

AI-Powered Incident Response: Automating Threat Detection and Mitigation

Kim Min Joon

Pohang University of Science and Technology (POSTECH), Pohang, South Korea

Corresponding E-mail: kim.joon@interviewiauniversity.com

Abstract

The escalating frequency and sophistication of cyberattacks have exposed the limitations of traditional, human-driven incident response mechanisms, which are often too slow, resource-intensive, and error-prone to counter modern threats effectively. This paper explores the transformative role of Artificial Intelligence (AI) in automating threat detection and mitigation within cybersecurity incident response workflows. It examines how machine learning algorithms, behavioral analytics, natural language processing, and autonomous response systems are being deployed to identify anomalies, classify threats, prioritize alerts, and execute containment measures in real time. The paper further discusses the architectural components of AI-driven incident response platforms, the benefits they offer in terms of speed, scalability, and accuracy, and the challenges associated with their adoption, including false positives, adversarial manipulation, data privacy concerns, and the need for human oversight. By synthesizing current developments and emerging trends, this paper argues that AI-powered incident response represents not merely an incremental improvement but a fundamental paradigm shift in how organizations defend against cyber threats. It concludes by outlining future directions, including explainable AI, federated learning, and human-AI collaboration, that will shape the next generation of intelligent security operations.

Keywords: Artificial Intelligence, Incident Response, Threat Detection, Automation, Machine Learning, Cybersecurity, Anomaly Detection, Mitigation, Security Operations Center, Behavioral Analytics

I. Introduction

The digital transformation of modern enterprises has dramatically expanded the attack surface available to malicious actors[1]. Organizations now operate across cloud infrastructures, distributed endpoints, Internet of Things (IoT) devices, and hybrid work environments, generating enormous volumes of security telemetry that traditional incident response teams struggle to process. Conventional incident response frameworks typically follow a linear sequence of preparation, identification, containment, eradication, recovery, and lessons learned, but this model assumes that human analysts can detect, triage, and respond to threats within acceptable timeframes. In reality, the average dwell time for an intruder—the period between initial compromise and detection—can extend for weeks or even months, during which attackers escalate privileges, exfiltrate data, and establish persistent backdoors. The sheer speed at which automated malware and ransomware propagate further compounds this problem, rendering manual response inadequate for containing fast-moving threats. Artificial Intelligence has

emerged as a critical enabler for closing this detection and response gap, offering the ability to analyze vast datasets, recognize subtle patterns, and initiate corrective actions within milliseconds. AI-powered incident response systems leverage supervised and unsupervised learning, deep neural networks, reinforcement learning, and natural language processing to augment or replace repetitive human tasks across the incident lifecycle. This paper examines the architecture, capabilities, benefits, and limitations of AI-driven incident response, with particular emphasis on automation of threat detection and mitigation. It argues that while AI cannot entirely replace human judgment, it fundamentally redefines the operating model of the Security Operations Center (SOC), shifting analysts from manual alert triage toward strategic oversight and threat hunting.

II. Foundations of AI in Threat Detection

Effective threat detection rests on the ability to distinguish malicious activity from benign behavior across heterogeneous data sources, including network flows, endpoint logs, authentication records, and application telemetry[2]. Traditional signature-based detection, while reliable for known threats, fails against zero-day exploits, polymorphic malware, and living-off-the-land techniques that abuse legitimate system tools. AI addresses this limitation through anomaly detection, in which models learn a baseline of normal behavior and flag statistically significant deviations. Unsupervised techniques such as clustering, isolation forests, and autoencoders are particularly valuable because they do not require labeled attack data, which is scarce and constantly outdated. Supervised learning, by contrast, excels when historical incident data is available, enabling classifiers to distinguish between true threats and false alarms with high precision. Deep learning architectures, including recurrent neural networks and transformers, capture temporal dependencies in sequential data, making them well suited to detecting multi-stage intrusions that unfold over time. Behavioral analytics extends these capabilities to user and entity behavior analytics (UEBA), profiling individual accounts and devices to identify credential misuse, lateral movement, or insider threats. Natural language processing further enriches detection by parsing threat intelligence reports, dark web chatter, and vulnerability disclosures to extract actionable indicators[3]. The integration of these techniques into a unified detection pipeline allows organizations to correlate weak signals across domains, reducing the likelihood that a stealthy intrusion escapes notice. Crucially, AI-based detection systems improve continuously as they ingest new data, adapting to evolving attacker tactics without requiring constant manual rule updates. This adaptive quality is essential in a threat landscape where adversaries deliberately modify their techniques to evade static defenses.

