Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Agentic AI for Autonomous Cyber Defense

Bharatveeranjaneya Reddy Devagiri

Osmania University, India

ARTICLE INFO

Received: 06 Feb 2026

Revised: 08 Feb 2026

ABSTRACT

In the face of increasingly diverse, large-scale and fast-moving threats, customary and reactive approaches that are heavily dependent on humans for monitoring and policy-based automation are no longer sufficient. Attackers are using automation, artificial intelligence and highly dynamic cloud-native infrastructure. As a result, security operations struggle to keep visibility current and to respond effectively. The article analyzes the future of autonomous cyber defense using an agentic AI, which can observe the environment, reason under uncertainty, plan and execute defense actions, and learn about attack diversity. Current AI-enabled cyber defense solutions largely exist in the domain of playbook-based automation and predefined static rule-based detections. Agentic AI, by virtue of its higher level of autonomy, can hypothesize malicious intent, simulate attack strategies, and implement risk-aligned defensive measures in near real-time. The paper presents an architectural model integrating the layers of perception, reasoning, planning, action, and learning for proactive and contextualized defense of distributed and heterogeneous environments. Furthermore, it discusses the role of advanced reasoning methods in balancing decision accuracy with explainability and the role of automated planning and action layers in translating high-level security goals to operational controls while preserving accountability with governance. The article also identifies the risks of over-automation, unexpected side effects, inaccuracies in models, adversarial behavior, and a lack of trust, transparency, and interpretability. The article argues that strong governance frameworks, including prioritizing human control, compliance with legal requirements and regulatory guidelines, continuous monitoring, and ethical design are essential for the safe, transparent, and accountable deployment of autonomous systems. This article argues for human-and-AI security models, which largely preserve the human responsibility and trust of existing models while enabling scalable and adaptable cyber-defenses, by framing agentic AI as a cyber-force multiplier rather than a cyber-replacement. Such models may be supported as a cyber-foundational capability if responsibly designed and governed and continuously aligned with organizational and social goals and values.

Keywords: Agentic AI in Cybersecurity, Autonomous Cyber Defense Systems, Artificial Intelligence for Cybersecurity, AI-Driven Threat Detection and Response, Human-AI Governance in Cyber Defense

1. Introduction

1.1 Defining Agentic AI in Cybersecurity Context

Agentic AI is the next evolution of cyber defense using AI. Whereas legacy and current approaches use playbooks and heuristics to respond to attack signatures, agentic AI can reason, plan, and automatically

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

respond to an ever-changing threat landscape [1]. Unlike reactive AI-based defense that seeks to automate detection rules and trigger scripted responses to them, agentic AI is an agent that continuously monitors its environment, simulates attack paths, and adjusts the cybersecurity posture of the system in near real-time. The precise combination of these AI workstreams will depend on the specific evolutionary transition from narrow AI to agentic AI. Machine learning can identify anomalies and discover patterns in the security telemetry data from systems, networks and devices; large language models can understand threat intelligence reports and reason about the context of security incidents; and reinforcement learning frameworks can be used to derive optimal defensive strategies through simulations and real-world attacks. On the other hand, decision-making architectures enable security tools to realize these workstreams by cleverly combining them to balance security goals with operational constraints [2]. In other words, the defining characteristic of agentic AI is not the technical capabilities of its workstreams, but its ability to exhibit emergent autonomous agency by independently hypothesizing adversarial intent, considering alternative defensive strategies aligned with the organization's appetite for risk, and autonomously engaging in coordinated defensive actions that go beyond its explicit incident response plans.

1.2 The Imperative for Autonomous Cyber Defense

However, contemporary threats are advanced and rapid, and humans cannot keep up with the volume of alerts. There is a pressing operational need for defensive actions to be taken entirely automatically, at machine speed, analyzing, deciding, and acting [1]. Automated discovery, exploit, and lateral movement techniques allow adversaries to breach enterprise networks in minutes, while customary security operations centers (SOCs) require hours or days to detect, investigate, and respond to complex attack chains. This temporal asymmetry renders customary human-in-the-loop security operations models ill-suited for protecting highly dynamic cloud and hybrid infrastructures composed of distributed microservices, ephemeral container-like workloads, and multi-tenant cloud environments. The attack surface widens greatly in cloud-native service-oriented architectures, and dynamic service discovery and continuous deployment challenge perimeter-based security architectures. Thorough AI-augmented security platforms have been demonstrated to detect 95% of attacks in threat scenarios with less than 2s latency for 10,000 events per second [2]. The volume of security telemetry from networks, endpoints, identities and applications has outstripped the ability of human analysts to process it, creating an information processing bottleneck, as well as analyst fatigue due to undifferentiated false positive alerts. Autonomous cyber defense overcomes this by moving from alert-based models, where security teams must react after an alert threshold is exceeded, to systems that self-protect by continuously monitoring their state, predicting paths of compromise, assessing organizational risk, and consistently responding to events in real time by modifying the operating environment.

