. When a threat is detected, these systems can automatically isolate affected systems, block malicious network connections, terminate suspicious processes, and collect forensic evidence for later analysis. This orchestration eliminates the delays inherent in manual coordination between different security tools, ensuring comprehensive responses that address all aspects of an incident simultaneously (Ismail et al., 2025).

Automated response systems excel at handling high-volume, low-complexity security events that would otherwise overwhelm human analysts. Security operations centers receive thousands of alerts daily, many representing false positives or minor issues that require simple remediation. AI systems can triage these alerts, automatically resolving routine incidents and escalating only those requiring human expertise. This capability allows security teams to focus their limited resources on complex investigations and strategic initiatives rather than repetitive tasks (Tariq et al., 2025).

Adaptive response mechanisms powered by reinforcement learning can optimize response strategies based on effectiveness feedback. These systems learn which containment measures successfully neutralize specific threat types with minimal business disruption, gradually developing sophisticated response playbooks. Predictive response capabilities represent an emerging frontier, with AI systems anticipating likely attack vectors and proactively strengthening defenses before attacks occur (Kinyua and Awuah, 2021).

2.3. Enhanced Security Operations Efficiency

Artificial intelligence dramatically improves the efficiency of security operations by automating time-consuming tasks and augmenting human analyst capabilities. Security teams face chronic talent shortages, with demand for skilled cybersecurity professionals far exceeding supply. AI tools help bridge this gap by handling routine analysis and enabling smaller teams to manage larger, more complex environments effectively (Ofusori et al., 2024).

Alert fatigue represents a significant challenge in security operations, with analysts receiving far more alerts than they can reasonably investigate. AI-powered triage systems can prioritize alerts based on risk severity, business context, and likelihood of representing genuine threats. By filtering out false positives and correlating related alerts into unified incidents, these systems reduce alert volumes by 70 to 90 percent while ensuring that critical threats receive immediate attention (Tariq et al., 2025).

Threat hunting activities benefit substantially from AI augmentation, with machine learning algorithms identifying suspicious patterns that warrant deeper investigation. Rather than manually searching through logs and network data, analysts can leverage AI to surface anomalies and potential indicators of compromise, focusing their expertise on validating findings and conducting detailed investigations. Vulnerability management processes become more strategic with AI-powered prioritization, assessing which vulnerabilities pose the greatest risk based on factors including exploitability, asset criticality, threat intelligence, and business context (Mahboubi et al., 2024).

2.4. Predictive Security Analytics

The ability to anticipate threats before they materialize represents one of AI's most valuable contributions to cybersecurity. Predictive analytics leverage historical data, threat intelligence, and machine learning algorithms to forecast likely attack vectors, identify emerging vulnerabilities, and assess evolving risk landscapes. This forward-looking capability enables organizations to shift from reactive security postures to proactive defense strategies (Achuthan et al., 2024).

Predictive models can identify systems and users most likely to be targeted based on historical attack patterns and current threat intelligence. By analyzing factors such as system configurations, patch levels, user roles, and previous security incidents, these models generate risk scores that guide security investments and monitoring priorities. Organizations can allocate defensive resources more effectively, concentrating protections on high-risk assets rather than applying uniform security measures across all systems (Le et al., 2025).

Threat forecasting capabilities enable security teams to anticipate attack campaigns before they reach their organizations. By analyzing global threat data, including attacks against other organizations in similar industries or regions, AI systems can identify emerging threat patterns and predict likely targets. User behavior analytics with predictive capabilities can identify accounts likely to be compromised or users at high risk of falling victim to social engineering, enabling targeted security controls and training (Almahmoud et al., 2023).

3. Emerging Threats from AI-Enabled Attacks

3.1. Adversarial Machine Learning

The same machine learning capabilities that strengthen defenses also enable sophisticated attacks against AI-powered security systems. Adversarial machine learning involves crafting inputs specifically designed to deceive AI models, causing them to misclassify malicious activities as benign or otherwise fail in their intended functions. These attacks exploit fundamental characteristics of machine learning algorithms, including their reliance on statistical patterns and their inability to truly understand the semantic meaning of data they process (Anthi et al., 2021).