III. Automating Incident Triage and Prioritization

One of the most persistent challenges in security operations is alert fatigue: the overwhelming volume of low-fidelity alerts generated by disparate tools causes analysts to miss genuinely critical incidents[4]. AI-powered triage addresses this problem by aggregating, deduplicating, and scoring alerts according to contextual risk factors such as asset criticality, user privilege level, threat intelligence matches, and historical outcomes. Machine learning models trained on past incident resolutions can predict which alerts are most likely to escalate into material breaches, allowing teams to focus attention where it matters most. Clustering algorithms group related alerts into coherent incidents, revealing campaigns that might otherwise appear as isolated events. Risk-scoring engines combine quantitative indicators with probabilistic reasoning to assign severity ratings that reflect both technical impact and business context.

Automation also extends to enrichment: AI agents can automatically query threat intelligence platforms, retrieve asset inventories, and pull endpoint forensic data, assembling a complete incident picture before a human analyst even opens the ticket[5]. This reduces mean time to acknowledge (MTTA) and mean time to respond (MTTR), two metrics that directly correlate with breach severity. Additionally, conversational AI assistants and natural language interfaces allow analysts to query incident data, request summaries, and receive recommended remediation steps without navigating multiple consoles. By offloading routine triage to intelligent systems, organizations free their skilled personnel to conduct proactive threat hunting and strategic security improvements, effectively multiplying the value of limited human expertise.

IV. Autonomous Mitigation and Response Actions

Detection and triage are only valuable if they translate into timely containment and remediation. AI-powered mitigation systems close the loop by executing response actions automatically when confidence thresholds are met. These actions may include isolating compromised endpoints from the network, disabling user accounts, blocking malicious IP addresses or domains at the firewall, terminating suspicious processes, and revoking access tokens. Playbook-driven automation, often implemented through Security Orchestration, Automation, and Response (SOAR) platforms, encodes expert-defined workflows that AI can trigger based on incident classification. Reinforcement learning offers a more adaptive approach, allowing response agents to learn optimal containment strategies through simulated attack environments, balancing the trade-off between swift action and operational disruption. For example, an AI system might quarantine a single workstation showing ransomware behavior while simultaneously alerting administrators, thereby preventing lateral spread without halting critical business functions. AI can also coordinate multi-step responses across cloud, on-premises, and endpoint environments, ensuring consistent policy enforcement[6]. However, autonomous mitigation raises important governance questions: How much authority should machines have to take destructive actions? What safeguards prevent an AI from isolating a production server erroneously? Leading implementations therefore adopt a tiered autonomy model, in which low-risk actions are fully automated, medium-risk actions require rapid human confirmation, and high-impact actions demand explicit approval. This graduated approach preserves the speed advantages of automation while retaining human accountability for consequential decisions.

V. Challenges, Risks, and Limitations

Despite its promise, AI-powered incident response is not a panacea. False positives remain a significant concern; an overzealous model may quarantine legitimate users or services, causing business disruption and eroding trust in automation. Adversarial machine learning presents a more insidious threat, as attackers can craft inputs specifically designed to deceive detection models, such as slightly perturbed malware samples that evade classification. Data quality and availability also constrain performance: models trained on incomplete, biased, or outdated datasets may generalize poorly to real-world conditions. Privacy and regulatory compliance complicate the collection and retention of telemetry, particularly in jurisdictions with strict data protection laws. Explainability is another critical gap; deep learning models often function as black boxes, making it difficult for analysts to justify automated decisions during audits or post-incident reviews. Integration challenges arise when AI tools must interoperate with legacy systems, proprietary formats, and fragmented security stacks[7]. Finally, the shortage of skilled

AI security professionals means many organizations lack the expertise to tune, validate, and maintain these systems effectively. Addressing these limitations requires robust model validation, continuous red-teaming, transparent documentation, and clear escalation paths that keep humans informed and in control.