2. Architectural Framework of Agentic AI Cyber Defense Systems

2.1 Perception Layer: Data Ingestion and Situational Awareness

The perception layer collects telemetry data from the network, endpoints, identity, cloud services, application programming interfaces (APIs), and external threat intelligence feeds. The architecture must overcome the challenge of real-time temporal alignment, semantic normalization, and correlation of the heterogeneous telemetry data to maintain a common operating picture of the situation. This layer builds models of multiple baselines (user, asset, workload) to detect deviations from normal behavior. This enables context enrichment (e.g., asset criticality, trust relationships) that can detect low-signal, multi-stage attacks by linking multiple low-fidelity signals into a high-fidelity attack indicator. Eventually, a

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

greater sensitivity to risky behaviors is achieved, albeit without a commensurate increase in the alert volume [3].

2.2 Reasoning and Decision-Making Layer

The reasoning layer conducts a security analysis of the perceived signals using probabilistic inference, learning-based cognition and symbolic reasoning. As an alternative to deterministic rules, the architecture uses the context of the threat to assess its likelihood, associated damage, and possibly the intent of the attack amid uncertainty. This allows for misconfiguration vs active exploitation and forward-looking analyses where opponent behaviors can be simulated. There are trade-offs between explainability vs. decision optimality, e.g., a richer model may have better accuracy and generalization but may not be easily understandable by human analysts who need to validate the decisions or post-mortem the incident [4].

2.3 Planning and Action Layer

The planning and action layer implements security decisions taken at the calculated level. It transforms abstract security goals into concrete security rules at the controls level in the infrastructure. Responses can be either low-impact security policy changes or containment actions with higher disruption rates. Architectural governance mechanisms allow certain authorized actions to execute automatically while requiring human approval for higher-impact responses. This balances responsiveness with accountability and allows for organizational diversity in tolerating differences in levels of compliance and different operational capabilities [3].

2.4 Learning and Adaptation Layer

The learning and adaptation layer closes the loop by incorporating feedback from the results of each incident, the evolution of the environment, and analyst action. It continuously improves detection models and decision policies to minimize false positives, adapt to concept drift, and respond to new attacks. To avoid destabilizing controls in production, the architecture isolates learning and experimentation from production enforcement. It enables timely incorporation of learning feedback. This constant learning helps the defense system to adjust to changes in threats and infrastructure and makes resilience a dynamic rather than a static design property [2].

Layer	Functions	Methods	Outcomes
Perception	Data ingestion and situational awareness	Temporal alignment, semantic normalization, multi-baseline modeling	High-fidelity attack indicators, increased sensitivity without alert volume growth
Reasoning & Decision-Making	Security analysis of signals	Probabilistic inference, learning-based cognition, symbolic reasoning	Threat assessment amid uncertainty, misconfiguration vs. exploitation distinction
Planning & Action	Implement security decisions	Automated/human-approved responses, governance mechanisms	Low-impact policy changes, high-disruption containment actions
Learning & Adaptation	Incorporate feedback and improve	Continuous model improvement, isolated experimentation	Minimize false positives, adapt to concept drift, dynamic resilience

Table 1: Architectural Framework of Agentic AI Cyber Defense Systems [3, 4]

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. Strategic Advantages and Operational Capabilities

3.1 Proactive Threat Prediction and Preemptive Mitigation

The ability to extrapolate threats and preemptively take defensive action may also be related to information security incident lifecycle and preparedness concepts in the NIST SP 800-61 Rev. 2 incident handling guide (continuous monitoring, situational awareness, and readiness in the context of effective incident handling) [5]. In addition, incident response is formalized by NIST SP 800-61 Rev. 2 in four phases (preparation; detection and analysis; containment, eradication, and recovery; and post-incident activity), which anticipatory defensive actions fit into before confirmed incident response [5]. Other independent works related to autonomous threat hunting have indicated that data-driven approaches can analyze telemetry and threat intelligence to infer attack paths and opponent behavior even before a compromise can be observed [6]. In this context, e.g. hypothesis-based analysis and behavior-based inference may detect incidents better and faster than signature-based approaches, as pre-incident detection is a primary focus of established incident response frameworks [5][6].