Evasion attacks represent the most common form of adversarial machine learning, where attackers modify malicious inputs to avoid detection by AI security systems. Malware authors can test their creations against commercial security products, iteratively modifying code until it evades detection. This process, once requiring significant expertise, has been automated through AI tools that systematically generate variants until finding versions that bypass security controls (Alotaibi and Rassam, 2023).

Poisoning attacks target the training data used to develop AI security models, introducing malicious examples that cause models to learn incorrect patterns. If attackers can influence training datasets, either by compromising data sources or exploiting systems that incorporate user feedback, they can degrade model performance or create specific blind spots. These attacks prove particularly concerning for security systems that employ continuous learning, automatically updating their models based on new data without rigorous validation (Tian et al., 2022).

Model inversion and extraction attacks enable adversaries to steal proprietary AI models or extract sensitive information from training data. By systematically querying AI systems and analyzing their responses, attackers can reconstruct approximate versions of underlying models or infer characteristics of training data. Backdoor attacks involve embedding hidden triggers in AI models that cause them to behave maliciously when specific conditions are met, remaining dormant until attackers choose to exploit them (Rigaki and Garcia, 2023).

3.2. AI-Powered Social Engineering

Artificial intelligence has dramatically enhanced the effectiveness and scale of social engineering attacks, enabling adversaries to craft highly personalized, convincing deception campaigns. Traditional phishing attacks relied on generic messages sent to large recipient lists, with success depending on small percentages of victims falling for obvious scams. AI-powered social engineering operates fundamentally differently, leveraging data analysis and natural language generation to create targeted, contextually appropriate messages that prove far more convincing (Schmitt and Flechais, 2024).

Large language models enable attackers to generate phishing emails, text messages, and social media content that closely mimics legitimate communications. These AI-generated messages exhibit proper grammar, appropriate tone, and contextually relevant content, eliminating the obvious errors that previously helped recipients identify phishing

attempts. The technology can adapt messaging to match the communication style of specific individuals or organizations, creating highly convincing impersonations (Basit et al., 2020).

Automated reconnaissance powered by AI enables attackers to gather extensive information about targets from public sources, including social media profiles, professional networking sites, corporate websites, and data breaches. Machine learning algorithms can analyze this information to identify optimal attack vectors, such as personal interests, professional relationships, current projects, and psychological vulnerabilities. This intelligence enables highly targeted spear-phishing campaigns tailored to individual recipients, dramatically increasing success rates (Almahmoud et al., 2023).

Voice synthesis technology has reached a point where AI-generated audio can convincingly impersonate specific individuals, enabling sophisticated vishing attacks. Attackers can clone voices from publicly available audio samples, then use synthesized speech to impersonate executives, IT support staff, or trusted contacts. Deepfake technology extends this capability to video, enabling attackers to create convincing fake video content for fraud and disinformation campaigns. Chatbots powered by AI enable attackers to conduct social engineering at scale, engaging with numerous targets simultaneously through natural language conversations (Triantafyllopoulos et al., 2025).

3.3. Automated Vulnerability Discovery and Exploitation

Artificial intelligence has transformed vulnerability research and exploitation, enabling both security researchers and malicious actors to discover and weaponize software flaws more efficiently. Machine learning algorithms can analyze source code, binary executables, and system behaviors to identify potential vulnerabilities, automating processes that previously required extensive manual analysis by expert researchers. This capability accelerates the discovery of security flaws, creating both opportunities for proactive patching and risks from faster weaponization (Shiri Harzevili et al., 2024).

Fuzzing techniques enhanced with AI can generate test inputs more intelligently than traditional random fuzzing approaches. Machine learning algorithms learn which input patterns are most likely to trigger crashes or unexpected behaviors, focusing testing efforts on promising areas rather than exhaustively testing all possibilities. This guided fuzzing discovers vulnerabilities more efficiently, reducing the time required to identify exploitable flaws in complex software systems (Bamohabbat Chafjiri et al., 2024).