VI. Future Directions and Human-AI Collaboration

The future of AI-powered incident response lies not in full autonomy but in sophisticated human-AI collaboration[8]. Explainable AI techniques, such as SHAP and attention visualization, will help analysts understand why a model flagged an event, building trust and enabling faster verification. Federated learning allows organizations to train shared models without exposing sensitive raw data, addressing privacy concerns while improving collective detection capabilities. Generative AI is beginning to assist in drafting incident reports, summarizing forensic findings, and simulating attacker behavior for training purposes[9]. Agentic AI systems, capable of planning and executing multi-step investigations, may soon function as virtual tier-one analysts working alongside human teams. As these technologies mature, governance frameworks must evolve in parallel, defining accountability, auditability, and ethical boundaries for automated response[10]. Ultimately, the most resilient security operations will combine machine speed with human judgment, leveraging AI to handle scale and repetition while reserving strategic, ethical, and novel decisions for people.

VII. Conclusion

AI-powered incident response represents a decisive shift from reactive, manual security operations toward proactive, intelligent defense. By automating threat detection through anomaly and behavioral analytics, streamlining triage and prioritization, and enabling autonomous or semi-autonomous mitigation, AI dramatically reduces the time attackers have to operate and the damage they can inflict. Yet automation introduces new risks—false positives, adversarial evasion, opacity, and governance gaps—that demand careful management. The most effective implementations will not pursue full autonomy but will instead cultivate a collaborative model in which AI amplifies human expertise, and humans retain ultimate accountability for high-stakes decisions. As cyber threats continue to grow in scale and sophistication, organizations that successfully integrate AI into their incident response workflows will gain a decisive advantage, while those that rely solely on manual processes will find themselves increasingly outmatched. The path forward requires investment not only in algorithms and platforms but also in people, processes, and policies that ensure AI serves as a trusted partner in defending the digital enterprise.

References:

- [1] G. Yang, J. Xu, W. Chen, Z.-H. Qi, and H.-Y. Wang, "Security characteristic and technology in the internet of things," *Nanjing Youdian Daxue Xuebao(Ziran Kexue Ban)/ Journal of Nanjing University of Posts and Telecommunications(Natural Nanjing University of Posts and Telecommunications(Natural*, vol. 30, no. 4, 2010.
- [2] C. T. Udeze, O. Falebita, and S. T. Fasasi, "AI-Based Cybersecurity Risk Prediction and Management," *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, vol. 16, no. 04, pp. 216-225, 2024.
- [3] R. H. Weber, "Internet of Things–New security and privacy challenges," *Computer law & security review*, vol. 26, no. 1, pp. 23-30, 2010.
- [4] S. Gupta, K. Vanteru, P. Gadiraju, and S. Reddy, "Edge AI Consensus: Blockchain-Secured Federated Learning for Trustless Intelligence Network," in *Proceedings of the 2025 2nd international Conference on Edge Computing, Parallel and Distributed Computing*, 2025, pp. 70-76.
- [5] C. T. Udeze, O. Falebita, and S. T. Fasasi, "How Artificial Intelligence Is Transforming Cybersecurity Incident Response," *Journal of Contemporary Science and Technology Management*, vol. 2, no. 03, pp. 95-116, 2026.
- [6] S. Gupta, K. Vanteru, S. Reddy, and B. Madupati, "AI-Enhanced Blockchain Networks for Climate Change Monitoring and Carbon Credit Verification," in *Proceedings of the 2025 4th International Conference on Frontiers of Artificial Intelligence and Machine Learning*, 2025, pp. 31-37.
- [7] A. E. Hassanien, T.-H. Kim, J. Kacprzyk, and A. I. Awad, "Bio-inspiring cyber security and cloud services: Trends and innovations," 2014.
- [8] C. T. Udeze, O. Falebita, and S. T. Fasasi, "Artificial Intelligence for Automated Cybersecurity Risk Assessment," *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, vol. 17, no. 04, pp. 21-32, 2025.
- [9] A. Konn, "The Convergence of Technology and Cybersecurity: Safeguarding Devices in a Cloud-Driven World," 2018.
- [10] Y. Maleh, "IT/OT convergence and cyber security," *Computer Fraud & Security*, vol. 2021, no. 12, pp. 13-16, 2021.