3.2 Machine-Speed Response and Temporal Advantage

Machine-speed response is critical because it operates within the framework of human-operated incident response. The slower the response cycle, the greater the loss of business and operational value, as in NIST SP 800-61 Rev. 2. An analysis and containment component is also essential to avoid increased impact once the indicators of compromise are identified [5]. The guide recognizes that automation can be a key enabler to accelerate response actions across complex, distributed systems [5]. Autonomous threat hunting frameworks extend this capability to all high-volume security data streams, using machines to continuously ingest, correlate, and reason over this data without needing human involvement. Doing so at machine speed allows real-time implementation of detection and containment logic and reduces the need for human triage of time-sensitive incidents. This may be valuable for fulfilling response time objectives stated in formal incident handling guidelines [5][6].

3.3 Scalability and Resource Optimization

A number of architectural benefits include scalability and resource efficiency in the incident response and threat hunting process. The growing complexity of the infrastructure, and a shortage of manpower, could hamper an organization's ability to implement and maintain an effective incident response process, according to NIST SP 800-61 Rev. 2, which recommends centralized coordination and automation [5]. More recently, autonomous threat hunting is a growing area of research, where artificial intelligence algorithms are introduced into threat hunting architecture to suppress low-confidence alerts and elevate higher-confidence threat hypotheses. This alleviates cognitive overload on human threat analyzers, who can then focus on oversight, investigation, and strategy activities. This could help security operations to scale without proportionally scaling the number of human operators, ameliorating one of the shortfalls of existing incident response frameworks [5] [6].

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Capability	Key Points
Proactive Threat Prediction & Preemptive Mitigation	Relates to NIST SP 800-61 Rev. 2 incident handling guide concepts. Data-driven approaches analyze telemetry and threat intelligence to infer attack paths and adversary behavior before compromise. Hypothesis-based and behavior-based inferences detect incidents better and faster than signature-based approaches.
Machine-Speed Response & Temporal Advantage	Operates within a human-operated incident response framework. Automation accelerates response actions across complex, distributed systems. Machines continuously ingest, correlate, and reason over high-volume security data without human involvement. Reduces human triage of time-sensitive incidents.
Scalability & Resource Optimization	Growing infrastructure complexity and manpower shortages hamper effective incident response. AI algorithms suppress low-confidence alerts and elevate higher-confidence threat hypotheses. Alleviates cognitive overload on human analysts who focus on oversight, investigation, and strategy. Security operations scale without proportionally scaling human operators.

Table 2: Strategic Advantages and Operational Capabilities [5, 6]

4. Critical Risks and Technical Challenges

4.1 Over-Automation and Unintended Operational Consequences

The main operational risk when automating defensive actions is the unclear definition of the authority and responsibility across the organizational and technical structures and the lack of human intervention in the incident response landscapes. These caveats on full automation are stated in NIST SP 800-61 Rev. 2. Partial automation of incident detection and response activities can be a further development. This cannot be achieved and is not desirable from a security, trust, and contextual perspective [7]. The standard notes that, while it may not be possible to reliably automate sharing or action in situations requiring context-specific judgment, controlled autonomy with escalation can allow for higher-level human intervention. Poorly governed autonomy can disrupt valid users and services of other parts of complex, interconnected infrastructures and create unintended impacts on dependent processes and systems. These risks support the need for clearly defined boundaries of autonomy and human decision-making points, which the standard supports in its recommendations for coordinated incident management and incident response in place of full autonomy.

4.2 Model Reliability: Errors, Hallucinations, and Contextual Misinterpretation

Model reliability in adversarial and noisy environments is one of the main technical challenges for automated systems, causing system and organizational reliability. An empirical characterization of the robustness of a deep learning classifier against cybersecurity attacks was recently published in Algorithms. It reported a drop in the model accuracy from 97.36% to 61.40% for the fast gradient sign method (FGSM) and to 62.28% for Gaussian noise evasion [8]. The corresponding PAVI confidence intervals are 38.60 % and 37.72 %, respectively, indicating a strong sensitivity to adversarial perturbations, especially in cases where the context is misinterpreted or the input is malformed. Errors in

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

these outputs can lead to system state divergence or inappropriate responses by autonomous defensive actions, which calls for validation, confidence thresholding, and failsafe controls.