Automated exploit generation represents a particularly concerning development, with AI systems capable of developing working exploits for discovered vulnerabilities. While exploit development traditionally required significant expertise and time, AI tools can automate substantial portions of this process, analyze vulnerability characteristics and generate exploit code. This capability dramatically reduces the time between vulnerability disclosure and active exploitation, compressing the window available for organizations to apply patches before attacks begin (Charmanas et al., 2023).

AI-powered attack tools can adapt their exploitation techniques based on target environment characteristics, increasing success rates across diverse systems. Automated attack frameworks incorporating AI capabilities enable less sophisticated adversaries to conduct complex attacks previously requiring expert knowledge. These tools democratize advanced attack techniques, lowering barriers to entry for cybercrime and expanding the threat landscape (Iturbe et al., 2024).

3.4. Algorithmic Bias and Security Blind Spots

The effectiveness of AI security systems depends fundamentally on the quality and representativeness of their training data, creating risks when datasets fail to adequately represent the full spectrum of legitimate activities and potential threats. Algorithmic bias in security contexts can create blind spots where certain attack types go undetected or generate excessive false positives that overwhelm security teams and erode trust in AI systems (Pasipamire and Muroyiwa, 2024).

Training data bias occurs when datasets used to develop AI security models fail to represent the diversity of real-world scenarios. If training data predominantly includes attacks against specific system types, industries, or geographic regions, resulting models may perform poorly when deployed in different contexts. This limitation proves particularly problematic for organizations in underrepresented sectors or regions, where AI security tools trained primarily on data from other contexts may miss relevant threats or generate excessive false alerts (Chawande, 2025).

Adversarial exploitation of bias represents a sophisticated attack vector where adversaries deliberately craft attacks that exploit known limitations in AI security models. By understanding which attack patterns are underrepresented in

training data, attackers can develop techniques specifically designed to evade detection. Concept drift poses ongoing challenges for AI security systems, as the threat landscape evolves continuously while models remain static unless actively updated (Lara-Gutierrez et al., 2025).

Fairness concerns arise when AI security systems treat different user populations inequitably, subjecting certain groups to disproportionate scrutiny or restrictions. Behavioral analytics systems that establish normal baselines may flag legitimate activities as suspicious simply because they differ from majority patterns, potentially discriminating against users with atypical but benign work patterns. Transparency limitations in complex AI models create challenges for security teams attempting to understand why systems make specific decisions, complicating incident investigation and raising accountability questions (Harrath et al., 2025).

4. Balancing Opportunities and Risks

4.1. Implementing AI Security Responsibly

Organizations seeking to leverage AI in cybersecurity must adopt thoughtful implementation approaches that maximize benefits while mitigating associated risks. Responsible AI deployment requires careful consideration of technical, operational, and ethical dimensions, ensuring that automated systems enhance rather than undermine security postures (Kaur et al., 2023).

Starting with clearly defined use cases and success metrics helps organizations focus AI implementations on areas where they provide genuine value. Rather than adopting AI broadly without clear objectives, successful organizations identify specific security challenges where AI capabilities offer meaningful advantages. Clear success metrics enable objective evaluation of whether AI systems deliver expected benefits and justify continued investment (Langer et al., 2024).

Maintaining human oversight of AI security systems proves essential for catching errors, handling edge cases, and ensuring accountability. While automation provides speed and scale advantages, human analysts bring contextual understanding, ethical judgment, and creative problem-solving capabilities that AI systems lack. Effective implementations establish clear boundaries between decisions that can be fully automated and those requiring human review (Frenette, 2023).

Continuous validation and testing of AI security models helps identify degrading performance, bias issues, and vulnerabilities to adversarial attacks. Organizations should implement rigorous testing protocols that evaluate model performance against diverse scenarios, including edge cases and adversarial examples. Regular retraining with current data helps address concept drift, while careful validation of training data helps prevent poisoning attacks (Salem et al., 2024).

Transparency and explainability should be prioritized when selecting and deploying AI security tools. Organizations should favor solutions that provide meaningful explanations for their decisions, facilitating incident investigation and supporting accountability when automated systems make errors. Vendor evaluation processes should assess not only the technical capabilities of AI security products but also the security of the AI systems themselves (Wen et al., 2024).

4.2. Developing AI-Aware Security Strategies

The rise of AI in both offensive and defensive contexts necessitate security strategies that explicitly account for AI-enabled threats while leveraging AI-powered defenses. Organizations must develop comprehensive approaches that address the full spectrum of AI-related security considerations.

Threat modeling should incorporate AI-specific attack vectors, including adversarial machine learning, AI-powered social engineering, and automated exploitation. Security teams should assess which AI-enabled attacks pose the greatest risks to their organizations based on threat actor capabilities, asset values, and existing security controls. This risk assessment guides investment in countermeasures and helps prioritize security initiatives (Paracha et al., 2024).

Defense in depth remains crucial in AI-augmented security environments, with multiple layers of controls providing resilience against both traditional and AI-enabled attacks. Organizations should not rely exclusively on AI security tools but rather integrate them into comprehensive security architectures that include traditional controls, human analysis, and diverse detection mechanisms (Salem et al., 2024).

Security awareness training must evolve to address AI-enabled threats, educating users about deepfakes, AI-generated phishing, and other emerging attack techniques. Updated training should emphasize verification procedures, such as confirming requests through alternative communication channels, rather than relying solely on message characteristics to identify threats. Incident response plans should address scenarios involving AI-enabled attacks and potential failures of AI security systems (Pedersen et al., 2025).

Collaboration and information sharing become increasingly important as AI transforms the threat landscape. Organizations should participate in industry groups, threat intelligence sharing communities, and security research initiatives focused on AI security. Collective defense proves particularly valuable against AI-enabled threats, as shared intelligence about adversarial techniques, model vulnerabilities, and effective countermeasures benefits the entire security community (Mohamed, 2025).

4.3. Building AI Security Expertise

The effective use of AI in cybersecurity requires specialized expertise that combines security knowledge with understanding of machine learning, data science, and AI system vulnerabilities. Organizations must invest in developing these capabilities through hiring, training, and partnerships (Ofusori et al., 2024).

Recruiting professionals with combined security and data science backgrounds helps organizations build teams capable of implementing and managing AI security systems effectively. These hybrid roles require understanding both cybersecurity principles and machine learning techniques, enabling practitioners to deploy AI tools appropriately and recognize their limitations. As demand for these skills exceeds supply, organizations may need to develop talent internally through training programs that upskill existing security professionals in AI technologies (Mohamed, 2023).

Partnerships with academic institutions and research organizations provide access to cutting-edge knowledge about AI security. Universities conducting research on adversarial machine learning and AI security can offer valuable insights and potentially collaborative opportunities. Continuous learning programs ensure that security teams remain current with evolving AI technologies and threats, as the rapid pace of AI development means that knowledge quickly becomes outdated.

Cross-functional collaboration between security teams, data science groups, and AI development teams helps ensure that security considerations are integrated throughout AI system lifecycles. Security professionals should be involved in AI system design, development, and deployment, providing input on threat models, security requirements, and validation approaches (Achuthan et al., 2024).

5. Conclusion

The integration of artificial intelligence into cybersecurity represents a transformative development that fundamentally alters both defensive capabilities and threat landscapes. Organizations now possess tools that can detect threats, respond to incidents, and analyze security data at scales and speeds that would be impossible through human effort alone. These capabilities provide genuine advantages in addressing the volume and sophistication of modern cyber threats, enabling security teams to manage increasingly complex environments with limited resources.