4.3 Adversarial Threats Against AI Systems

AI systems have existing vulnerabilities in the training and inference phases. During these attacks, adversarial attacks can pose a risk to autonomous security components. According to MDPI, the kind of attack on AI systems can vary in terms of degradation in performance, but accuracy degradation is higher in adversarial perturbation techniques compared to other manipulation techniques [8]. This also illustrates that the right inputs can steer detection and decision-making without touching the inner workings of the system. Related issues include the risk of poisoning training data, manipulating inputs, and exploiting logic in agentic AI systems. These risks can be reduced using continual evaluation of AI subsystems and multi-layered defense lines, thus ensuring that operations remain reliable.

4.4 Explainability, Auditability, and Trust Deficits

Explainability and auditability are important to trust automated cybersecurity solutions, especially in highly regulated, sensitive, or safety-critical environments. NIST SP 800-61 Rev. 2 requires documentation, traceability, and repeatability of the incident response process to enable post-incident review, evidence preservation, and compliance review [7]. Alternatively, decision-making systems with no interpretability or explainability may breach these objectives, as defensive actions taken by machines would be beyond scrutiny, as the reasoning behind them could not be interpreted. Incident investigators or auditors would also be unable to assess whether actions adhered to regulations and organizational policies. These limitations result in a trust deficit, hampering the adoption of autonomous security capabilities even when they are effective.

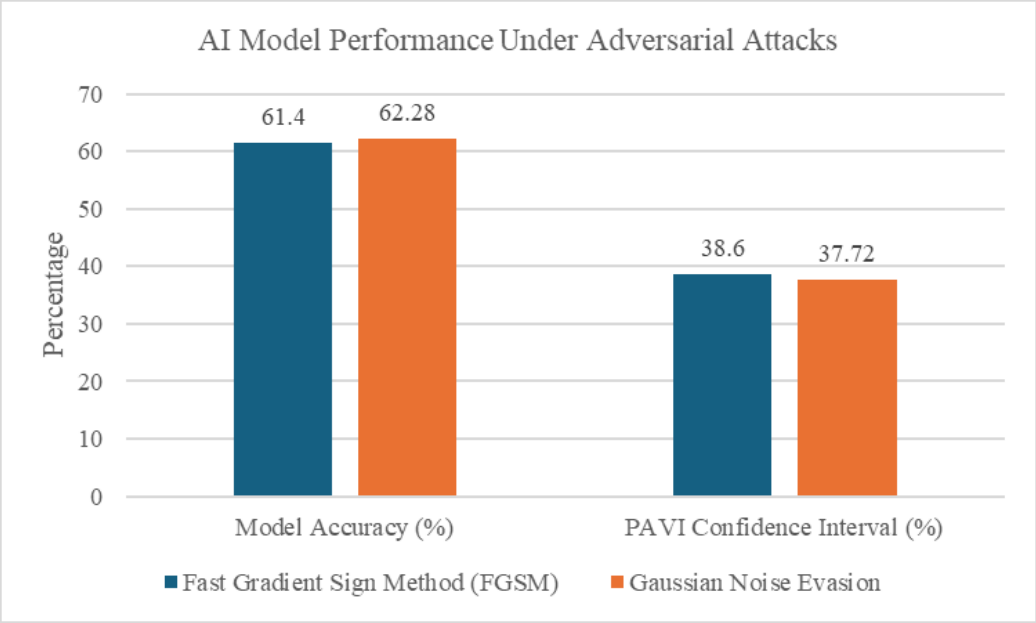

Fig 1: AI Model Performance Under Adversarial Attacks [7, 8]

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. Governance Frameworks and Ethical Implementation

5.1 Human-in-the-Loop Controls and Oversight Mechanisms

Human-in-the-loop HITL controls are integral to models of governance that strike a balance between the benefits of automation and ex post accountability and those that apply to the increasing autonomy of automation. A tiered framework for human-AI teaming distinguishes five levels of AI autonomy and three levels of trust and operator engagement, relative to task complexity and operational risk. This allows governance to scale alongside an increasing level of AI capability while maintaining appropriate human oversight where necessary. [10] In a survey of 194 cybersecurity practitioners, only 13 % of respondents reported that they trust automatically generated AI recommendations without human oversight. This study found that, for instance, 92 % of respondents believed that automation reduces MTTD/MTTR. Hence, even though the operational benefit of automation is obvious, human oversight fulfills the requirement of situational awareness, responsibility, and accountability for high levels of autonomy and should be preserved. In an autonomy stratification, the use of escalation protocols and human decision checkpoints could allow governance frameworks to manage and control cyber defense efficiently.