However, the same technologies that strengthen defenses also empower adversaries with sophisticated attack capabilities. AI-enabled threats including adversarial machine learning, automated social engineering, and intelligent exploitation tools create new challenges that traditional security approaches cannot adequately address. The democratization of AI technologies means that these advanced capabilities are available not only to elite adversaries but to a growing population of threat actors with varying skill levels and motivations.

Success in this evolving landscape requires balanced approaches that leverage AI's defensive potential while recognizing and mitigating its risks. Organizations must implement AI security tools thoughtfully, maintaining human oversight, ensuring transparency, and continuously validating system performance. Security strategies must explicitly account for AI-enabled threats, incorporating appropriate countermeasures and updating training programs to address emerging attack techniques. Building organizational expertise that spans both cybersecurity and AI technologies proves essential for effective implementation and management of AI security systems.

The future of cybersecurity will undoubtedly involve increasing AI integration, with both defenders and attackers leveraging ever more sophisticated technologies. Organizations that approach this transformation strategically, combining technological capabilities with human expertise and ethical frameworks, will be best positioned to maintain

security in an increasingly automated threat environment. The key lies not in viewing AI as either a silver bullet solution or an insurmountable threat, but rather as a powerful tool that requires careful implementation, continuous oversight, and integration into comprehensive security strategies that account for both its capabilities and limitations.

As AI technologies continue evolving, the cybersecurity community must remain vigilant, adaptive, and collaborative. Sharing knowledge about AI security challenges, effective countermeasures, and emerging threats benefits all organizations facing common adversaries. Research into AI security, adversarial machine learning, and defensive techniques must continue advancing to stay ahead of threat actors. Regulatory frameworks and industry standards should evolve to address AI-specific security considerations, providing guidance for responsible implementation while avoiding approaches that stifle beneficial innovation. The rise of AI in cyber defense ultimately represents both opportunity and challenge, requiring organizations to navigate complex tradeoffs and make thoughtful decisions about technology adoption, resource allocation, and risk management in an increasingly complex digital environment.

Compliance with ethical standards

Disclosure of conflict of interest

There is no conflict of interest to be disclosed.

Statement of Ethical Approval

This article does not contain any studies with human participants or animals performed by the author.

References

[1] Abbasi, M., Shahraki, A., and Taherkordi, A. (2021). Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A survey. Computer Communications, 170, 19–41. https://doi.org/10.1016/j.comcom.2021.01.021

[2] Achuthan, K., Ramanathan, S., Srinivas, S., and Raman, R. (2024). Advancing Cybersecurity and privacy with Artificial Intelligence: Current Trends and Future Research Directions. Frontiers in Big Data, 7. https://doi.org/10.3389/fdata.2024.1497535

[3] Adawadkar, A. M., and Kulkarni, N. (2022). Cyber-security and Reinforcement Learning — A brief survey. Engineering Applications of Artificial Intelligence, 114, 105116. https://doi.org/10.1016/j.engappai.2022.105116

[4] Almahmoud, Z., Yoo, P. D., Alhussein, O., Farhat, I., and Damiani, E. (2023). A holistic and proactive approach to forecasting cyber threats. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-35198-1

[5] Alotaibi, A., and Rassam, M. A. (2023). Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and Defense. Future Internet, 15(2), 62. https://doi.org/10.3390/fi15020062

[6] Anthi, E., Williams, L., Rhode, M., Burnap, P., and Wedgbury, A. (2021). Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems. Journal of Information Security and Applications, 58, 102717. https://doi.org/10.1016/j.jisa.2020.102717

[7] Bamohabbat Chafjiri, S., Legg, P., Hong, J., and Tsompanas, M.-A. (2024). Vulnerability detection through machine learning-based fuzzing: A systematic review. Computers andamp; Security, 143, 103903. https://doi.org/10.1016/j.cose.2024.103903

[8] Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., and Kifayat, K. (2020). A comprehensive survey of AI-enabled phishing attacks detection techniques. Telecommunication Systems, 76(1), 139–154. https://doi.org/10.1007/s11235-020-00733-2

[9] Charmanas, K., Mittas, N., and Angelis, L. (2023). Exploitation of vulnerabilities: A topic-based machine learning framework for explaining and predicting exploitation. Information, 14(7), 403. https://doi.org/10.3390/info14070403

[10] Chawande, S. (2025). Adversarial Machine Learning and securing AI Systems. World Journal of Advanced Engineering Technology and Sciences, 15(1), 1344–1356. https://doi.org/10.30574/wjaets.2025.15.1.0338

[11] Frenette, J. (2023). Ensuring human oversight in high-performance AI systems: A Framework for control and accountability". World Journal of Advanced Research and Reviews, 20(2), 1507–1516. https://doi.org/10.30574/wjarr.2023.20.2.2194

[12] Harrath, Y., Adohinzin, O., Kaabi, J., and Saathoff, M. (2025). Bridging domains: Advances in explainable, automated, and privacy-preserving AI for Computer Science and cybersecurity. Computers, 14(9), 374. https://doi.org/10.3390/computers14090374

[13] Ismail, Kurnia, R., Brata, Z. A., Nelistiani, G. A., Heo, S., Kim, H., and Kim, H. (2025a). Toward robust security orchestration and automated response in security operations centers with a hyper-automation approach using agentic artificial intelligence. Information, 16(5), 365. https://doi.org/10.3390/info16050365

[14] Ismail, Kurnia, R., Brata, Z. A., Nelistiani, G. A., Heo, S., Kim, H., and Kim, H. (2025b). Toward robust security orchestration and automated response in security operations centers with a hyper-automation approach using agentic artificial intelligence. Information, 16(5), 365. https://doi.org/10.3390/info16050365

[15] Iturbe, E., Llorente-Vazquez, O., Rego, A., Rios, E., and Toledo, N. (2024). Unleashing offensive artificial intelligence: Automated attack technique code generation. Computers andamp; Security, 147, 104077. https://doi.org/10.1016/j.cose.2024.104077

[16] Kaur, R., Gabrijelčič, D., and Klobučar, T. (2023). Artificial Intelligence for cybersecurity: Literature review and future research directions. Information Fusion, 97, 101804. https://doi.org/10.1016/j.inffus.2023.101804

[17] Kinyua, J., and Awuah, L. (2021). AI/ML in security orchestration, automation and response: Future research directions. Intelligent Automation andamp; Soft Computing, 28(2), 527–545. https://doi.org/10.32604/iasc.2021.016240

[18] Langer, M., Baum, K., and Schlicker, N. (2024). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. Minds and Machines, 35(1). https://doi.org/10.1007/s11023-024-09701-0

[19] Lara-Gutierrez, A., Fernandez-Gago, C., and Onieva, J. A. (2025). A framework for drift detection and adaptation in AI-driven anomaly and threat detection systems. International Journal of Information Security, 24(5). https://doi.org/10.1007/s10207-025-01118-9

[20] Le, T. D., Le-Dinh, T., and Uwizeyemungu, S. (2025). Cybersecurity analytics for the enterprise environment: A systematic literature review. Electronics, 14(11), 2252. https://doi.org/10.3390/electronics14112252

[21] Mahboubi, A., Luong, K., Aboutorab, H., Bui, H. T., Jarrad, G., Bahutair, M., Camtepe, S., Pogrebna, G., Ahmed, E., Barry, B., and Gately, H. (2024). Evolving techniques in cyber threat hunting: A systematic review. Journal of Network and Computer Applications, 232, 104004. https://doi.org/10.1016/j.jnca.2024.104004

[22] Mohamed, N. (2023). Current trends in AI and ML for cybersecurity: A state-of-the-art survey. Cogent Engineering, 10(2). https://doi.org/10.1080/23311916.2023.2272358

[23] Mohamed, N. (2025). Artificial Intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 67(8), 6969–7055. https://doi.org/10.1007/s10115-025-02429-y