5.2 Regulatory Compliance and Policy Alignment

Regulatory compliance and policy alignment can ensure that autonomous cybersecurity systems are ethical by restricting their decisions to legally and ethically acceptable bounds. However, survey results by practitioners show that 41 % of respondents believe there is a gap in governance, as regulations are too slow to react to the new requirements posed by AI-augmented incident response [9]. A human-AI collaborative level of autonomy taxonomy for compliance depicts levels of autonomy with corresponding levels of trust and oversight. It enables the injection of regulatory and enterprise risk management constructs into autonomous decision-making authority [10]. Structuring compliance checkpoints around levels of autonomy and human oversight allows automated systems to only operate within policy-compliant governance architectures, thereby limiting the potential for legal or ethical violations in practice.

5.3 Continuous Monitoring, Auditing, and Accountability Infrastructure

Monitoring and auditing systems represent an empirical basis for accountability for autonomous systems. This is possible through the assessment of performance and effects at different levels of automation. Survey results show a wide endorsement from practitioners for modernizing governance systems with an audit-friendly mechanism that stores decision outcomes and trust levels, which is useful for a systematic logging and evaluation strategy of security operations using AI [9]. The layered model of autonomy can provide a framework for accountability by explicitly mapping human-analyst and AI-agent roles during the task to decision rights, duties and boundaries within a SOC function [10]. Data for MTTD/MTTR reduction and trust readouts for each layer of autonomy can be leveraged for governance realignment and post-incident reviews.

5.4 Responsible AI Principles and Ethical Design

Governance frameworks for autonomous cybersecurity systems require establishing accountability, transparency, fairness, and ethics as core principles. The distribution of responsibilities between humans and AI, as well as the autonomy level, trust, and oversight, should be considered to ensure explainability through feedback and limited decision-making autonomy [10]. While practitioners are motivated to automate operations, they do not fully trust unsupervised AI decisions. This motivates stronger ethical principles around bias mitigation, explainable decisions, and strong human control in automated governance processes [9]. The extent to which decision authority and scope of control are relinquished at

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

different levels of autonomy could provide a useful operational indicator of the ethics of the existing governance processes for enabling responsible AI system development.