[24] Ofori, H. K., Bell-Dzide, K., Brown-Acquaye, W. L., Lempogo, F., Frimpong, S. O., Agbehadji, I. E., and Millham, R. C. (2025). Application of machine learning and Deep Learning techniques for enhanced insider threat detection in cybersecurity: Bibliometric Review. Symmetry, 17(10), 1704. https://doi.org/10.3390/sym17101704

[25] Ofusori, L., Bokaba, T., and Mhlongo, S. (2024). Artificial Intelligence in cybersecurity: A comprehensive review and future direction. Applied Artificial Intelligence, 38(1). https://doi.org/10.1080/08839514.2024.2439609

[26] Paracha, A., Arshad, J., Farah, M. B., and Ismail, K. (2024). Machine learning security and privacy: A review of threats and countermeasures. EURASIP Journal on Information Security, 2024(1). https://doi.org/10.1186/s13635-024-00158-3

[27] Pasipamire, N., and Muroyiwa, A. (2024). Navigating algorithm bias in AI: Ensuring fairness and trust in Africa. Frontiers in Research Metrics and Analytics, 9. https://doi.org/10.3389/frma.2024.1486600

[28] Pedersen, K. T., Pepke, L., Stærmose, T., Papaioannou, M., Choudhary, G., and Dragoni, N. (2025). Deepfake-driven social engineering: Threats, detection techniques, and defensive strategies in corporate environments. Journal of Cybersecurity and Privacy, 5(2), 18. https://doi.org/10.3390/jcp5020018

[29] Rigaki, M., and Garcia, S. (2023). A survey of privacy attacks in Machine Learning. ACM Computing Surveys, 56(4), 1–34. https://doi.org/10.1145/3624010

[30] Salem, A. H., Azzam, S. M., Emam, O. E., and Abohany, A. A. (2024). Advancing cybersecurity: A comprehensive review of AI-Driven Detection Techniques. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00957-y

[31] Schmitt, M., and Flechais, I. (2024). Digital deception: Generative artificial intelligence in Social Engineering and phishing. Artificial Intelligence Review, 57(12). https://doi.org/10.1007/s10462-024-10973-2

[32] Shiri Harzevili, N., Boaye Belle, A., Wang, J., Wang, S., Jiang, Z. M., and Nagappan, N. (2024). A systematic literature review on Automated Software Vulnerability Detection using machine learning. ACM Computing Surveys, 57(3), 1–36. https://doi.org/10.1145/3699711

[33] Song, Y., Zhang, D., Wang, J., Wang, Y., Wang, Y., and Ding, P. (2025). Application of deep learning in malware detection: A Review. Journal of Big Data, 12(1). https://doi.org/10.1186/s40537-025-01157-y

[34] Tariq, S., Baruwal Chhetri, M., Nepal, S., and Paris, C. (2025). Alert fatigue in security operations centres: Research challenges and opportunities. ACM Computing Surveys, 57(9), 1–38. https://doi.org/10.1145/3723158

[35] Tian, Z., Cui, L., Liang, J., and Yu, S. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. ACM Computing Surveys, 55(8), 1–35. https://doi.org/10.1145/3551636

[36] Triantafyllopoulos, A., Spiesberger, A. A., Tsangko, I., Jing, X., Distler, V., Dietz, F., Alt, F., and Schuller, B. W. (2025). Vishing: Detecting Social Engineering in spoken communication — a first survey and urgent roadmap to address an emerging societal challenge. Computer Speech andamp; Language, 94, 101802. https://doi.org/10.1016/j.csl.2025.101802

[37] Wen, S.-F., Shukla, A., and Katt, B. (2024). Artificial Intelligence for System Security Assurance: A systematic literature review. International Journal of Information Security, 24(1). https://doi.org/10.1007/s10207-024-00959-0

[38] Ye, X., Luo, F., Cui, H., Wang, J., Xiong, X., Zhang, W., Yu, J., and Zhao, W. (2025). Research on insider threat detection based on personalized federated learning and behavior log analysis. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-04029-w