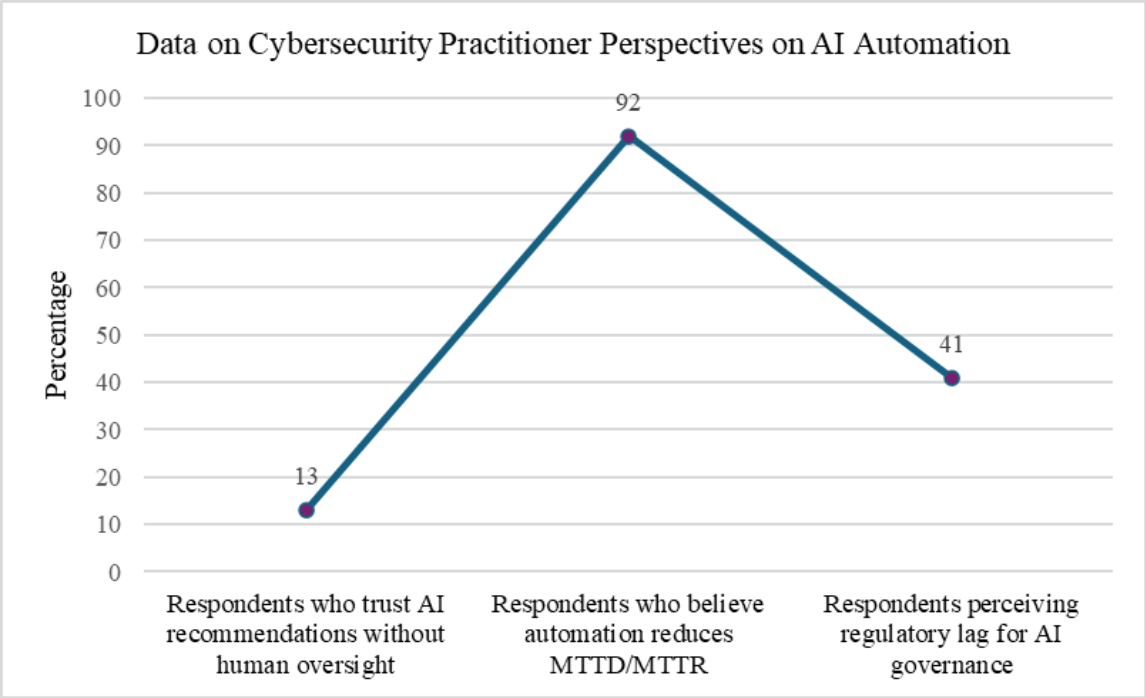

Fig 2: Data on Cybersecurity Practitioner Perspectives on AI Automation [9, 10]

Conclusion

Agentic AI is the next generation of cyber defense systems integrating perception, reasoning, planning, and learning (PRPL). With PRPL capabilities, these systems operate at machine speed to address complex and dynamic threats, accelerating the detection and response process, enabling greater scalability against expanding attack surfaces, and achieving greater resiliency by responding to new and emerging adversarial behaviors. However, the reviews also highlight the challenges for the governance of these benefits, given practitioner distrust around full autonomy for decision-making despite the strong endorsement for the operational benefits of clever automation. This provides a counterpoint for the importance of accountability, transparency, and responsible stewardship in decision-making beyond operational value for gaining general acceptance for smart automation technologies. Structured human-AI collaboration models can formalize the levels of autonomy, the trust margins, and the human intervention points at which decisions are applied. These models can be integrated as part of the joint development of autonomous capabilities to ensure compliance with laws, regulations, ethics, and organizational policies and principles while also providing the benefits of human judgment for higher-stakes security decisions. Continuing monitoring and auditability further tie together these concerns, with each of these methods enabling, respectively, observable and verifiable outputs, which can enable adaptive governance and post-hoc analysis, and further establishing agentic AI as a force multiplier for human expertise, rather than a replacement, within an appropriate boundary of control. As governance

systems evolve around these capabilities, agentic AI will ultimately form a trusted digital guardian able to conduct cybersecurity operations that are proactive, scalable, and resilient while maintaining organizational control, ethicality and stakeholder trust in more autonomous defense environments.

References

[1] Giovanni Apruzzese et al., "The Role of Machine Learning in Cybersecurity," ACM Digital Marketing, 2023. Available: https://dl.acm.org/doi/10.1145/3545574

[2] Akshay Mittal, "AI-Augmented DevSecOps Pipelines for Secure and Scalable Service-Oriented Architectures in Cloud-Native Systems," ResearchGate, 2025. Available: https://www.researchgate.net/publication/394980888_AI-Augmented_DevSecOps_Pipelines_for_Secure_and_Scalable_Service-Oriented_Architectures_in_Cloud-Native_Systems

[3] Kelley Dempsey et al., "Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations," NIST Special Publication 800-137, September 2011. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-137.pdf

[4] Anna L. Buczak et al., "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Xplore, 2015. Available: https://ieeexplore.ieee.org/document/7307098

[5] Alex Nelson et al., "Incident Response Recommendations and Considerations for Cybersecurity Risk Management," National Institute of Standards and Technology (NIST), 2005. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf

[6] Siva Raja Sindiramutty, "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence," arXiv, December 2023. Available: https://arxiv.org/pdf/2401.00286.pdf

[7] Paul Cichonski et al., "Computer Security Incident Handling Guide," National Institute of Standards and Technology (NIST), 2012. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf

[8] Sarfraz Brohi and Qurat-ul-ain Mastoi, "AI Under Attack: Metric-Driven Analysis of Cybersecurity Threats in Deep Learning Models for Healthcare Applications," Algorithms, 2025. Available: https://www.mdpi.com/1999-4893/18/3/157

[9] Olufunsho I. Falowo and Jacques Bou Abdo, "Empirical Study on Automation, AI Trust, and Framework Readiness in Cybersecurity Incident Response," Algorithms, 2026. Available: https://www.mdpi.com/1999-4893/19/1/62

[10] Ahmad Mohsin et al., "A Unified Framework for Human–AI Collaboration in Security Operations Centers with Trusted Autonomy," arXiv, 2024. Available: https://arxiv.org/pdf/2505.23397